ҚАЗАҚСТАН РЕСПУБЛИКАСЫ БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
«АЛМАТЫ ЭНЕРГЕТИКА ЖӘНЕ БАЙЛАНЫС УНИВЕРСИТЕТІ»
коммерциялық емес акционерлік қоғамы
«Ақпараттық қауіпсіздік жүйелері» кафедрасы

«ҚОРҒАУҒА ЖІБЕРІЛДІ»
Кафедра меңгерушісі с.ғ.к., доцент Бердібаев Р. Ш.
_____ «_____» _____ 2019 ж.
(қолы)

ДИПЛОМДЫҚ ЖОБА

Тақырыбы: «Тыйым салынған контентті табу үшін Интернет желісіндегі ақпараттық объекттерді интеллектуалдық талдау құралдарын құру»
Мамандығы: 5В100200 – «Ақпараттық қауіпсіздік жүйелері»
Орындаған: Нұрекен Елдос Тобы: СИБк-15-1
Ғылыми жетекші: т.ғ.қ., доцент, Омар Тұрғанбек Қалиұлы

Кеңесшілер:
Экономикалық бөлім бойынша:
э.ғ.к., профессор Арисбаева М.Т.
(ғылыми дәрежесі, атағы, аты-жөні)
_____ «17» 05 2019 ж.
(қолы)

Тіршілік қауіпсіздігі бөлімі бойынша:
аға оқытушы Торғаев Д.О.
(ғылыми дәрежесі, атағы, аты-жөні)
_____ «22» 04 2019 ж.
(қолы)

Есептеу техникасын қолдану бойынша:
т.ғ.к., доцент Омар Т.Қ.
(ғылыми дәрежесі, атағы, аты-жөні)
_____ «20» 05 2019 ж.
(қолы)

Мөлшер бақылаушы:
аға оқытушы Аскарова Ж.Б.
(ғылыми дәрежесі, атағы, аты-жөні)
_____ «30» 05 2019 ж.
(қолы)

Пікір беруші:
_____
(ғылыми дәрежесі, атағы, аты-жөні)
_____ «_____» _____ 201__ ж.
(қолы)

Алматы 2019

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
«АЛМАТЫ ЭНЕРГЕТИКА ЖӘНЕ БАЙЛАНЫС УНИВЕРСИТЕТІ»
коммерциялық емес акционерлік қоғамы

Басқару және ақпараттық технологиялар институты
Ақпараттық қауіпсіздік жүйелері кафедрасы
5В100200 – «Ақпараттық қауіпсіздік жүйелері» мамандығы

Дипломдық жобаны орындауға берілген

ТАПСЫРМА

Студент: Нуренен Елдос
(аты-жөні)

Жобаның тақырыбы: "Тайғақ сайтқан контентті табу үшін Интернет желісіндегі ақпараттық объектілерді интеллектуалдық талдау құралдарын құру

2018 ж. «26» 10 № 124 университет бұйрығымен бекітілді.

Аяқталған жұмысты тапсыру мерзімі: «____»_____20___ ж.

Жобаға алғашқы деректер (талап етілетін зерттеу (жоба) нәтижелерінің параметрлері және зерттеу нысанының алғашқы деректері): _____
Тайғақ сайтқан контентті анықтау пәтінің жұмыс істеу принциптерін зерттеу. Мәтіннің толқалқалған анықтау әдістерін қарастырып, қолданыстағы пәтіндерге шолу жүргізу. Деректерді жинау және талдау бойынша қиындық болатын алгоритмдерді зерттеп, оң алгоритмдердің артықшылықтары мен кемшіліктерін ажырату. Тайғақ сайтқан контентті табу үшін Интернет желісіндегі ақпараттық объектілерді талдау пәтінін әзірлеу. Әзірленген пәтінді тестілеуден өткізіп, толқалқалауға анықтау нәтижелілігіне мен әсерінің статистикасын жинау

Диплом жобасындағы әзірленуі тиіс мәселелер тізімі немесе диплом жобасының қысқаша мазмұны: _____
1. Тайғақ сайтқан ақпараттан мәтіліктің қорытылу жүйесі
2. Тайғақ сайтқан контентті анықтау үшін тәсілдердің шолуы
3. Мәтін толқалқалған талдау пәтіндеріне шолу
4. Нейрондық желілердің жұмыс істеу принциптері

5. Тоналдылықты талдау
6. Нүктені жіктеу үшін деректерді Технологиялар
7. Деректерді жинау және талдау алгоритмдерін жіктеу
8. Пайдаланушы контентті анықтайтын жүйені Python тілінің құралдарының көмегімен іске асыру

Графикалық материалдардың (міндетті түрде дайындалатын сызбаларды көрсету) тізімі:

1. ID-мен оқыту модулінің блок-сұлбасы
2. Пакеттерді жинау блок сұлбасы
3. Хаффман алгоритмінің архитектурасы
4. CBOW және Skip-gram архитектурасы

Негізгі ұсынылатын әдебиеттер:

1. Саймон Хайкин. Нейронные сети: полный курс Neural Networks: A Comprehensive Foundation. - 2-е изд. - М.: "Вильямс", 2006 - С. 1104.
2. Пан2Б. Thumbs up? Sentiment Classification using Machine learning Techniques / Б. Пан, Л. Ли. - М: Вильямс, 2002. - 312с.
3. М.В. Киселькина, Е.В. Котельников, Метод автоматической классификации текстов по тональности, основанный на словаре терминов.

Жоба бойынша жобаның бөлімдеріне қатысты белгіленген кеңесшілер

Бөлімдері	Кеңесшілері	Мерзімі	Қолы
Есептеу техникасы бойынша	Омар Т.Ж.		
Экономика бөлімі	Жиенбаева Ж.Т.	09.03 - 15.05.19	
Әкімшілік қауіпсіздігі бөлімі	Торгаев Э.Э.	08.04 - 22.04	

Диплом жобасын дайындау
КЕСТЕСІ

Бөлімдердің атауы, әзірленетін мәселелердің тізімі	Ғылыми жетекшіге ұсыну мерзімдері	Ескерту
1. Тыйым салынған ақпараттан желінің қорғау жүйесі	15.02.2019	
2. Тыйым салынған контентті анықтау үшін жасалатын шаралар	21.02.2019	
3 Мәтін тоналдығын талдау жүйелеріне шолу	02.03.2019	
4. Нейрондық желілердің жұмыс істеу принциптері	15.03.2019	
5. Тоналдылықты талдау	25.03.2019	
6. Жүйені әзірлеу үшін қолданылған технологиялар	03.04.2019	
7. Деректерді жинау ж/е талдау алгоритмдерін әзірлеу	12.04.2019	
8. Тыйым салынған контентті анықтайтын жүйені Python тілінің құралдарының көмегімен іске асыру	25.04.2019	
9. Техникалық-экономикалық негіздеме	08.05.2019	
10. Өміртіршілік қауіпсіздігі	20.05.2019	

Тапсырманың берілген уақыты «_____»____________________20___ж.

Кафедра меңгерушісі _______________ (т.ғ.к., доцент Бердібаев Р.Ш.)
(қолы) (аты-жөні)

Жобаның
ғылыми жетекшісі _______________ (т.ғ.к., доцент Омар Т.К.)
(қолы) (аты-жөні)

Орындалатын тапсырманы
қабылдаған студент _______________ (Курекен Е.)
(қолы) (аты-жөні)

АҢДАТПА

Интернет желісінің қарқынды дамуы нәтижесінде пайдаланушылар күн сайын жүздеген мың пікірді әлеуметтік желілерде, блогтарда, форумдарда, мамандандырылған алаңдарда жариялайды және бұл пікірлерді біздің қауіпсіздігіміз үшін толық көлемде өңдеу қажет. Бұл дипломдық жобаның мақсаты - тыйым салынған контентті табу үшін ақпараттық объектілерді интеллектуалды талдау құралдарын әзірлеу. Дипломдық жұмысты орындау барысында Python бағдарламалау тілінің пакеттеріне толық талдау жүргізілді және мәліметтерді жинау мен тазартуды тиімді жүзеге асыруға мүмкіндік беретін модульдерді таңдау жүргізілді. Тыйым салынған контентті анықтау үшін терең оқытудың нейрондық желілері негізінде контентті интеллектуалды талдайтын моделі әзірленді, ол python тілі және оның кітапханалары арқылы жүзеге асырылды.

АННОТАЦИЯ

В результате стремительного развития сети Интернет пользователи ежедневно публикуют сотни тысяч мнений в социальных сетях, блогах, на форумах, специализированных площадках и эти мнения необходимо для нашей безопасности обрабатывать в полном объеме. Цель данного дипломного проекта - разработка средств интеллектуального анализа информационных объектов для обнаружения запрещенного контента. В процессе выполнения дипломной работы был проведен подробный анализ библиотек на языке программирования python и произведен выбор модулей, позволяющих эффективно реализовать сбор и очистку данных. Для определения запрещенного контента была разработана модель интеллектуального анализа контента на основе нейронных сетей глубокого обучения, которая реализована посредством языка python и ее библиотек.

ANNOTATION

As a result of the rapid development of the Internet, users publish hundreds of thousands of opinions daily on social networks, blogs, forums, specialized sites and these opinions need to be processed in full for our security. The purpose of this diploma project - the development of means of intellectual analysis of information objects for the detection of prohibited content. In the course of the thesis was carried out a detailed analysis of the libraries in the programming language python and made the choice of modules to effectively implement the collection and cleaning of data. To determine the prohibited content, a model of content mining based on deep learning neural networks was developed, which is implemented through the python language and its libraries.

Мазмұны

Кіріспе

Әрбір екі жыл сайын әлемде деректер саны екі есе өседі. Әрбір Интернет пайдаланушы қандай да бір өнім, оқиға, адам және т. б. туралы интернет-ресурстарда үлкен көлемде өз пікірін қалдыру мүмкіндігі бар. Сондықтан дүниежүзілік тордағы барлық ақпарат пайдалы емес, мазмұнның бір бөлігі қолайсыз болып табылады. Оған ҚР аумағында тыйым салынған ұйымдардың сайттарын немесе есірткілік заттарды қолдануды насихаттайтын лицензиясыз контентті тарататын беттерді жатқызуға болады. Кәмелетке толмағандардың денсаулығы мен дамуына зиян келтіруі мүмкін ақпараттан қоршау аса маңызды болып табылады. Шынында да, билік өкілдері оған ерекше көңіл бөледі. Саяси қайраткерлер мен билік органдары адамдардың өздерін және олардың саясатын қалай қабылдайтынын түсіну үшін тоналдылықты талдауды жиі қолданады.

Дәстүрлі бұқаралық ақпарат құралдарын талдау үлкен пайда бере алады, бірақ бұл объективті көрініс пен істің нақты жағдайын туғызбайды. Ең толық ақпарат алу үшін әлеуметтік медиа мониторингін жүргізу қажет. Осылайша, жұмыстың өзектілігі интернет пайдаланушыларының көптеген ретсіз пікірлерді өңдеу мен талдауда қажеттілікпен қамтамасыз етіледі. Сарапшылардың мұндай ақпаратты өңдеу физикалық мүмкін емес. Алайда, қазіргі уақытта мұндай жүйелер өте қымбат болып табылады және көптеген шағын компаниялар оларды, әсіресе өз қажеттіліктеріне байланысты сирек зерттеулер үшін бере алмайды. Бұл міндетті шешу көптеген коммерциялық ұйымдар үшін практикалық құндылыққа ие болса да, ол салыстырмалы түрде жаңа болып табылады және өзінің кемшіліктері бар. Мұндай жүйелердің негізінде табиғи тілді өңдеу жатыр. Ең жиі шешілетін міндет — мәтіннің эмоциялық түсін анықтау. Бұл тапсырманың толық шешімі жоқ және орыс тілі үшін жеткілікті зерттелмеген.

1 Жұмыстың өзектілігі

1.1 Тыйым салынған ақпараттан жеткіліксіз қорғалу мәселесі

Интернет-пайдаланушылардың қалаусыз ақпараттан жеткіліксіз қорғалу мәселесі қазіргі уақытта үлкен өзектілікке ие болды. Осы саладағы террористік іс-әрекетті насихаттауға қарсы әрекет ету, лицензиясыз материалдардың таралуына қарсы күрес және кәмелетке толмағандарды қолайсыз контенттен қоршау өте маңызды. Қалаусыз ақпараттан қорғау әрекеттерін екі үлкен топқа бөлуге болады: кәмелетке толмағандарды қалаусыз материалдардан қоршау және заңнаманы бұзатын контентті блоктау. Екі бағыттардың өзектілігін билік өкілдері жақсы түсінеді және түрлі нормативтік-құқықтық актілерде баяндайды.

Бағдарламалық жасақтаманы әзірлеушілерде қажетсіз ақпараттан қорғау қажеттілігін түсінеді. Қазіргі уақытта "қауіпсіз іздеу" функциясы көптеген әлеуметтік желілерде бар (мысалы, Вконтакте) және web-сервистерде (Google-да Яндекс, Safe Search-да отбасылық іздеу). Қажетсіз және қауіпті ақпаратты детектеуді, сондай-ақ контентті сүзуді қамтамасыз ететін "Спутник" браузері сияқты ресейлік әзірлемелер бар. Сонымен бірге, ата-аналық бақылау модулі, кәмелетке толмағандарды қолайсыз ақпараттан қоршауды қамтамасыз етуге үлкен көмек көрсетеді. Сонымен қатар, қажетсіз ақпаратпен байланысты жеке міндет-таратылатын жарнамалық мазмұнды анықтау пайдаланушыларға электрондық поштадағы спаммен күресте өлкен үлес көрсетіп жатыр. Келтірілген мысалдар тұжырымдалған мәселе өзекті екенін көрсетеді

1.2 Тыйым салынған конетнтті анықтау үшін жасалатын шаралар

Қажетсіз ақпараттан қорғау мәселесінің үлкен өзектілігіне байланысты, қазіргі уақытта бірқатар шешімдер ұсынылған.

Бірінші, бірақ мүмкін тәсілдердің жалғыз емесі - қолмен мазмұнды тексеру және талдау. Бұл тәсіл сөзсіз артықшылыққа ие, нәтиженің жоғары сапасын қамтамасыз етеді, бірақ өте көп еңбек ресурстарын қамтиды. Бұл ақпараттың едәуір көлемін тиімді өңдеуге мүмкіндік бермейді. Үлкен деректер массивтерін өңдеуге қабілетті автоматты жүйелерді құру қажеттілігі қадағаланады. Мұндай жүйелер әртүрлі түрде жасалуы мүмкін.

Мәліметті тексеруге негізделген шешімдерінің бірі болып тұрақты өрнектердің көмегімен қойылатын ережелері саналады. Олардың негізінде: онлайн-ресурсқа рұқсат беру немесе бұғаттау туралы шешім қабылданады.

Тағы бір опция Яндекс немесе Google сияқты сайттардағы қауіпсіз іздеу режимін автоматты түрде қамтитын бағдарламалар. Олардың негізгі кемшілігі - тым шектеулі қолдану.

Басқа жүйелер алдын ала белгілі бір "ақ" тізімнің негізінде жұмыс жасайды. Қарама-қарсы тәсіл "тыйым салынбаған барлық нәрсе – рұқсат етілген", яғни "қара" тізімдерді пайдалануда.

Алайда, көбінесе барлық аталған тәсілдер өзінің тиімсіздігін көрсетеді: жасалатын "қара" және " ақ " тізімдер блокталатын контентті рұқсат

етілгеннен ерекшеленетін нұсқалардың алуан түрлілігі үшін барлығын жаппайды. Бұл жағдайға мысал болып Интернет желісіндегі ақпараттың жоғары өзгергіштігі: жаңа сайттардың жиі пайда болуын және бір веб-беттің ішінде мазмұнды жылдам жаңартуды келтіруге болады.

Жағымсыз ақпаратты анықтаудың қиындауының факторы болып, әртүрлі қарама-қайшы және өзгермелі деректер көлемі, оның ішінде web-ресурстарды құру ерекшеліктері саналады. Әдетте олар күрделі иерархиялық құрылымы бар және көптеген элементтерден тұрады, мысалы мәтіндік және графикалық мазмұн. Бағдарламалық жағымсыз мазмұн тек бір ғана мәтіндік белгілердің негізінде анықталмайды. Жиі сайттың бағытын анықтау үшін ақ тізімі секілді тәсілдер жеткіліксіз.

Алдын ала белгілі бір ережелер немесе тізімдер, өзін-өзі оқи алмайды. Алайда, мұндай жағдайда машиналық оқыту әдістерін (Data Mining) қолдануға болады. Сонда негізгі идея болып, зерттелетін объектіні алдын ала белгілі жиындардың біріне жатқызу қажеттілігі саналады.

Жіктеу міндетін дұрыс шешу өте маңызды және көптеген салаларда елеулі табыстарға алып келеді. Мысалы, коммерция саласында пайдаланылатын автоматты клиенттерді түсініп жекелендіру олардың белгілі бір мақсатты аудиторияға тиесілігін тану, компанияларға неғұрлым икемді маркетингтік саясатты жүргізуге мүмкіндік береді. Пластикалық карталармен операциялар кезінде қауіпсіздікті қамтамасыз ету үшін маңызды электрондық төлем жүйелерінде кеңінен жіктеу пайдаланылады.

Жіктеу әдістері клиенттің іс-әрекеттерінің негізінде оны екі санаттың біріне жатқызу керек, заңды пайдаланушы немесе қаскүнем, осылайша алаяқтық жағдайларын анықтау жүзеге асырылады. Менеджер екі санаттың қайсысының өкілі: "төлем төлеуге қабілетті" немесе "төлем жасауға қабілетсіз" - несие алуға өтініш білдірген клиент болып табылатындығын анықтауда, қолданудың тағы бір нұсқасын көрсетеді. Осы операцияны орындау үшін барлық қол жетімді талдау жүргізіледі, талдау нәтижелері негізінде шешім қабылданады [1].

Қажетсіз ақпараттан қорғау мәселесіне қатысты ата-ана жүйесінің жұмысы мысал бола алады, мүмкін web-беттерді санаттар бойынша бөлетін және олардың жағымсыздығын анықтайды ("ересектерге арналған сайттар", " алкоголь», "қару", "есірткі" және т. б.).

Интернет-пайдаланушылардың қолайсыз ақпараттан қорғалуын арттырудың негізгі нұсқасы болып машиналық оқыту ағымындағы жіктеу әдістері саналады. Web-беттерді санаттау және талдау негізінде жүзеге асырылуы мүмкін.

Мәтін бойынша жіктеу әдісі кеңінен қолданылады, ол екі тізбекті кезеңнен тұрады. Бірінші кезеңде деректерді классификатор қабылдайтын нысанға ауыстыру арқылы дайындау жүргізіледі. Осы кезеңдегі іс-қимыл реттілігінің мысалдарының бірі белгілеу және web-беттердің тестілік мазмұнын алу, стемминг операциясын орындау арқылы тыныс белгілерін,

стоп-сөздерді сылтау, есімдік және т. б. элементтерін алып тастау . (Naive Bayess, SVM және т.б.).

Көбінесе тестке бөлінген әдістер және оқыту үлгісі қолданады (supervised method). Көрнекі мысал болуы ретінде SVM әдісін сипаттауға болады. Алайда, Ko зерттеушілері мәтін бойынша ресурстардың аз шығындарымен жіктеуге, сондай-ақ оқыту таңдауларын жасауға арналған Seo алдын ала оқытусыз әдісін ұсынады (unsupervised method). Онда құжат ұсыныстарға бөлінеді, содан кейін әрбір ұсынысқа алдын ала дайындалған кілт сөз тізімдері мен ұсыныстар ұқсас метрикасы (sentence similarity measure) негізінде санат салыстырылады.

Мәтіндік жіктеудің әр түрлі нұсқаларының қызықты мысалдарының бірі антиспам техникасы болып табылады. Қарау барысында техника беттегі сөздердің жалпы санын, сөздің орташа ұзындығын, web-бет сөздерінің неғұрлым жиі кездесетін сөздердің жиынтығына тиістілігі, статистиканы есептеу n-грамм (N-граммдық комбинациялар) негізінде санаттандыру нұсқаларын анықтауды ұсынады.

Басқа балама-құжаттарды жинау ретінде қараудан өту лексикалық деректер базасынан алынатын олардың мәндерін талдауға арналған сөздер. Бұл мысалы, орыс тілінде "коса" сөзі өрілген шашты, бау-бақша құралын немесе тас жиегін көрсете алады. Ұқсас ағылшын сөзі "base" әскери лагерь немесе термин бейсболда. Алайда жүргізілген эксперименттер сөздердің мағынасының ұқыптылық анықтау көлемін арттырады.

Алайда, мәтіндік классификация web-беттердің ерекшеліктерін ескермейді. HTML-құжат, басқа құжаттарға сілтемелермен байланысты, суреттер және басқа да мәтіндік емес элементтерден тұрады. Сондықтан URL талдауына негізделген әдіс қарқынды дамып жатыр. Егер параркша мүмкін оқырмандардың қызығушылығын тудырмаса, интернеттегі бетке қолданушылар сирек кіреді. Яғни сайттың мекен-жайы қандай да бір түрде оның тақырыбын көрсетуі керек. Талдау тәсілдерінің бірі болатын құрамдас бөліктерге URL-ды бөлу. Мұндай тәсіл фишинг сайттарынан қорғау мақсатында URL-ды талдау кезінде іске асырылған.

Сондай-ақ, сайт мекенжайының қандай да бір бөлігі қандай позицияда екенін анықтау қажет. Мысалы, авторлар "paypal" фрагменті бар келесі 2 сілтемелерді береді: http://www.paypal.com/ және http://www.paypal.com.hostin gcompany.com/.

Осылайша, URL-дың әрбір фрагменті өзі мен оның позициясын қамтитын екі өлшемді Вектор түрінде ұсынылады, одан кейін оқытылған жіктегішке кіруге беріледі.

Тағы бір әдіс хост аты ұзындығы мен әр түрлі таңбалар санын (мысалы, нүктелер) және осы таңбалар арасында жасалған URL фрагменттерін талдау. Сонымен қатар, хост туралы ақпарат негізінде белгілер қолданылады (географиялық ерекшеліктер, тіркеу күні, TTL шамасы және т.б.). Барлық осы атрибуттар қандай да бір классификатор арқылы өңделеді (Naïve Bayess, SVM, Logistic Regression).

URL-ды фрагменттерге бөлу опцияларының бірі энтропияны қолдану. Мұндай тәсіл бірнеше сөздер біріктірілген домендердің атауын құрамдас бөліктеріне бөлуге мүмкіндік береді, мысалы, "activatealert". Қалған элементтер арасында ең аз энтропиясы бар бөлшектердің бірі ең ықтимал жаңа фрагмент болады.

Бұл әдіс санаттаудың жақсы нәтижелерін көрсете алады жеке міндеттерді шешу кезінде ("спам" / "обычное письмо", "phishing" / " benign»), алайда, жалпы жағдайда, санаттың еркін саны мен құрамы кезінде жіктеу сапасы төмендейді. Басты себеп - бұл шындығында, Интернеттегі беттің мекен-жайы оның мазмұнымен сәйкес келмейді.

1.3 Мәтін тоналдығын талдау жүйелеріне шолу

Мәтіннің тоналдығын талдау мәтіннен эмоционалды боялған лексика мен авторлардың мәтінде сөз болып отырған объектілерге қатысты эмоционалды қарым-қатынасын алуға мүмкіндік береді. Қазіргі заманғы жүйелердің көпшілігі екілік бағалауды пайдаланады – "оң сентимент" немесе "теріс сентимент", алайда кейбір жүйелер тональдылықтың күшін көруге мүмкіндік береді. Қазіргі әлемде қандай да бір жағдайларда біздің таңдауымызға басқа адамдардың пікірі жиі әсер етеді-біз оны интернет – дүкенде тапсырыс берместен бұрын тауар туралы пікірлерді оқимыз, сайлауда қандай да бір кандидат үшін дауыс бермес бұрын басқа адамдардың пікірін білеміз, біз баратын жоғары оқу орнын, жұмыс орнын және мейрамхананы ұзақ және мұқият таңдаймыз. Бұл ақпарат маркетологтар, әлеуметтанушылар және басқа да көптеген мамандар үшін маңызды қызығушылық тудырады. Бұдан басқа, интернет-ресурстардың иелері үшін пайдаланушылардың пікірін білу өмірлік маңызды – бұл сіздің порталыңызда жасалған жаңалықтарға, сіздің сайтыңыздағы жаңа жаңалықтарға қатысты пікір немесе сіздің интернет-дүкендегі тауарды пайдаланушылардың бағалауы. Жоғарыда айтылғандардың барлығы мәтіннің тоналдығын талдау міндетін өзекті етеді. Алайда, бұл тапсырманың келешегі мен өзектілігіне қарамастан, орыс тіліндегі мәтіннің үнсіздігін талдай алатын жүйелердің салыстырмалы аз саны бар.

"SentiStrength" – m.Thelwall, K. Buckley, G. Paltoglou және D. Cai әзірлеген жүйе. Бастапқыда, бұл жүйе ағылшын тіліндегі қысқа құрылымсыз бейресми мәтіндерді талдау үшін әзірленген. Алайда, ол басқа тілдердегі мәтіндермен жұмыс істеу үшін, оның ішінде орыс тіліндегі мәтіндерге де теңшеуге болады. Нәтиже екі баға түрінде беріледі – мәтіннің оң құрамын бағалау (+1-ден +5-ке дейінгі шкала бойынша) және теріс құрамды бағалау (-1-ден -5-ке дейінгі шкала бойынша). Бұдан басқа, бағалауды басқа түрде ұсыну мүмкіндігі бар: бинарлық бағалау (позитивті/негативті мәтін) Тернарлық бағалау (позитивті/негативті/бейтарап) -4-тен +4-ке дейінгі бірыңғай шкала бойынша баға арқылы Алгоритм әрбір шкала үшін мәтінде үнсіздіктің ең жоғары мәнін іздеуге негізделген (яғни, барынша теріс бағасы бар сөздерді және барынша оң бағасы бар сөздерді іздеу). Алгоритмнің

жұмысы кезінде сөздердің қарапайым өзара әрекеттесуі ескеріледі (мысалы, күшейткіш-сөздер олар әрекет ететін сөз үшін үнсіздіктің мәнін күшейтеді – "өте зұлымдық" жай ғана "зұлымдық" емес, теріс бағаға ие болады) және идиоматикалық өрнектер.Жүйенің кемшіліктері: жүйе орыс тілі үшін теңшелетін болса да, онда жүзеге асырылған алгоритмдер оның ерекшелігін, оның ішінде орыс морфологиясын ескермейді, бұл бірқатар проблемаларға әкеледі. Мысалы, орыс тіліндегі жүйенің толыққанды жұмыс істеуі үшін деректер банкінде әрбір сөз үшін барлық сөз формалары болуы қажет. Сонымен қатар, жүйе тек мәтіннің жалпы тоналдығын ғана санайды.[2]

"Аналитикалық курьер" және "X-files" жүйесінің құрамындағы мәтіннің үнсіздігін талдау компоненті – "Ай-Теко"компаниясымен әзірленген. Мәтіннің үнсіздігін анықтау компоненті сөздіктер мен ережелерге негізделген әдісті жүзеге асырады. Бұл жүйе пайдаланушыға белгіленген ұсыныстар массивін береді. Ұсыныстарда тоналдылық объектілері (бар болса) және оларға қатысты тоналдылық беретін сөздер тізбегі бар. Бұдан басқа, табылған сөздер тізбегі негізінде әрбір сөйлем үшін жалпы тоналдылық есептеледі. Жалпы тоналды есептеу үшін бірқатар арнайы ережелер қолданылады. Мысалы ("Доктор Смит тұмау ауруымен ауыратын науқасты емдеді" ұсынысы үшін), оң етістіктің "емдеу" теріс тізбекпен үйлесуі (бұл жағдайда "тұмау ауруымен ауыратын адам") оң етістіктің (біздің мысалда – "дәрігер Смитке") жататынын көрсететін ереже бар. Тональдығы тернарлық шкала бойынша бағаланады (оң/теріс/бейтарап). Жүйе бірнеше кезеңдерде жұмыс істейді: мәтінді алдын ала өңдеу, табылған сөздерді бөлу және жіктеу, табылған сөздерді бір-бірімен байланысты тізбекке біріктіру. Тоналдылық объектілерін бөлу жүйенің кемшіліктері: мәтінді сандық бағалаудың болмауы.

"Ваал" – Шалак Владимир әзірлеген жүйе. Бұл жүйе "адамның сана-сезіміне жеке сөздер мен мәтіннің фонетикалық құрылымының негізсіз эмоциялық әсерін" бағалауға арналған. Жүйенің жұмысы мәтінді жиілік сөздікке айналдыруға және кейбір сөздерді белгілі бір психолингвистік санаттарға жатқызуға негізделген. Талдау нәтижесі пайдаланушыға осы мәтінге/сөзге ("тегіс – сопақ", "күшті – әлсіз") және т.б. қатысты бірқатар критерийлер бойынша бағалар жиынтығы түрінде беріледі. Сонымен қатар, бұл өнімді психолингвистика саласындағы мамандар болып табылмайтын адамдардың пайдалануы мүмкін емес.

RCO Fact Extractor жүйесінің құрамындағы тональдылықты талдау компоненті-RCO компаниясы әзірлеген жүйе. Мәтіннің үнсіздігін талдау үшін жүйе ережелерге негізделген тәсілді қолданады. Бұл жүйе мәтіннің синтаксистік құрылымын және әр түрлі сөздердің өзара әрекеттесуін ескереді. Құрамдауыштың жұмысы бес кезеңнен тұрады: барлық нысандардағы объект туралы барлық ескертулерді тану, атап өтудің толық, қысқа және басқа да нысандарын және конструкциялардың толық синтаксистік талдауын қоса алғанда, мақсатты объектімен байланысты барлық оқиғалар мен белгілер көрсетілетін айқын байқалатын позицияларды және әрбір пропозиция үшін эмоционалдық-коннотативтік жағдайларды сипаттайтын пропозицияларды

бөлу және жіктеу "позитив-негатив" тоналдығы туралы шешім қабылдау оның құрамында эмоциялық-коннотативтік, тоналды және бейтарап сөздер, терістеу құралдары мәтіннің жалпы тоналдығын бағалау оған кіретін барлық пропозициялардың тоналдылықтары негізінде өз жұмысы үшін компонент RCO компаниясында әзірленген мәтінді синтаксистік талдау және атауларды теңестіру модульдерін пайдаланады. Жүйенің кемшіліктері: мәтінді сандық бағалаудың болмауы.

2. Зияткерлік талдау әдістері мен құралдары

2.1 Машиналық оқытудың ақпарат қауіпсіздігінде рөлі

Кейбір сарапшылар ақпараттық қауіпсіздік саласында машиналық оқыту және жасанды интеллект технологияларын қолдану үшін жаңа құралдарды пайдалану тәжірибесін бастап жатыр. Сондықтан қажетті технологиялар мен шектерді таңдау әдістері, көптеген ақпараттық қауіпсіздікті қамтамасыз ету үшін арналған өнімдерде іске қосылады. Эльман Бейбутов (IBM) ойынша, мүмкіндігінше көп деректерді жинап өңдеу маңызды, ал деректерді дұрыс құрылымдау және автоматтандырылған қорғау құралдары тиімді жұмыс істеу үшін өңдеу жолдарын түсіну керек деп санайды. Мысалы, IBM-де бүгінде Watson суперкомпьютерінің қуатын пайдаланатын бірқатар бағдарламалар жасалған. Бастапқыда бұл жоба денсаулық сақтау саласы үшін іске қосылды, бірақ қазіргі кезде Watson деректерді құрылымдай алады және осы суперкомпьютердің тиімділігі жаңа мүмкіндіктерінің арқасында арта түсті.

Классикалық машинамен оқыту алгоритмдері (мұғаліммен немесе мұғалімсіз) қарапайым деректер мен түсінікті белгілер болған жағдайда қолданылады. Мысалы, шетелде қолма-қол ақша алғаннан кейін төлем картасын бұғаттау. Мұнда бәрі оңай, әдетте сіздің барлық транзакцияларыңыз үй аймағында өтеді, ал мұнда аномалия – кенеттен (егер сіз сапар туралы өз банкіңізді алдын ала ескертпесе) елден тыс қолма-қол ақша алу болып саналады. Қолданылатын машиналық оқытудың барлық алгоритмдерінің 50% -ға дейін ескірген классикалық алгоритмдер болып табылады. Сол алгоритмдер көмегімен қажетті тапсырманы тез шешуге болады.

Барлық классикалық алгоритмдердің 75% – ға дейін-мұғаліммен оқыту, яғни белгіленген немесе таңбаланған деректермен жұмыс істейді. Мысалы, модельдер: бұл спам, және бұл жоқ; бұл DDoS, және бұл жоқ; бұл алаяқтық (фрод) және бұл жоқ. Мұғаліммен оқыту арқылы сіз жаңа деректерді оңай жіктей аласыз, оларда аномальды нәрсе анықтай аласыз. Мұндай алгоритмдер арқылы бұрын белгісіз зиянды кодтың жүктелуін, спам - және фишингтік шабуылдарды, DGA-домендерді (автоматты түрде жасалатын зиянды домендерді), командалық серверлермен ботнеттермен коммуникацияларды анықтауға болады. Мұғаліммен ең танымал алгоритмдер жіктеу және регрессияны атауға болады. Жіктеу санатты болжауға, ал регрессия – мәнді болжауға мүмкіндік береді. Егер шабуылдар саны өсуі кезінде қорғаныс керек болса, онда сізге регрессия қажет, ал егер сіз жарты жылдан кейін қандай шабуылдар көп болатынын түсінгіңіз келсе, жіктеу қажет болады. Екі түрдің әрқайсысы мұғаліммен бірге машиналық оқыту алгоритмдерінің ішкі жинағына бөлінуі мүмкін. Айталық, жіктеуге шешімдер ағаштары, random forest немесе SVM жатады. Олардың көмегімен, атап айтқанда, SQL Injection шабуылдарын немесе күдікті HTTP трафигін анықтауға болады.

Бірақ кіріс деректері анықталмаған кезде не істеу керек? Қорғаныс жүйесі бір тіркелгіге кірудің төрт сәтсіз әрекеттерін белгілеген жағдайда бұл айқын бұзушылық, өйткені бағдарлама ережелерінде үш сәтсіз әрекеттен

кейін шектеу қарастырылған. Төрт және одан да көп сәтсіз әрекеттерді анықтау үшін машиналық оқыту қажет емес. Бірақ әр түрлі географиялық нүктелерден бір есептік жазбаның белсенділігін, бір тәулік ішінде зиянды белсенділікті білдіруі мүмкін. Мысалы, іссапарға аттанған жағдайында қорғалатын жүйеге әртүрлі қалардан кірсеңіз, айталық, Мәскеу, Лондон, Нью-Йорк және Чикаго әуежайларынан. Мұндай сценарийлерді банктер жиі алаяқтық деп санайды. Өйткені қандай орын үйреншікті екенін алдын ала білмейміз, ал қандай орын жоқ. Мұнда мұғалімсіз оқыту және оның алгоритмдерінің бірі – кластерлердің ұқсас оқиғаларын біріктіруге мүмкіндік беретін кластерлеу көмектеседі. Стандартты емес кіру орнының пайда болуы (кластерлерге түспейтін) аномалия болып табылады және есептік жазбаны ұрлау белгісі болып саналады. Бұл тәсіл мұғаліммен оқытудан гөрі дәл емес.

Мұғалімсіз оқуды жақсы іске асыратын басқа сценарий-ақпарат ағуын анықтау немесе әкімші саботаж. Сіз бұлттан алып жазылатын немесе бір компьютер арқылы жергілікті желі арқылы жүктейтін қарапайым және аномалды файлдар санының жазылуы арасындағы айырмашылығын айта алмайсыз. Сіз түрлі пайдаланушылар мен пайдаланушылар топтарының осы белгіні кластерлерге біріктіріп, сол арқылы қалыпты және ауытқушылық мінез-құлықты анықтай отырып, өзара салыстыруға ғана мүмкіндігіңіз бар. Мысалы, пайдаланушылар күніне шамамен 100 Мбит деректерді Интернетке түсіреді, бірақ бір күнде бір пайдаланушы 10 Гбит-ден астам жүктеп алды. Бұл машинамен оқытусыз анықталатын, айқын аномалия. Алайда, машиналық оқыту бірнеше белгілерді біріктіруге көмектеседі (мысалы, деректер көлемі, уақыт, хаттама, деректер түрі, алушының мекенжайы) және деректерді ұрлау үшін жазылған бағдарламаларды анықтайды.

Нейрожелілер – бұл соңғы уақытта үлкен танымалдыққа ие болатын мұғалімсіз машиналық оқыту алгоритмдерінің бірі. Әдетте олар өте күрделі датасеттер (биометриядағы бет бейнелері, сондай-ақ құжаттардың дауысы немесе бейнелері) немесе датасеттегі модельді таңдау белгілерін белгілеу қиын жерде қолданылады. Нейрожелінің негізгі идеясы – оның ішкі қабаттарына датасетте не маңызды және оқу процесінде одан не алынуы тиіс туралы өз пайымдауларын жасау мүмкіндігі. Жоғарыда жазылған барлық мысалдар нейросетрлермен табылуы мүмкін, бірақ әдетте оларды неғұрлым күрделі сценарийлерде қолданады – жалған құжаттарды тану, биометрия үшін қауіп-қатерлермен күрес, дауыстық коммуникациялардағы ақпараттың жылыстауын іздеу, қауіпсіздік бойынша мәтіндерді тану және т.б. Нейросетьдердің елеулі кемшіліктерінің бірі кері байланыстың жоқтығы, яғни кіріс деректерінен дәл осындай нәтиже алынғанын түсіндірудің мүмкін еместігін атауға болады [3].

Қазіргі замандағы ақпараттық қауіпсіздік бірқатар қиындықтарға тап болады, олардың арасында үлкен оқиғалар ағындарын, сараптаманың төмендеуі мен персоналдың жетіспеуін атап өту керек. Бұл ретте, қабылданатын қорғау шараларына қарамастан, шабуылдар саны өсуде. Қазіргі уақытта қауіптерді анықтаудың орташа кезеңі 200 күнге жуық құрайды, бұл

пайдаланылатын қорғаныс құралдарының реактивтілігінің нәтижесі болып табылады. Сондықтан бүгін, зиянды белсенділікке қарсы күрестің жаңа әдістерін қолдану маңызды.

Соңғы алты жыл ішінде киберқауіпсіздік нарығында жасанды интеллектпен байланысты 220-дан астам шабуыл тіркелді. Бұл бағыт қазіргі уақытта ең көп таралған мәмілелердің бестігіне кіреді, ал ақ нарығының көптеген ойыншылары (мүмкін, отандық ойыншыларды қоспағанда) өз өнімдеріне интеграцияланатын машиналық оқыту технологияларына белсенді инвестиция салады. Бірақ қарапайым тұтынушы өзінің массасында жасанды интеллектінің барлық артықшылықтарын белсенді пайдалана алмайды. Ол үшін дұрыс өңделген датасеттер жоқ, не ең маңыздысы бар талдау үлгілерін өз бетінше әзірлеуге немесе қолдануға қабілетті білікті талдаушылар жоқ. Алайда, сатып алынатын немесе пайдаланылатын шешімдерде машиналық оқыту модельдерін пайдалану үшін де осы технологияның нені білдіретінін түсіну қажет.

Дегенмен, машиналық оқыту панацея емес. Біріншіден, датасеттерге де, алгоритмдерге де бағытталған шабуылдардың тұтас класы бар, бұл қате шешімдерге, өткізіп алған шабуылдарға немесе жалған іске қосылуларға әкелуі мүмкін. Екіншіден, зиянкестер де өзінің криминалдық қызметінде машиналық оқыту әдістерін қолдана бастайды – зиянды бағдарламалар жасау, пайдаланушылардың мінез-құлқын талдау, дербес деректерді құрастырушы бағдарламаны әзірлеу, осалдықтарды іздеу, фишинг, парольдерді таңдау, жеке тұлғаны ауыстыру, қорғау жүйелерін аралау және т. б. Сондықтан ақпараттық қауіпсіздікке машиналық оқытуды қолданусыз киберқауіпсіздіктің қазіргі заманғы жүйесін елестету мүмкін емес.

2.2 Жасанды нейрондық желілері

Жасанды нейрондық желі (ЖНЖ) – математикалық модель, сондай-ақ биологиялық нейрондық желілерді тірі ағзаның жүйке жасушаларының желілерін ұйымдастыру және қызмет ету принципі бойынша құрылған оның бағдарламалық немесе аппараттық іске асырылуы. Бұл ұғым мидағы процестерді зерттеу кезінде және осы процестерді модельдеу кезінде пайда болды. Алғашқы әзірленген желілердің бірі болып У. Маккалок пен У. Питтс әзірлеген нейрондық желілері болды. Оқыту алгоритмдерін әзірлегеннен кейін, алынған модельдерді практикалық мақсаттарда қолдана бастады, болжамдау міндеттерінде, бейнелерді тану үшін, басқару міндеттерінде және т. б.

ЖНЖ – біріктірілген және өзара әрекеттесетін қарапайым процессорлар (жасанды нейрондар) жүйесі. Мұндай процессорлар әдетте қарапайым (әсіресе жеке компьютерлерде қолданылатын процессорлармен салыстырғанда). Мұндай желінің әрбір процессоры тек қана ол мезгіл-мезгіл алатын сигналдармен және ол мезгіл-мезгіл басқа процессорларға жіберетін сигналдармен айналысады. Сонымен бірге, басқарылатын және өзара іс-

қимылмен ерекшеленетін үлкен желіге қосылғанда, мұндай қарапайым процессорлар бірге өте күрделі тапсырмаларды орындауға қабілетті.

Машиналық оқыту тұрғысынан нейрондық желі бейнелерді тану әдістерінің, дискриминанттық талдаудың, кластерлеу әдістерінің және т. б. жеке жағдайы болып табылады.

Математикалық тұрғыдан алғанда, нейрондық желілерді оқыту-бұл сызықты емес оңтайландырудың көппараметрлік міндеті.

Кибернетика тұрғысынан, нейрондық желі адаптивті басқару міндеттерінде және робототехника үшін алгоритмдер ретінде пайдаланылады.

Есептеу техникасы мен бағдарламалаудың дамуы тұрғысынан нейрондық желі – тиімді параллелизм мәселесін шешу тәсілі.

Ал жасанды интеллект тұрғысынан, ЖНЖ коннективизмнің философиялық ағымының негізі және компьютерлік алгоритмдердің көмегімен табиғи интеллектіні құру (модельдеу) мүмкіндігін зерттеу бойынша құрылымдық тәсілдеменің негізгі бағыты болып табылады.

Нейрондық желілер осы сөздің әдеттегі мағынасында бағдарламаланбайды, олар оқиды. Оқыту мүмкіндігі-дәстүрлі Алгоритмдер алдындағы нейрондық желілердің басты артықшылықтарының бірі. Техникалық оқыту нейрондар арасындағы байланыс коэффициенттерін табу болып табылады. Оқыту барысында нейрондық желі кіріс деректері мен шығу арасындағы күрделі тәуелділікті анықтауға, сондай-ақ қорытуды орындауға қабілетті. Бұл дегеніміз, табысты оқыту жағдайында желі оқыту үлгісінде болмаған деректердің, сондай-ақ толық емес және "ойластырылған", ішінара бұрмаланған деректердің негізінде дұрыс нәтижені қайтара алады. Кіріс деректері ретінде графтар, мәтіндер, деректер базасына сұраныс нәтижелері және т. б. түрінде ұсынылатын неғұрлым күрделі жағдайлар да кездеседі. Әдетте, олар деректерді алдын ала өңдеу және белгілерді алу арқылы бірінші немесе екінші жағдайға келтіріледі.

Бейнелер ретінде өз табиғаты бойынша әртүрлі объектілер: мәтін символдары, бейнелер, дыбыстардың үлгілері және т.б. болуы мүмкін. Үлгі әдетте белгілер мәндерінің векторы ретінде ұсынылады. Бұл ретте барлық белгілердің жиынтығы үлгі жататын классты бір жақты анықтауы тиіс. Егер белгілер жеткіліксіз болса, желі бір үлгіні бірнеше класстармен салыстыра алады. Желіні оқыту аяқталғаннан кейін оған бұрын белгісіз бейнелерді көрсетуге және белгілі бір класқа жататындығы туралы жауап алуға болады.

Кластеризация деп класстардың саны да, белгілері де алдын ала белгісіз болған кезде класстарға көптеген кіру сигналдарын бөлу түсініледі. Оқытудан кейін мұндай желі кіріс сигналының қай класына жататынын анықтауға қабілетті. Желі сондай-ақ кіріс сигналы бөлінген класстардың ешқайсысына қатысы жоқ екендігін анықтаса, бұл оқыту таңдауында жоқ жаңа деректердің белгісі болып табылады. Осылайша, мұндай желі бұрын белгісіз жаңа сигнал кластарын анықтай алады. Желі бөлінген класстар мен пәндік салада бар кластар арасындағы сәйкестікті адам белгілейді. Кластерлеуді, мысалы, Кохоненнің нейрондық желілері жүзеге асырады.

Кохоненнің қарапайым нұсқасында нейрондық желілер үлкен болмайды, сондықтан оларды гиперслоздарға (гиперколонкаларға) және ядроларға (микроколонкаларға) бөледі. Егер адамның миымен салыстырсақ, параллель қабаттардың мінсіз саны 112-ден аспауы тиіс. Бұл қабаттар өз кезегінде гиперколонка (гиперколонка) құрайды, онда 500-ден 2000 микроколонка бар. Бұл ретте әрбір қабат осы қабаттарды басып өтетін көптеген гиперполонокқа бөлінеді. Микроколонкалар шығуда нәтиже ала отырып сандармен және бірліктермен кодталады. Қажет болса, артық қабаттар мен нейрондар жойылады немесе қосылады. Нейрондар мен қабаттардың санын таңдау үшін өте ыңғайлы. Мұндай жүйе нейрондық желілерге икемді болуға мүмкіндік береді. Мұндай желінің шоғыры шығу қабатындағы нейрондардың саны, әдетте, анықталатын класстар санына тең болуымен сипатталады. Бұл ретте нейрондық желінің шығысы мен ол ұсынатын сынып арасында сәйкестік белгіленеді. Желі қандай да бір бейнені көрсеткенде, оның шығу жолдарының бірінде бейненің осы класқа тиесілі екендігі туралы белгі пайда болуы тиіс. Сонымен қатар басқа шығуларда бұл классқа тән емес деген белгі болуы тиіс. Егер екі немесе одан да көп шығуда класына тиістілік белгісі болса, желі өз жауабында "сенімді емес" деп есептеледі. желілер осы сөздің әдеттегі мағынасында бағдарламаланбайды, олар оқиды. Оқыту мүмкіндігі-дәстүрлі Алгоритмдер алдындағы нейрондық желілердің басты артықшылықтарының бірі. Техникалық оқыту нейрондар арасындағы байланыс коэффициенттерін табу болып табылады.[4]

Нейрон қабілетін жинақтау және бөлу жасырын тәуелділіктер арасындағы кіріс және шығыс деректер. Сонымен қатар, қазіргі уақытта, желі тұрақты мәндердің және қазіргі кездегі қандай да бір факторлардың негізінде қандай да бір бірізділіктің болашақ мәнін болжауға қабілетті. Алдыңғы өзгерістер шын мәнінде қандай да бір дәрежеде болашақты алдын ала анықтағанда ғана болжау мүмкін екенін атап өткен жөн. Мысалы, өткен аптада баға белгілеулер негізінде акциялардың баға белгілеулерін болжау табысты болуы мүмкін, ал соңғы 50 жылда деректер негізінде ертеңгі лотерея нәтижелерін болжау ешқандай дәлдің бермейді.

Нейрондық желілер үздіксіз функцияларды аппроксимациялай алады, жалпыланған аппроксимациялық теорема дәлелденген. Сызықтық операциялар мен каскадты қосылыстардың көмегімен ерікті сызықсыз элементтен берілген дәлдікпен кез келген үздіксіз функцияны есептейтін құрылғыны жүзеге асыруға болады. Бұл дегеніміз, нейронның сызықты емес сипаттамасы еркін болуы мүмкін. Бірақ, кез келген желі әмбебап аппроксиматор болып қалады және құрылымды дұрыс таңдау кезінде кез келген үздіксіз автоматтың жұмыс істеуін жеткілікті түрде дәл аппроксимациялау мүмкін.

Егер деректер бір-бірімен тығыз байланысқан болса, нейрожелілердің әр түрлі параметрлер арасындағы өзара байланысты анықтауға қабілеттілігі үлкен өлшемдік деректерді жинақы көрсетуге мүмкіндік береді. Кері процесс – ақпарат бөлігінен бастапқы деректер жинағын қалпына келтіру-ассоциативті

жады деп аталады. Ассоциативті жады, сондай-ақ, аяқталған, зақымдалған кіріс деректерінен бастапқы сигнал, бейнені қалпына келтіруге мүмкіндік береді. Гетероассоциативті жады есебін шешу мазмұнға қатысты жадты жүзеге асыруға мүмкіндік береді.

2.3 Табиғи тілді өңдеу

Табиғи тілді өңдеу (Natural Language Processing, NLP) – жасанды интеллект пен математикалық лингвистиканың жалпы бағыты. Ол компьютерлік талдау және табиғи тілдерді синтездеу мәселелерін зерттейді. Жасанды интеллектке қатысты талдау тілдің түсінігі сауатты мәтіннің генерациясын білдіреді. Бұл мәселені шешу компьютер мен адамның өзара әрекеттесуінің ыңғайлы нысанын құруды білдіреді.

Табиғи тілді өңдеу білім базасын толтыру, сұрақтарға жауаптарды қалыптастыру және диалог жүргізу мақсатында мәтіндерді түсінуге бағытталған, тілді тану мен генерациялау, жіктеуді, мәтіндерден білімді экстракциялауды және басқа да әрекеттерді қамтиды. Мәтіндерді өңдеу міндеті алғаш рет елуінші жылдары американдық лингвист Ноам Хомскийдің табиғи тілдің грамматикасы бойынша жұмыстарында қарастырылды,онда компьютерлік лингвистиканың негізгі парадигмасы – контекстік-тәуелсіз грамматика сипатталған. Мәтіндерді терең өңдеудің алғашқы тәсілдері әдетте осындай грамматиканы қолданып, сондай-ақ талдау ағашынан ережелер жинағы мен арнайы дайындалған лексиконның көмегімен білімнің кейбір логикалық көрінісіне аудару жолымен тілді талдауға келіп жетті. Осыдан кейін логикалық көріністі білім базасына қосуға және оған түрлі операцияларды орындауға, сұрақтарға жауап беруге, бекітулерді тексеруге және т. б. әрекеттерді істеуге болады. Алайда, бұл тәсілді практикалық қолдану кезінде әлем туралы жалпы қабылданған (яғни қарапайым, базалық) білімді есепке алу қажеттілігімен байланысты қиындықтар туындады. Мысалы, сапарды жоспарлаудың зияткерлік ассистенті өте қарапайым деректерді қажет етеді. Ол түнде адамдар ұйықтайтынын түсіну керек (түнде кейбір өткелдерді жоспарламау үшін), солтүстік теңіздерде қыста адамдардың көпшілігі шомылмайтынын білу керек және т. б. Барлық осы мәліметтерді шешімдер қабылдау кезінде сақтау және ескеру қажет болды және 1984 жылы жалпы білімнің үлкен базасын құру мен қорытындыларды қалыптастырудың қолайлы тетігін әзірлеу бойынша CYC (ағылшын тілінен encyclopedia) жобасы бастау алды. Жоба әлі күнге дейін бар, бірақ шын мәнінде сәтсіздікке ұшырады – бүгінде бұл білім базасын бірнеше зерттеу жұмыстары ғана пайдаланады.

90-жылдардың басында Машиналық оқыту әдістері дами бастады және бір уақытта статистикалық лингвистика бойынша бірқатар жұмыстар жасалды. Машиналық оқытуда мәтіндерді өңдеуге байланысты әртүрлі есептерді шешу үшін жіктеу алгоритмдері өзін өте жақсы көрсетті. Спам детекциясы, тақырып бойынша құжаттарды сұрыптау, атаулы мәндерді бөлу. Компьютерлік Лингвистикада сөйлеу бөліктерін анықтау Марковтың

жасырын тізбегі және максимальды энтропия моделі сияқты статистикалық әдістердің арқасында жоғары дәлдікке ие болды. Ықтимал мән мәтіндік тәуелсіз грамматика негізінде парсерлер пайда болды, ал IBM корпорациясында статистикалық машиналық аударма бойынша ауқымды жоба өтті. Уақыт өте келе, терең оқыту (deep learning) негіздері қаланды, ол тек жақында ғана жоғары өнімді жүйелер саласындағы прогреске және оқыту үшін пайдаланылатын деректердің үлкен көлемінің пайда болуына байланысты алғашқы өсуді берді. Терең оқыту – көп деңгейлі ("терең") нейрондық желілерді үлкен көлемде қолданады. Машиналық оқыту үшін белгілерді бөлуге және бір мезгілде сол белгілер бойынша тікелей оқытуға мүмкіндік беру қажет.

2010 жылы лексикаландырылған ықтимал грамматика моделі ұсынылды, ол грамматикалық талдаудың дәлдігін 93% - ға дейін арттыруға мүмкіндік берді, бұл, әрине, идеалдан алыс. Талдау дәлдігі-бұл дұрыс құрылған грамматикалық байланыстардың пайызы және ұзын сөйлем дұрыс бөлшектелу ықтималдығы. Сонымен қатар, жаңа алгоритмдер мен тәсілдердің арқасында, терең оқытумен қоса грамматикалық талдау жылдамдығы артты. Сонымен қатар, барлық жетекші алгоритмдер мен модельдер зерттеушілердің көпшілігіне қолжетімді болды және NLP (Natural Language Processing) үшін терең оқыту саласындағы ең танымал жұмыс Томас Миколовтың алгоритмі болды [5].

Қазіргі таңда табиғи тілмен жұмыс істейтін зерттеушілерде зияткерлік жүйелерді құру үшін көптеген құралдар бар (кестені қараңыз), оларды шартты түрде үш сыныпқа бөлуге болады: жеке сөзбен жұмыс істеу әдістері, сөйлемдермен жұмыс істеу әдістері және бірнеше сөйлемдерден ерікті мәтіндерді өңдеу әдістері.

Дәстүрлі түрде сөйлемдегі сөздер сөздік элементтері ретінде өңделді. Бірақ бұл әдіс қиындықтарға тап болды, егер әр түрлі сөйлесу формаларын, мысалы, техникалық жаргонды есепке алсақ, онда сөздердің көлемі, тіпті ағылшын тілінде де үлкен және кең қолдану саласы үшін толық семантикалық сөздікті құрастыру өте қиын міндет болды. Машиналық оқытудың көптеген стандартты пакеттеріне кіретін word2vec алгоритмі үлкен белгісіздік корпустарда (әр түрлі жанрлар мен стильдерде жазылған әр түрлі тақырыптар бойынша көптеген әр түрлі мәтіндер) сөздердің сапалы түсініктерін үйретететін үлкен танымалдыққа ие болды. Сөздердің дәстүрлі көріністерінен айырмашылығы, бұл жерде тілдің нейроверо екіталай моделі қолданылады (осыдан және терең оқытумен байланыс) – әрбір сөз кішкентай (толық сөздіктің көлеміне қатысты) кеңістіктегі заттық сандардың векторы, мысалы, 300 өлшемдегі өлшеммен беріледі. Бастапқы векторларға кездейсоқ мәндер беріледі. Одан әрі оқу процесінде сөз үшін вектор таңдап алынады (скалярлық туындыға қарай осы алгоритм жағдайында), ұқсас контекстерде кездесетін басқа сөздердің векторларына барынша ұқсас. Контекст ретінде алдыңғы және кейінгі сөздердің шағын терезесі алынады, мысалы, бес сөз. Бұл өте қарапайым тәсіл қызықты нәтижелер береді. Сөз косинусына жақын жиі

семантикалық жақын сөздер орналасады. Көптеген табиғи тілді өңдеу үшін қызықты қатынастар векторларға кодталған. Атақты мысал: "Париж" сөзінің векторынан "Франция" сөзінің векторын алып, "Италия" векторын қосса, онда "Рим" векторына өте жақын вектор пайда болады. "Астана" қатынасы сөз векторларына кодталған.

Сонымен қатар, мәтіндерді өңдеу есептерінде терең модельдер сирек қолданылады – мысалы, word2vec-те терең нейрожелі , алайда бұл алгоритм терең оқыту парадигмасына салынады. Ол өзі мұғалімсіз оқыту режимінде белгілерді табады, осылайша адам миындағы оқыту процессін қайталайды. Мысалы, жаңа тілді зерттей отырып, біз бірнеше рет белгісіз сөзді кездестіреміз және алдымен оның мағынасын білмейміз, бірақ содан кейін оны қолдану мәнмәтінінен ұғынуды бастаймыз. Бұл ретте біз сөздің қатаң анықтамасын бере алмаймыз, ал жақындық пен ұқсастық интуитивті ұғымдарға сүйенеміз.

Word2vec оқитын қосымша қарым – қатынастар (белгілер) мәтіндерді өңдеу үшін пайдалы болуы мүмкін, бірақ өкінішке орай, оқытудан кейін векторларда қандай қарым – қатынастар бар және олар қаншалықты сенімді кодталғанын олар қатынастардың барлық мәні үшін орындала ма, түсініксіз. Рас, сөздердің векторлық түсініктерін онтологиялармен толықтыруға мүмкіндік беретін әдістер бар. Мысалы, мақалада зерттеушілер кез келген қатынастар мен таксономиялар векторлық көріністерге сенімді кодтайтын векторларды оқыту әдісін ұсынды.

Стандартты word2vec алгоритмі омонимиямен байланысты мәселелерді шешуге мүмкіндік бермейді. Мысалы, ағылшын тілінде шамамен 40% - да бір жазу кезінде әртүрлі мағынасы бар омонимді сөздер қолданылады ("машина": автомобиль, компьютер, механикалық құрылғы). Бұл табиғи тілді өңдеудегі өте күрделі міндет, оны шешу үшін омонимдерді автоматты түрде анықтау және жеке мағыналар үшін жеке векторларды құру мақсатында word2vec алгоритмінің модификациясы, сондай-ақ омонимнің берілген контекстке қатысты дұрыс мағынасын анықтау процедурасы ұсынылған. Ресей зерттеушілері орындаған жұмыста омонимикалық болуы мүмкін сөздер үшін векторларды оқыту әдісі сипатталған. Басқа әдістерден айырмашылығы, бұл алгоритм әр түрлі омонимдік сөздердің мағыналары бойынша оқытылады.

Табиғи тілді өңдеудің көптеген әдістері тек синтаксистік құрылымнан шығаруға болатын синтаксис пен семантиканы елемей, сөздердің түсініктерін ғана табысты пайдаланады. Мәтіндерді ұсынудың мұндай үлгісі "bag of words" деп аталады. Сөздердің тәртібін ескермейтін қарапайым сөздер жиынтығынан тұрады. Мысалы, векторлық көріністер жағдайында модель жаттығатын корпус сөздерінің векторларын кластерлерге біріктіруге және мұндай кластерлерді қарапайым жіктеу міндеттері үшін пайдалануға болады. Егер тапсырма сапалы семантикалық көріністерді алудан тұрса, онда сөйлемнің синтаксистік құрылымымен жұмыс істейтін мәтіндерді өңдеу құралдары қажет болады. Мысалы отель туралы клиенттер қалдырған пікірді талдау кезінде келесі үлгіні кездестіруге болады: "Отель жағымды, бірақ жуынатын

бөлмесі кішкентай". Ұсыныстың құрылымын талдаусыз, біз әрбір сын есімнің қай сөзге жататынын түсіне алмаймыз.

Грамматикалық талдау кезінде әлі де нақты мәселелер туындайды. Біріншіден, мұнда көп сөйлеу бөліктерін тану сапасына байланысты, ол өте жоғары болуы тиіс (97-98%). Алайда ұзын сөйлемдерде сөздің дұрыс емес белгілі бір бөлігін кездестіруге болады, бұл талдау қателіктеріне әкеледі. Екіншіден, грамматикалық талдау шамамен 90-93% дәлдігін береді (дұрыс анықталған қарым-қатынас пайызы), ал бұл, өз кезегінде, ұзақ сөйлемде әрдайым талдау қателері болады дегенді білдіреді. Мысалы, талдау дәлдігі 90% болғанда, бірыңғай қатесіз сөйлемнің ұзындығы 10 сөзді талдау ықтималдығы небәрі 35% құрайды. Жақын арада талдау сапасын жақсартуға үміт бар, бірақ жиі дұрыс грамматикалық талдау, сөйлем семантикасын түсінуді білдіреді. Ағылшын тілінде бұл жиі қиындық тудырады. Мысалы, "i saw a man with a hammer" сөйлемінде біз адамды балға арқылы көрдік немесе балға бар адамды көрдік деп санауға болады. Яғни бұл сөйлемнің екі түрлі грамматикалық талдауы болуы мүмкін. Әрине, егер ең дәл грамматикалық талдау қажет болса, онда бірнеше ықтимал нұсқаларды қалдыру, содан кейін әртүрлі факторлардың жиынтығы бойынша дұрыс нұсқаны анықтау қажет.

Терең оқыту әдістері сөйлемдермен жұмыс істеуге және Wodr2vec класстарынан алынған векторлардың реттілігі арқылы сөйлемді модельдеу және оны машинамен оқыту алгоритмдерінде қолдану мүмкіндігін береді. Машиналық оқытудың стандартты алгоритмдері атрибуттардың бекітілген жиынтығымен жұмыс істейді және оларды осындай модельге бейімдеуге болмайды. Бірақ онымен рекуррентік нейрондық желілер өте жақсы жұмыс істейді. Олар кіруде бір сөзді векторлық көрініс қабылдайды және бірнеше ішкі деңгейлерге бөледі, ал шығуда классификатор немесе регрессор құрылады. Нейрожелілерден айырмашылығы, рекуррентік желінің ішкі деңгейлері (кейде жоғарғы деңгей) желіге кері қосылған, яғни алдыңғы сөзге ауысқан желі күйі кейінгі сөзге қосымша кіру ретінде желіге берілетін болады. Осылайша, нейрожеліде "жады" пайда болады, ол оған сөйлемнен сөздерді дәйекті өңдеуге және әрбір сөзге немесе барлық сөйлемге қатысты бірден жеке болжам жасауға мүмкіндік береді. Мысалы, желі бір ретпен ұсыныс береді, ал желі ағымдағы қадамды анықтау үшін өзінің алдыңғы күйін пайдаланады. Бірақ практикада қарапайым рекуррентті желілер алдыңғы сөздер туралы жады желінің жаттықтыру және пайдалану кезінде тез жоғалатын болғандықтан өте жақсы жұмыс істемейді. Сондықтан, әдетте арнайы жады элементтері – LTSM (Long Term Short Memory) қолданылады, ол жадты жазу, оқу және тазалау кезінде анықтайды көптеген нейрондар мен басқарушы элементтерді білдіреді. Бұл элементтер жадқа ұзақ ретпен есептеу кезінде өзгермеуге мүмкіндік береді және оқыту кезінде қатені дұрыс атрибуттауға мүмкіндік береді. Басқару элементтері де нейрожеліні оқыту процесінде оқиды.

Рекуррентті нейрожелілер, әсіресе LTSM тілді үлгілеу, машиналық аударма және басқада әр түрлі міндеттерді шешу кезінде өзін жақсы көрсете

білді, бірақ желілердің бұл класында айтарлықтай кемшілік бар - олар сөйлемдегі сөздердің тәртібін ғана пайдаланады және оларды дәстүрлі аспаптармен алынған, грамматикалық құрылымдармен жұмыс істеуге мәжбүрлеуге болмайды. Шын мәнінде, рекуррентті желілерді әр міндет үшін нөлден бастап тіл грамматикасын "үйрету" керек. Сонымен қатар, рекурренттік желі аралық фразалар үшін түсінік құрмайды, сондықтан сөйлемдерді құрайтын әр түрлі фразалардың сапалы ұсынымдары қажет болатын есептер үшін рекурсивті нейрондық желілер қолданылады.

Рекуррентті, рекурсивті желілер сөйлемдегі сөздердің бірізділігінің үстінен емес, сөйлемнің тәуелділік грамматикасының негізінде әр сөйлем үшін оны талдау үшін екілік ағаш құрады. Рекурсивті желінің жұмысын келесідей елестетуге болады. Алдымен ол талдау ағашының жапырақтарын өңдейді (ағаштың жапырақтары-сөйлемнің екі сөзіне және олардың арасындағы грамматикалық тәуелділіктің түріне сілтегіштер), алынған жапырақтарды сөз векторы сияқты өлшемдік вектормен алмастырады. Әрі қарай жұмыс істеуді жалғастырады, бірақ енді жапырақтар сөзбен емес, фразалармен біріктіріледі, сөйлемнің векторлық түсініктері құрылады. Сонымен, талдау ағашы бар, ағаштың әрбір түйінін нейрондық желіге ауыстыра отырып, біздің ағаш сияқты топологиясы бар рекурсивті желі салуға болады. Әрине, барлық көбейтілген желілердің жалпы параметрлері бар, яғни оқыту және пайдалану кезінде біз бір желімен жұмыс істейміз. Классификация немесе регрессия түріндегі болжамдар жоғарғы торапты қоса алғанда, көбею желісінің кез келген түйіндерінің үстінен болуы мүмкін.

Оқыту кезінде рекурсивті желі тек толық сөйлемдер үшін ғана емес, сонымен қатар барлық сөйлемдер үшін де сапалы түсініктерді жасауды үйренуі мүмкін. Бұл ретте нейрожелі грамматикалық талдау қателіктерінің әсерін әлсіретуі мүмкін, әсіресе рекурсивті нейрожелі үйренетін міндетке әсер етеді. Осылайша, біз сөздерге де, сөйлемдегі барлық сөздерге де семантикалық жақындық өлшемін аламыз. Сонымен қатар, рекурсивті нейрондық желіге LTSM жады элементтерін қосуға және өте сапалы векторлық көріністерді алуға болады [6].

Ұсыныстардың векторларын алу үшін басқа тәсіл әрбір сөйлем, параграф немесе тұтас құжат үшін сөйлемнің немесе параграфтың әрбір сөзінің контекстін болжауға қатысатын жеке вектор жаттығудан және оқыту процесінде алдын ала болжамдарды барынша жақсартатын векторлар таңдалады. Алынған векторлардың сапасы бойынша бұл әдіс (оны әдетте doc2vec деп атайды) рекурсивті нейрожелілермен ұштасады, бұл ретте оқыту үшін белгіленген оқыту үлгісі қажет емес. Рас, бұл әдістің екі маңызды кемшілігі бар: оған үлкен ұсыныстар немесе тұтас параграфтар қажет, ол қысқа фразалар деңгейінде жұмыс істемейді; ол нейрожеліден гөрі қымбат,- ұсыныстың әрбір векторы жеке оңтайландырылады.

Сөздер мен сөйлемдерді үлгілеудің тағы екі тәсілі – сөздердің символдық көріністерімен немесе аралас көріністермен жұмыс істейтін орама нейрожелілер туралы атап өткен жөн. Әдетте түйме нейрожеліге кіруде

барлық сөйлемдер бөлек сөздердің векторлық көріністерінің матрицасы түрінде беріледі. Орама желісі кіріс мәліметтерінің үстіңгі жағында терезелердің біріздігінде қолданылатын тіркелген өлшемнің қосалқы желілерінің еркін ұзын тізбегін өңдейді. Осылайша, ұю операциясы эмуляцияланады және желі ұюда қолданылатын сүзгішке оқытылады. Ұю желілері рекурсивті желілер деңгейінде жақсы нәтижелер көрсетеді.

Қызықты нәтижелерді сөздердің символдық көрінісінен жоғары жұмыс істейтін нейрожелілер, сондай-ақ аралас модельдер көрсетті. Мысалы, сөйлеу бөліктерін анықтау тапсырмасында символды нейрожелілер басқа әдістерден асып түсті,бұл ретте оқыту үшін әртүрлі белгілердің қыңыр құру талап етілмейді. Бұл желілер мәліметтердің сиретілу мәселесін шешуге көмектесіп, сөздердің векторлық түсініктерін толықтыру үшін табысты қолданылады. Символдық ұғымдары бар нейрожелілер тек белгілі бір сөзбен жұмыс істеу үшін ғана емес, сөз морфологиясын пайдалануды үйренеді.

Рекуррентті және рекурсивті нейротейлердің көмегімен мәтіндерді автоматты түрде өңдеуге байланысты қарапайым есептерді тиімді шешуге болады: жіктеу, тоналдықты анықтау, атаулы мәндерді және қарапайым фактілерді бөлу және т. б.

Бірнеше сөйлемнен тұратын мәтіндерді өңдеу туралы сөз келгенде, оларды тәуелсіз мәндер ретінде емес, пікірлердің өзара байланысқан тізбегі ретінде қарастыруды талап етеді, барлық қолда бар технологиялар елеулі мәселелерге тап болады. Бұл жағдайда келесі ұсыныстармен байытылатын және түрлендірілетін семантикалық контекст пайда болады және оны модельдеу өте қиын. Компьютерлік Лингвистикада қарапайым және жақсы формальды міндет бар кореференттілікті немесе анафорикалық қатынастарды шешу (мысалы, "кітап үстелде жатыр. "Кітап" және "ол" сөздері бірдей мағынаға сілтеме жасайды). Өкінішке орай, бүгінгі күні бұл міндет шешілмеген, ал қолда бар әдістер әлі де қанағаттанарлықсыз нәтиже береді. Алайда, егер қолданбалы аймақты қатты шектесе, онда бірнеше ұсыныстар мен диалогтар тұратын мәтіндер үшін өте қолайлы семантикалық модельдерді құруға болады.

Терең оқытудың жаңа әдістерінің арқасында бүгінгі таңда сөздерге және ұсыныстарға сапалы семантикалық түсінік алуға болады. Қазір өз семантикалық сөздіктер мен білім базаларын құру үшін аз күш қажет, сондықтан мәтіндерді автоматты өңдеу жүйесін әзірлеу оңай болды. Алайда ұсыныстар немесе бейнелер, сондай-ақ диалогтар дәйектілігі түрінде ұсынылған өзара байланысты оқиғаларды түсіну міндетін толыққанды шешуден әлі де алыс. Қазіргі таңда белгілі барлық әдістер тілді түсінудің міндеттерін шешуде немесе пәндік саланы елеулі шектеуде табысты жұмыс істейді. Мысалы, torpater компаниясында мәтіндерді автоматты түрде өңдеуге арналған әр түрлі құралдардың көмегімен электрондық коммерция объектілері туралы пікірлердегі бағалау пайымдауларын анықтау бойынша тапсырма шешіледі. ConceptNet және FrameNet сияқты кең қамтылған семантикалық сөздіктерді құрудағы прогресті атап өткен жөн, сондай-ақ мәтіннен

сөздіктерге сөздерді тіркестіруді жүзеге асыратын машиналық оқыту әдістері жүзеге асырады. Рас, әртүрлі семантикалық рөлдерді қамтитын FrameNet сөздік жағдайында автоматты байланыстыру сапасы әлі де төмен, дәлдігі 60% аспайды.

Data Mining өндіру немесе деректерді қазудеп аударылады. Жиі Data Mining жанында "деректер базасындағы білімді анықтау" (knowledge discovery in databases) және "деректерді зияткерлік талдау" деген сөздер кездеседі. Оларды Data Mining синонимдері деп санауға болады. Барлық көрсетілген терминдердің пайда болуы деректерді өңдеу құралдары мен әдістерінің дамуындағы жаңа ораммен байланысты.

90-шы жылдардың басына дейін осы саладағы жағдайды қайта болжауға ерекше қажеттілік болған жоқ. Барлығы қолданбалы статистика деп аталатын бағыт шеңберінде өз кезектерімен жүрді. Теоретиктер аналитикаға байланысты конференциялар мен семинарлар өткізді, жинақтармен мол мақалалар мен монографиялар жазды.Сонымен қатар, практиктер теориялық экзерсистерді қолдану талпыныстары көп жағдайда нақты міндеттерді шешу үшін бедеулік болып табылатындығын әрдайым білді. Бірақ практиктердің біраз уақытқа дейін жүргізген ізденістеріне ерекше назар аудармауға болады. Олар негізінен шағын жергілікті деректер базасын өңдеудің жеке мәселелерін шешті.

Уақыт өте келе, адамдарға деректерді жазу және сақтау технологияларының жетілдірілуіне байланысты әртүрлі облыстарда ақпараттық кендердің орасан зор ағындары құлады. Кез келген кәсіпорынның қызметі (коммерциялық, өндірістік, медициналық, ғылыми және т.б.) енді оның қызметінің барлық егжей-тегжейін тіркеумен және жазумен сүйемелденеді. Бұл ақпаратпен не істеу керек" деген сөз " шикі деректер ағындары өнімді өңдеусіз ешкімге қажет емес қоқыс үйіндісін құратыны анық болды.

Мұндай қайта өңдеуге қойылатын қазіргі заманғы талаптардың ерекшелігі мынадай:

1) деректердің шексіз көлемі бар;
2) деректер әртүрлі (сандық, сапалық, мәтіндік) болып табылады;
3) нәтижелер нақты және түсінікті болуы тиіс;
4) шикі деректерді өңдеу құралдары пайдалану оңай болуы тиіс.

Ұзақ уақыт бойы Деректерді талдаудың негізгі құралы рөліне үміткер дәстүрлі математикалық статистика пайда болған мәселелердің алдында ашық түрде жүзеге асырылды. Басты себеп-фиктивті шамаларды операцияларға әкелетін іріктеме бойынша орташалану тұжырымдамасы (аурухана бойынша пациенттердің орташа температурасының типі, сарайлар мен лачугалар мен т.б. тұратын көшедегі үйдің орташа биіктігі). Математикалық статистика әдістері, ең бастысы, алдын ала тұжырымдалған гипотезаларды тексеру үшін (verification-driven data mining) және деректерді жедел аналитикалық өңдеудің негізін құрайтын "өрескел" барлау талдауы үшін (online analytical processing, OLAP) пайдалы болды.

Data Mining (discovery-driven data mining) заманауи технологиясының негізіне деректердегі көпаспективті өзара қарым-қатынастардың фрагменттерін көрсететін шаблондар (паттерндер) концепциясы алынған. Бұл шаблондар адамға түсінікті түрде ықшам көрсетілуі мүмкін деректерді іріктеуге тән заңдылықтар болып табылады. Шаблондарды іздеу талданатын көрсеткіштердің мәндерін іріктеу құрылымы және бөлу түрі туралы априорлық болжамдардың шектерімен шектелмеген әдістермен жүргізіледі.

Data Mining әдістері: ассоциация, бірізділік, жіктеу, кластерлеу және болжаумен байланысты. Егер оқиға байланысты тізбектен тұрса, онда дәйектілік туралы айтуға болады. Мысалы, үйді сатып алғаннан кейін бір айдың ішінде 45% жаңа ас үй плитасы да сатып алынады, ал екі апта ішінде 60% жаңа қоныстанушылар тоңазытқышпен жабдықталады [7].

Жіктеме көмегімен белгілі бір объект тиесілі топты сипаттайтын белгілер анықталады. Бұл жіктелген объектілерді талдау және кейбір ережелер жиынтығын қалыптастыру арқылы жасалады.

Кластерлеу топтан өзгеше, топтың өздері алдын ала қойылмаған кезде қолданылады. Кластерлеу арқылы Data Mining құралдары әртүрлі біртекті деректер топтарын өз бетінше бөледі.

Болжаудың барлық мүмкін болатын жүйелерінің негізі уақытша қатар түрінде мәліметтер базасында сақталатын тарихи ақпарат болып табылады. Егер мақсатты көрсеткіштердің мінез-құлқының динамикасын барабар көрсететін үлгіні табу мүмкін болса, олардың көмегімен болашақта мінез-құлқын болжауға болатын жүйені әзірлеуге ықтималдық бар.

Data Mining қолданбалы статистиканың жетістіктері, бейнелерді тану, жасанды интеллект әдістері, деректер базасының теориясы және т.б. негізінде пайда болған және дамып келе жатқан мультидисциплинарлық сала болып табылады. Сондықтан Data Mining түрлі қолданыстағы жүйелерінде іске асырылған әдістер мен алгоритмдердің көптігі. Мұндай жүйелердің көпшілігі бірден бірнеше тәсілдерді біріктіреді. Дегенмен, әдетте, әрбір жүйеде басты ставка жасалатын қандай да бір негізгі компонент бар. Төменде көрсетілген негізгі компоненттердің жұмыс негізінде жіктелуі келтірілген. Бөлінген класстарға қысқаша сипаттама беріледі.

Пәндік-бағытталған аналитикалық жүйелер өте әртүрлі. Қаржы нарықтарын зерттеу саласында кең таралған осындай жүйелердің ішкі классы "техникалық талдау" деп аталады. Ол баға динамикасын болжаудың және нарық динамикасының әртүрлі эмпирикалық үлгілеріне негізделген инвестициялық портфельдің оңтайлы құрылымын таңдаудың бірнеше ондық әдістерінің жиынтығын білдіреді. Бұл әдістер жиі қарапайым статистикалық аппаратты пайдаланады, бірақ өз саласының қалыптасқан ерекшелігін (кәсіби тіл, әр түрлі индекстердің жүйелері және т.б.) барынша ескереді. Нарықта осы класстың көптеген бағдарламалары бар.

Барлық белгілі статистикалық пакеттердің соңғы нұсқалары дәстүрлі статистикалық әдістермен қатар Data Mining элементтерін қамтиды. Бірақ олардың негізгі назары әлі де классикалық әдістемелерге – корреляциялық,

регрессиялық, факторлық талдауға және т.б. бөлінеді. Статистикалық талдау үшін пакеттердің ең жаңа егжей-тегжейлі шолуы ЦЭМИ беттерінде келтірілген. Бұл класс жүйелерінің жетіспеушілігі пайдаланушының арнайы дайындығына қойылатын талап деп саналады. Сондай-ақ, қуатты заманауи статистикалық пакеттер қаржы мен бизнесте жаппай қолдану үшін тым ауыр салмақты болып табылады.

Data Mining-та олардың қолданылуын шектейтін статистикалық пакеттердің айтарлықтай принципті кемшілігі бар. Пакеттердің құрамына кіретін көптеген әдістер статистикалық парадигмаға сүйенеді, онда басты фигуранттар іріктеменің орташаланған сипаттамалары қызмет етеді. Бұл сипаттамалар, жоғарыда көрсетілгендей, нақты күрделі өмірлік феномендерді зерттеу кезінде жиі фиктивті шамалар болып табылады.

Ең қуатты және кең таралған статистикалық пакеттердің мысалдары ретінде SAS (SAS Institute компаниясы), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA және т.б. атауға болады.

Бұл архитектурасы нейрондардан жүйке құрумен ұқсас жүйелердің үлкен класы (қазір белгілі болғандай, әлсіз). Ең көп таралған архитектуралардың бірінде, қателіктің кері таралуымен көп қабатты перцептронда иерархиялық желі құрамындағы нейрондардың жұмысы имитацияланады, онда жоғары деңгейдегі әрбір нейрон төменде жатқан қабаттағы нейрондарының шығуымен қосылған. Ең төменгі қабаттың нейрондарында кіріс параметрлерінің мәндері беріледі, олардың негізінде қандай да бір шешім қабылдау, жағдайдың дамуын болжау және т. б. Бұл мәндер келесі қабатқа берілетін сигналдар ретінде қарастырылады, сандық мәндерге (таразыға), нөмірараліқ байланыстарға тәуелді әлсірейді немесе күшейді. Нәтижесінде нейронның ең жоғарғы қабатының шығуында кейбір мән өндіріліп, жауап ретінде қарастырылады. Барлық желінің кіріс параметрлерінің енгізілген мәндеріне реакциясы. Сонымен қатар, ол үшін ең алдымен кіріс параметрлерінің мәндері және оларға дұрыс жауаптар белгілі бұрын алынған деректердің "өшірілуі" қажет. Жүзеге асырылатын процесс желі жауаптарының белгілі дұрыс жауаптарға барынша жақындығын қамтамасыз ететін, желі аралық байланыстардың таразысын іріктеуден тұрады.

Нейрожелілік парадигманың негізгі кемшілігі оқыту үлгісінің өте үлкен көлеміне ие болу қажеттілігі болып табылады. Тағы бір маңызды кемшілік-бұл тіпті дренаждалған нейрондық желі - қара жәшік. Бірнеше жүздеген еларалық байланыстардың салмағы ретінде тіркелген білім талдап, интерпретацияланбайды.

Case based reasoning – CBR жүйелерінің идеясы бір қарағанда өте қарапайым. Болашаққа болжам жасау немесе дұрыс шешімді таңдау үшін, бұл жүйелер бұрынғы қолма-қол жағдайдың жақын аналогтарын табады және олар үшін дұрыс болған жауапты таңдайды. Сондықтан бұл әдіс "жақын көрші" (nearest neighbour) әдісі деп аталады. Соңғы уақытта memory based

reasoning термині де тарату алды, ол шешім жадында жинақталған барлық ақпараттың негізінде қабылданады.

CBR жүйелері әртүрлі есептерде жақсы нәтижелерді көрсетеді. Олар ең бастысы алдыңғы тәжірибені жинақтайтын қандай да бір модельдер немесе ережелер жасамайды , шешімді таңдауда олар қол жетімді тарихи деректердің барлық массивіне негізделеді, сондықтан CBR жүйесінің нақты қандай факторларының негізінде өз жауаптарын құратынын айту мүмкін емес.

Басқа минус "жақындық" шарасын таңдау кезінде CBR жүйесі ерікті түрде рұқсат етуі. Бұл өлшемге қанағаттанарлық жіктемеге немесе болжамға жету үшін жадыда сақтау қажет көптеген прецеденттердің көлемі байланысты.CBR пайдаланатын жүйелердің мысалдары – Kate tools (Acknosoft, Франция), Pattern Recognition Workbench (Unica, АҚШ).

Ағаштар шешімі ең танымал тәсілдердің бірі шешу есептер Data Mining. Олар ағаш түрі бар "егер...." (if-then) типті жіктеуші ережелердің иерархиялық құрылымын жасайды. Шешім қабылдау үшін кейбір объектіні немесе жағдайды қандай классқа жатқызуға болады, оның тамырынан бастап осы ағаш тораптарында тұрған сұрақтарға жауап беру талап етіледі. Егер жауап оң болса, келесі деңгейдің оң торабына өту жүзеге асырылады, егер теріс болса – сол торапқа; содан кейін тиісті торапқа байланысты сұрақ қайтадан жасалады.

Тәсілдің танымалдығы көрнекілігі мен түсініктікліктіне байланысты. Бірақ шешім ағаштары деректерде "үздік" (ең толық және дәл) ережелерді таба алмайды. Олар белгілерін жүйелі түрде қарап шығудың классикалық принципін жүзеге асырады және логикалық тұжырымның иллюзиясын ғана жасай отырып, нақты заңдылықтардың сынықтарын тізбектейді.

Сонымен қатар, көптеген жүйелер осы әдісті қолданады. Ең танымал See5/C5.АҚШ), KnowledgeSeeker (ANGOSS, Канада), Clementine (Integral Solutions, Ұлыбритания), Sipina (University of Lyon, Франция), IDIS (Information Discovery, АҚШ), KnowledgeSeeker (ANGOSS, Канада). Бұл жүйелердің құны 1-ден 10 мың АҚШ долларына дейін өзгереді [8].

Data Mining нарығында жалпы мойындалған PolyAnalyst – ресейлік әзірлеу жүйесінің мысалында осы тәсілдің қазіргі заманғы жағдайын бейнелеуге болады. Бұл жүйеде мақсатты айнымалылардың басқа айнымалылардан тәуелділігінің түрі туралы гипотезасы қолданылады. Бағдарламаларды құру процесі эволюция принципі секілді құрылады(бұл тәсіл генетикалық алгоритмдерге сәл ұқсас). Жүйе аса немесе аз қанағаттанарлық көрінетін тәуелділікті тапқанда, оған шағын модификациялар енгізе бастайды және салынған еншілес бағдарламалардың ішінен дәлдікті арттыратындарды іріктейді. Осылайша, жүйе бірнеше генетикалық бағдарламалар желісін "өсіреді", олар бір-бірімен ізделіп тәуелділікті білдіру дәлдігінде бәсекелеседі. PolyAnalyst жүйесінің арнайы модулі табылған тәуелділікті жүйенің ішкі тілінен пайдаланушыға түсінікті тілге (математикалық формулалар, кестелер және т.б.) аударады.

Эволюциялық бағдарламалаудың басқа бағыты белгілі бір түрдегі функциялардың пішінінде мақсатты айнымалылардың тәуелділігін

іздестірумен байланысты. Мысалы, осы типтегі ең сәтті алгоритмдердің бірінде – аргументтерді топтық есепке алу әдісі (МГУА) тәуелділік полином түрінде іздейді. Қазіргі уақытта Ресейде сатылатын МГУА жүйелерінің ішінен Ward Systems Group компаниясының NeuroShell жүйесінде іске асырылған.

Data Mining генетикалық алгоритмдерді қолданудың негізгі саласы емес. Оларды әртүрлі комбинаторлық есептер мен оңтайландыру есептерін шешудің қуатты құралы ретінде қарастыру қажет. Дегенмен, генетикалық алгоритмдер қазір Data Mining әдістерінің стандартты құралдарына кірді, сондықтан олар осы шолуға енгізілді.

Деректерді сақтау қоймаларын құру кезінде оған түсетін ақпаратты тазалауға аз көңіл бөлінеді. Сақтау көлемі неғұрлым көп болса, соғұрлым жақсы. Бұл ақаулы тәжірибе және ең жақсы тәсілі-деректерді қоқыс үйіндісіне айналдыру. Деректерді тазалау қажет. Өйткені, ақпарат әртүрлі форматта және әр түрлі дерек көздерден жиналады. Ақпаратты жинаудан кейін көптеген тыныс белгілерінің, смайликтардың болуы тазалау процесін әсіресе өзекті етеді.

Үлкен есеппен жұмыс істегенде қателер әрқашан жіберіледі және олардан толық құтылу мүмкін емес. Осы дәлсіздіктерді шешуге қандай уақыт бөлу қажет деген сұрақ туындайды. Жалпы, кез келген тәсілмен қателердің санын қолайлы деңгейге дейін азайтуға ұмтылу керек. Талдау үшін қолданылатын әдістер дәл емес мәліметтерді жинаса, үлкен қиыншылықтар туады. Сонымен қатар, мәселенің психологиялық аспектісін ескеру қажет. Егер талдаушы деректер қоймасынан алынған цифрларға сенімді болмаса, оларды пайдаланбауға тырысады және басқа көздерден алынған мәліметтерді пайдаланады. Сонда мұндай қойма не үшін қажет?

Үлгілердің сәйкессіздігі сияқты қателерді қарастырмаймыз,енгізу форматтары мен кодтау айырмашылықтары. Яғни, ақпарат әртүрлі көздерден келіп түсетін мәлімет, онда бір фактіні белгілеу үшін түрлі келісімдер қабылдануы қажет. Мұндай қателіктің тән мысалы – адамның жынысын белгілеу. Бір жерде ол М/Ж ретінде белгіленеді, бір жерде 1/0, бір жерде True/False. Мұндай қатемен қайта кодтау және типтерді келтіру ережелерінің тапсырмасы арқылы күреседі. Мұндай мәселелер қазіргі уақытта толықтай шешілмеді. Осындай қарапайым тәсілдермен шешілмейтін жоғары тәртіптегі мәселелер қызықтырады.

Мұндай қателердің нұсқалары өте көп. Сонымен қатар, белгілі бір пәндік салаға немесе міндеттерге ғана тән қателер бар. Бірақ міндетке тәуелді емес қарастырайық:

1) Ақпараттың қарама-қайшылығы;
2) рұқсатнамалар осы;
3) аномальды мәндер;
4) шу;
5) деректерді енгізу қателері.

Осы мәселелердің әрқайсысын шешу үшін ойластырылған әдістер бар. Әрине, қателерді қолмен басқаруға болады, бірақ үлкен көлемде деректерді

өңдеу өте қиын болады. Сондықтан, адамның ең аз қатысуымен автоматты режимде осы тапсырмаларды шешу нұсқаларын қарастыру қажет.

Алдымен қарама-қайшылық деп санауды шешу керек. Бір қызығы, бұл міндет ерекше емес. Қарама-қайшылық деп саналатынды анықтау үшін әрекеттердің бірнеше нұсқасы бар.

Бірнеше қарама-қайшы жазбалар табылған кезде, оларды жою. Әдіс қарапайым, сондықтан оңай іске асырылады. Кейде бұл жеткілікті. Бұл жерде асықпау маңызды.

Қарама-қайшы деректерді түзету. Әрбір қарама-қайшы оқиғалардың пайда болу ықтималдығын есептеп, ең ықтималдығын таңдауға болады. Бұл қайшылықтармен жұмыс істеудің ең сауатты және дұрыс әдісі.

Қарама-қайшылықпен күрес өте маңызды мәселе. Бұл жалпы, көптеген деректер сақтау базалары үшін актуалды мәселе. Болжамдау әдістерінің көпшілігі деректер біркелкі тұрақты ағынға түседі деген болжамнан туындайды. Іс жүзінде бұл өте сирек кездеседі. Сондықтан деректерді сақтау базаларын қолданудың талап етілетін салаларының бірі болжау іске асырылған сапасыз немесе елеулі шектеулермен көрсетіледі. Осы құбылыспен күресу үшін келесі әдістерді қолдануға болады.

Аппроксимация. Яғни, егер қандай да бір нүктеде деректер болмаса, оның төңірегін деректерді алып және белгілі формулалар бойынша осы нүктедегі мәнді есептеп, қоймаға тиісті жазбаны қосады. Бұл реттелген деректер үшін жақсы жұмыс істейді. Мысалы, күнделікті өнімді сату туралы ақпарат.

Ең шынайы мәнді анықтау. Ол үшін нүктенің маңайы емес, барлық деректер алынады. Бұл әдіс реттелмеген ақпарат үшін қолданылады, яғни зерттелетін нүкте төңірегіндегі не екенін анықтай алмайтын жағдай.

Жалпы картинадан қатты шығып кеткен оқиғалар жиі кездеседі. Және мұндай мәндерді түзету дұрыс. Бұл болжау құралдары процестердің табиғаты туралы ештеңе білмейді. Сондықтан кез келген аномалия қалыпты мән ретінде қабылданады. Осының салдарынан болашақ бейнесі қатты бұрмаланады. Кейбір кездейсоқ сәтсіздік немесе табыс заңдылық болып саналады.

Бұл шабуылға қарсы күрес әдісі бар робасты бағалау. Бұл әдістер күшті наразылыққа төзімді. қолда бар деректерді рұқсат етілген шекаралардан шығатын нәрселерге бағалау қажет.

Талдауда әрдайым шуға тап боламыз. Шу ешқандай пайдалы ақпарат бермейді, тек суретті анық көруге кедергі келтіреді. Бұл құбылыспен күресудің бірнеше әдістері бар.

Спектрлік талдау. Оның көмегімен жоғары жиілікті деректерді бөле аламыз. Қарапайым айтқанда, бұл негізгі сигналға қатысты жиі және елеусіз тербелістер. Спектрдің енін өзгерте отырып, шудың қандай түрін таңдауға болады.

Авторегрессионды әдістер. Бұл өте кең таралған әдіс уақытша қатарларды талдау кезінде белсенді қолданылады және Шу процесін

сипаттайтын функцияның табылуына әкеледі. Осыдан кейін шуды жойып, негізгі сигнал қалдыра аласыз.

Жалпы, бұл жеке жұмыс тақырыбы, себебі мұндай қателердің түрлерінің саны тым үлкен, мысалы, қате, деректердің саналы бұрмалануы, форматтардың сәйкес келмеуі, және бұл мәліметтерді енгізу бойынша қосымшаның ерекшеліктерімен байланысты типтік қателерді есептемегенде. Олардың көпшілігімен күресу үшін қолданылған әдістер бар. Кейбір нәрселер анық, мысалы, деректерді қоймаға енгізер алдында пішімдерді тексеруге болады. Кейбір нәзік. Мысалы, әр түрлі тезаурустар негізінде қателерді түзетуге болады. Бірақ, кез келген жағдайда, мұндай қатеден тазарту керек.

Лас деректер өте үлкен мәселе болып табылады. Іс жүзінде олар деректер қоймасын құру бойынша барлық күш-жігерді енгізе алады. Сонымен қатар, әңгіме бір реттік операция емес, осы бағыттағы тұрақты жұмыс туралы болып отыр. Таза жерде емес, тазаланған жерде. Ең жақсы нұсқа-қоймаға түсетін барлық деректер өтетін шлюзді құру.

Жоғарыда сипатталған мәселелерді шешу нұсқалары жалғыз емес. Сараптамалық жүйелерден бастап, нейрожелімен аяқталатын өңдеудің көптеген әдістері бар. Осы технологиялар көптеген іске асырылған freeware компоненттер мен бағдарламалар. Ең бастысы, оларды сауатты пайдалана алу. Тазалау әдістері пәндік аймаққа қатты байланысты екенін ескеру қажет. Деректер қоймасын ұйымдастыру қызметі мен мақсаты іс жүзінде барлығы байланысты. Біреу үшін шудан тазарту өте құнды ақпарат алуға мүмкіндік береді. Егер де тапсырма туралы априорлы ақпарат болса, онда деректерді тазарту сапасын тәртіпке арттыруға болады. Сонымен қатар, бұл шлюзді қолда бар деректер көздерімен сапалы біріктіру қажет.

Сүзгілеу механизмдері OLAP сияқты деректерді сақтау қоймасының ажырамас атрибуты болуы тиіс. Әйтпесе, жиналған қоқыстың тауында пайдалы деректер табу мүмкін емес. Оның көлемі ұлғайған сайын пайдаланушылар міндетті түрде бірдей пікірге келеді.

2.4 Тоналдылықты талдау

Табиғи тілдің өңделуі туралы бүгінгі күні көп айтады – бұл концепция жасанды интеллектті одан әрі дамыту үшін негіз қалаушы болып саналатын ғылыми ортада ғана емес, сонымен қатар біздің әріптестеріміз, студенттер және IT-индустриясындағы қазіргі заманғы жағдайға қызығушылық танытатын адамдар да бар.

Бұл ең қызықты және танымал ғылыми бағыттардың ішінде sentiment analysis атауына ие болды, ғылыми қазақ тіліне аударғанда "мәтіндердің тоналдығын талдау" дегенді білдіреді. Мәтіннің үнсіздігін талдау – бұл мәтіннің эмоционалды боялған лексикасын автоматты түрде анықтауға арналған контент-талдау әдістерінің класы, сондай-ақ мәтінде сөз болып отырған объектілер бойынша автордың пікірлері [9].

Мәтіннің эмоциялық бағалауын Автоматты анықтаудың қазіргі жүйелерінде жиі бір өлшемді эмотивтік кеңістік қолданылады. Позитив

немесе негатив (жақсы немесе жаман) жіктелуі қолданылады. Алайда, көп өлшемді кеңістіктерді пайдаланудың табысты жағдайлары белгілі.

Үндестік талдаудағы негізгі міндет осы құжаттың полярлығын жіктеу, яғни құжатта немесе ұсыныста көрсетілген пікір оң, теріс немесе бейтарап болып табылады ма деген анықтама болып табылады. Мысалы, "зұлым"," қайғылы "және" бақытты "сияқты эмоционалдық жағдай.

Бинарлық шкала бойынша классификациялау. Құжаттың полярлығын екілік шкала бойынша анықтауға болады. Бұл жағдайда құжаттың полярлығын анықтау үшін бағалаудың екі сыныбы қолданылады: оң немесе теріс. Бұл тәсілдің ең кемшілігі-құжаттың эмоциялық құраушысы әрдайым бір мәнді анықтауға болмайды, яғни құжатта оң және теріс баға белгілері болуы мүмкін. Бұл саладағы ерте жұмыстар тауардың шолу полярлығын және Фильмдер туралы пікірлерді айырудың әртүрлі әдістерін қолданатын Терни және Панг еңбектерін қамтиды. Бұл құжат деңгейіндегі жұмыс үлгісі.

Көп жолақты шкала бойынша жіктеу. Көп жолақты мектеп бойынша құжаттың полярлығын жіктеуге болады. Олар 3 немесе 4 балдық шкала бойынша рейтинг болжау жағына қарай "оң немесе теріс" деген бағалаудан киноға пікірлерді жіктеудің негізгі міндетін кеңейтті. Сонымен қатар Снайдер мейрамханалардың шолуларына терең талдау жасап, тағам және атмосфера сияқты әртүрлі қасиеттерінің рейтингтерін болжап көрсетті.

Басқа зерттеу бағыты-субъективтілік/объективтілік идентификациясы. Бұл міндет әдетте осы мәтінді екі сыныптың біріне жатқызу ретінде анықталады: субъективті немесе объективті. Бұл мәселе кейде полярлықты топтастырудан гөрі күрделі болуы мүмкін: сөз бен сөздердің субъективтілігі олардың контекстіне байланысты болуы мүмкін, ал объективті құжат субъективті сөйлемдерді (мысалы, адамдардың пікіріне негізделген жаңалық мақала) қамтуы мүмкін. Сонымен қатар, Су атап өткендей, нәтижелер көп жағдайда мәтіндердің аннотациясы аясында қолданылатын субъективтілікті анықтауға байланысты. Панг қалай болса да, полярлықты жіктеу алдында құжаттан объективті ұсыныстарды жою нәтижелердің дәлдігін арттыруға көмектескенін көрсетті.

Толық талдау моделі функцияның/аспектінің негізінде талдау деп аталады. Бұл модель ұялы телефонда, сандық камерада немесе банкте әртүрлі функциялармен немесе мәндердің аспектілерімен көрсетілген пікірлер мен көңіл-күйді анықтауға сілтеме жасайды. Сипат/аспект-ұялы телефон экраны немесе камераны түсіру сапасы сияқты тоналдыққа зерттелетін нысан атрибуты немесе компоненті. Бұл мәселе бірқатар міндеттерді шешуді талап етеді, мысалы, өзекті мәндерді сәйкестендіру, олардың функцияларын/аспектілерін алу және әрбір функция/аспект бойынша айтылған пікір оң, теріс немесе бейтарап болып табылады ма. Осы шотқа неғұрлым егжей-тегжейлі пікірталастар NLP бойынша анықтамалықтан, "нақтылық пен субъективтілікті талдау" тарауында табылуы мүмкін.

Компьютерлік оқыту элементтерін қолдана отырып, компьютерлік мәтіндерді автоматты түрде талдау жасай алады, мысалы, жасырын

семантикалық талдау, тірек векторлар әдісі, "сөз қапшығы" және осы саладағы семантикалық бағыт. Неғұрлым күрделі әдістер көңіл-күй иесін (яғни адам) және мақсатын (яғни сезімге қатысты мәні) анықтауға тырысады. Мәтінмәнді ескере отырып, пікірді анықтау үшін сөздер арасындағы грамматикалық қарым-қатынастарды пайдаланады.

Грамматикалық байланыстылық қарым-қатынасы мәтінді терең құрылымдық талдау негізінде алынады. Тоналды талдау екі жеке санатқа бөлінуі мүмкін, қолмен (немесе сарапшылар) және автоматты тоналды талдау. Олардың арасындағы ең елеулі айырмашылықтар жүйенің тиімділігі мен талдау дәлдігінде жатыр. Бұл веб-беттерді, онлайн-жаңалықтарды, интернет желісіндегі дискуссиялық топтардың мәтіндерін, онлайн-шолуларды, веб-блогтар мен әлеуметтік медиаларды қоса алғанда, үлкен мәтін массивтерін өңдеуге мүмкіндік береді.

WordNet-Affect-бұл семантикалық тезаурус, онда эмоцияларға байланысты ұғымдар ("эмоционалдық концепттер", ағылш. "affective concepts"), эмоционалдық құрамы бар ("эмоциялық сөздер", ағылш. "affective words"). WordNet-Affect WordNet синсеттерінің "эмоциялық тұжырымдамаға" сәйкес келетін әрбір синсет "эмоциялық сөздердің"көмегімен ұсынылуы мүмкін ішкі жиынынан тұрады.

Осылайша, WordNet-Affect ағылшын тілі үшін WordNet негізінде (сондай-ақ WordNet-Affect және басқа тілдер үшін нұсқалары бар) синонимдер жиынтығын таңдау және әртүрлі эмоционалдық ұғымдарға жатқызу арқылы жасалды. Атап айтқанда, синсеты зат, жалпы, сын есімдерді, үстеу білдіретін сипаттамасы эмоциялар, қолмен талдағыш көмегімен белгіленген, корпус арнайы эмоциялық белгілерді (affective labels, A-labels). Бұл эмоционалдық белгілер көңіл-күйді, эмоционалдық пікірлерді немесе эмоцияларды тудыратын жағдайларды білдіретін әр түрлі жағдайларды сипаттайды.

Сондай-ақ, WordNet-Affect-те олардың эмоционалдық валенттілігіне сәйкес синсеттерді бөлу үшін қосымша эмоционалдық белгілер қолданылады. Ол үшін төрт қосымша эмоционалдық белгілер анықталады: оң, теріс, бір мәнді емес және бейтарап. Біріншісі оң гедонистік сигналдардың (немесе рахат) болуымен сипатталатын эмоциялық күй ретінде анықталатын оң эмоцияларға сәйкес келеді. Ол қуаныш немесе хобби сияқты синсеттерді қамтиды. Жағымсыз белгілер теріс гедонистік сигналдармен (немесе ауырсыну) сипатталады, мысалы, ашу немесе қайғы. Эмоционалдық жағдайларды білдіретін синсеттер, олардың валенттілігі семантикалық контекстке байланысты бірдей емес деп аталады. Ақырында, психологиялық жағдайларды анықтайтын және әрдайым бірдей емес деп қарастырылатын синсеттер, бірақ валенттікпен сипатталмайтын бейтарап болып табылады.

SenticNet-эмоциялық түсініктердің жиынтығымен жұмыс істеу үшін тағы бір семантикалық тезаурус. SenticNet 2010 жылы Массачусет технологиялық институтының медиа-зертханасында іске қосылған жоба болып табылады. Сол уақыттан бері SenticNet жобасы одан әрі дамуды алды

және мәтіннің эмоциялық құрамдас бөлігін талдауға арналған және data mining-тен адамның компьютермен өзара әрекеттесуін ұйымдастыруға дейінгі есептер спектрін қамтитын зияткерлік қосымшаларды жобалау үшін қолданылады. SenticNet-тің басты мақсаты табиғи тілдің көмегімен берілетін тұжырымдамалық және эмоционалдық ақпаратты машиналық тану процедурасын жеңілдету болып табылады. Егер SentiWordNet және WordNet-Affect сияқты басқа да лексикалық тезаурустарды SenticNet-пен салыстырсақ, олардың басты айырмашылықтары sentiwordnet және WordNet-Affect сөздерді және эмоционалдық ұғымдарды синтаксистік деңгейде байланыстыруды қамтамасыз етеді, мысалы, "мақсатқа жету", "жаман сезім", "ерекше оқиғаны тойлау", "өзін-өзі ұстаудан айырылу" немесе "жетінші аспанда болу". senticnet ұғымын семантикалық деңгейде байланыстырады.

Соңғы нұсқасы SenticNet 2 болып табылады. SenticNet 1 нұсқасына қарағанда, ол жай ғана OpenMind, SenticNet 2 корпусынан шамамен 5700 ұғымға үндестік мәнін береді, семантика мен "sentics" (яғни когнитивті және "эмоциялық" ақпарат) 14000-нан астам түсініктермен байланыстырады және SenticNet 1-ге қарағанда табиғи тілдегі мәтіннің терең және көп қырлы талдауын жүргізуге мүмкіндік береді. SenticNet 2 "sentic-есептеулер", парадигма көмегімен құрылған, ол табиғи тілде пікірлерді тану, түсіндіру және өңдеу үшін ИИ және семантикалық паутина әдістерін қолданады.[1]

SenticNet деректерін машинамен оқылатын, компьютерлік бағдарламалармен өңдеуге ыңғайлы түрде ұсыну үшін деректер XML синтаксисін қолдану арқылы RDF-триплеттерге кодталады. "Love" концептіне арналған XML файлының үлгісін жоба сайтынан келесі сілтеме бойынша көруге болады. Мысалы, егер қолданба жұмысы барысында "туған күн" деген ұғым болса, онда SenticNet оны "оқиғалар" жоғары деңгейдегі ұғымға жатқызады және семантикалық жақын ұғымдардың жиынтығымен байланыстырады, мысалы, "тәтті", "достық тосын сый" немесе "клоун" (іздеу нәтижелерін жақсарту үшін қосымша/контекстік ақпарат көзі ретінде пайдаланылуы мүмкін). SenticNet сондай-ақ "sentic-вектор" ұғымын Pleasantness (жағымды), Attention (назар аударыңыз), Sensitivity (сезімталдық) and Aptitude (қабілет) сияқты шамалардың сандық мәндерімен, сондай-ақ тоналдылық шамасын (мәтіннің тоналдығын талдау сияқты есептер үшін), негізгі және қосымша көңіл-күй, сондай-ақ эмоциялық жақын ұғымдардың жиынтығын, мысалы, "мереке" немесе "ерекше оқиға" (мәтіннің тоналдығын анықтау сияқты есептер үшін) сияқты мағыналармен салыстырады.

Сөздерді жіктеу үшін әр сөзге "оң", "теріс" немесе "бейтарап"деген баға сәйкес келетін үндік сөздік пайдаланылады. Соңғы нәтиже алу үшін екі баға мәнін есептеу керек: оң мәтін және теріс. Мәтіннің оң құрамдас бөлігін табу үшін мәтіннің барлық оң терминдерінің тональдылығының сомасын олардың салмағын ескере отырып табу қажет. Мәтіннің теріс құрамдас мәні ұқсас. Барлық мәтіннің тоналдылығын қорытынды бағалау үшін осы құрамдастардың қатынасын мына формула бойынша есептеу қажет: {\displaystyle T=P/N} T=P/N, мұндағы T – тоналдықтың қорытынды бағасы, P

– мәтіннің оң құрамдас бөлігінің бағасы және N – мәтіннің теріс құрамдас бөлігі. Меньшиковтың бабына сәйкес, Т мәні бірлікке жақын мәтін, егер 1 – ден сәл жоғары болса, бейтарап болып саналады. Егер 1 жоғары болса, онда өте оң. Кері дұрыс және теріс үнсіздік мәтіндері үшін. Егжей-тегжейлі осы әдіс қаралды жұмыстарға Голдберга және Пономаревой.

2.5 Класстеризациялау

Кластерлеу – ұқсас объектілер тобына біріктіру-деректерді талдау және Data Mining саласындағы іргелі міндеттердің бірі болып табылады. Ол қолданылатын қолданбалы салалар тізімі, кең: кескіндерді сегменттеу, маркетинг, Алаяқтықпен күрес, болжау, мәтіндерді талдау және т.б. Қазіргі кезеңде кластерлеу деректерді талдау кезінде бірінші қадам болып табылады. Ұқсас топтарды бөлгеннен кейін басқа әдістер қолданылады, әрбір топ үшін жеке модель құрылады.

Кластерлеу міндеті статистика, образдарды тану, оңтайландыру, Машиналық оқыту сияқты ғылыми бағыттарда тұжырымдалған. Сондықтан көптеген синонимдер кластер – класс, таксон, қоюландыру түсініктері.

Қазіргі уақытта объектілер топтарын кластерлерге бөлу әдістерінің саны өте үлкен – бірнеше ондаған Алгоритмдер және олардың тағы да көп модификациялары. Алайда бізді Data Mining-да қолдану тұрғысынан кластерлеу алгоритмдері қызықтырады.

Data Mining кластерлеу деректерді талдау, аяқталған аналитикалық шешімді құру кезеңдерінің бірі болған кезде құндылыққа ие болады. Аналитика жиі ұқсас объектілер тобын бөліп алу, олардың ерекшеліктерін зерттеу және әрбір топ үшін барлық деректерде бір ортақ модель құруға қарағанда жеке модель құру оңай. Мұндай тәсілдерді клиенттердің, сатып алушылардың, тауарлардың топтарын бөліп және олардың әрқайсысы үшін жеке стратегияны әзірлей отырып, маркетингте үнемі пайдаланады.

Data Mining технологиясы жиі кездесетін деректер келесі маңызды ерекшеліктерге ие:

1) деректер қоры кестелерінің жоғары өлшемдігі (мың өрістер) және үлкен көлемі (жүздеген мың және Миллион жазбалар);
2) деректер жиынтықтары сандық және категориялық атрибуттардың көп санын қамтиды.

Барлық атрибуттар немесе объектілердің белгілері сандық (numerical) және категориялық (categorical) болып бөлінеді. Сандық атрибуттар – бұл кеңістікте реттелуге болатын, тиісінше санаттылық – реттелмейтін. Мысалы, "Жас" атрибуты сандық, ал "түс" категориялық. Мәндердің атрибуттарына қосу таңдалған шкаланың түрін өлшеу кезінде болады, ал бұл жалпы алғанда, жеке тапсырма болып табылады [10].

Кластерлеудің көптеген алгоритмдері объектілерді бір-бірімен салыстыруды кейбір жақындық (ұқсастық) шарасы негізінде болжайды. Жақындық өлшемі-объектілердің жақындығының ұлғаюымен шектеуі бар және артатын шама. Ұқсастық шаралары арнайы ережелер бойынша "ойлап

табады", ал нақты шараларды таңдау міндетке, сондай-ақ өлшеу шкаласына байланысты. Сандық атрибуттар үшін жақындық шарасы ретінде келесі формула бойынша есептелетін евклидтық қашықтық өте жиі қолданылады:кластеризация алгоритмдері тво кейбір жақындық (ұқсастық) шараларының негізінде объектілерді өзара салыстыруды болжайды. Жақындық өлшемі-объектілердің жақындығының ұлғаюымен шектеуі бар және артатын шама. Ұқсастық шаралары арнайы ережелер бойынша "ойлап табады", ал нақты шараларды таңдау міндетке, сондай-ақ өлшеу шкаласына байланысты. Сандық атрибуттар үшін жақындық шарасы ретінде өте жиі мына формула бойынша есептелетін евклидтық қашықтық қолданылады:

$$D(x,y) = \sqrt{\sum_i (x-y)^2} \qquad (2.1)$$

Data Mining-те үлкен деректер массивтерін өңдеу қажеттілігі мүмкіндігінше кластерлеу алгоритмін қанағаттандыруы тиіс талаптарды қалыптастыруға алып келді. Оларды қарастырайық:

1) Деректер базасы бойынша өту жолдарының ең аз ықтимал саны;
2) Компьютердің жедел жадының шектеулі көлемінде жұмыс істеу;
3) Алгоритмді кейінірек есептеуді жалғастыру үшін аралық нәтижелерді сақтай отырып тоқтатуға болады;

Деректер базасынан объектілер тек бір бағыттағы курсор режимінде (яғни, жазбалар бойынша навигация режимінде) алынуы мүмкін кезде Алгоритм жұмыс істеуі тиіс.

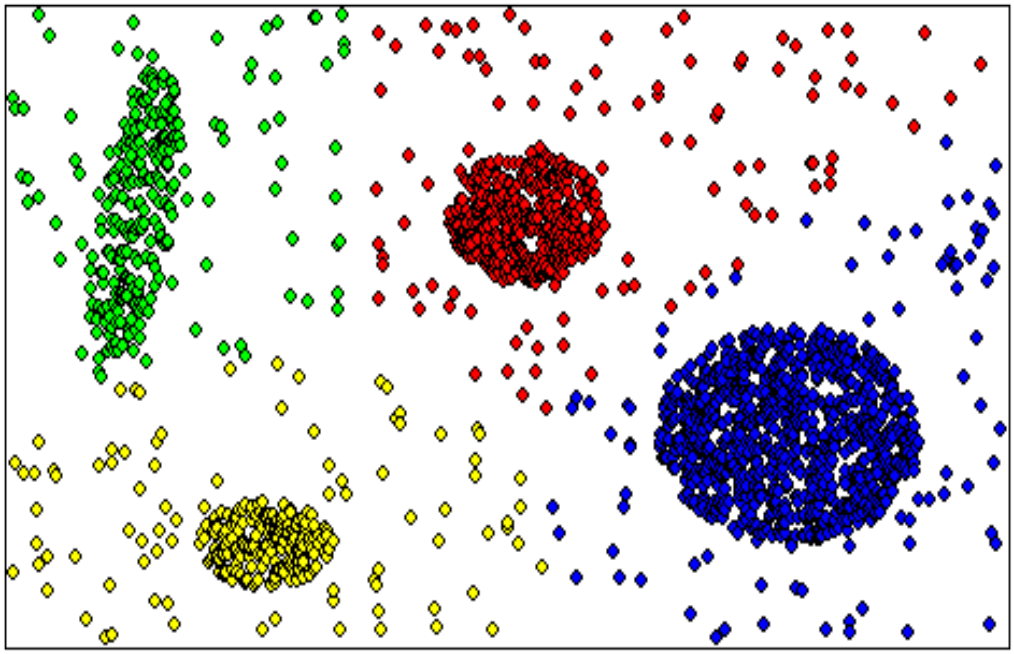
Осы талаптарды қанағаттандыратын алгоритмді (әсіресе екіншісін) масштабталатын (scalable) деп атаймыз. Масштабталу-алгоритмнің есептеу күрделілігіне және бағдарламалық іске асыруға байланысты маңызды қасиеті. Кеңірек анықтама бар. Егер деректер базасындағы жазбалардың санын арттыра отырып, жедел жадының өзгермейтін сыйымдылығында оның жұмыс уақыты сызықтық өссе, Алгоритм масштабталатын деп аталады.

Бірақ көп деректерді өңдеу қажет емес. Сондықтан кластерлік талдау теориясының қалыптасу кезеңінде алгоритмдердің ауқымдылығы мәселелеріне іс жүзінде назар аударылмаған. Барлық өңделетін деректер жедел жадқа ене алады деп болжалды,басты назар кластеризация сапасын жақсартуға әрдайым көңіл бөлінді. Кластерлеудің Жоғары сапасы мен масштабтаудың арасындағы тепе-теңдікті сақтау қиын. Сондықтан Data Mining арсеналында микромассивтерді кластерлеудің тиімді алгоритмдері (microarrays), сондай-ақ аса үлкен деректер қорын өңдеу үшін (large databases) масштабталуы тиіс.кластерлеу алгоритмдерінің тво кейбір жақындық (ұқсастық) шараларының негізінде объектілерді өзара салыстыруды болжайды.

Кластерлерге бөлу тәсілі бойынша алгоритмдер екі түр болады: иерархиялық және иерархиялық емес. Классикалық иерархиялық Алгоритмдер тек категориялық атрибуттармен жұмыс істейді, онда салынған кластерлердің толық ағашы құрылады. Мұнда кластерлер иерархиясын

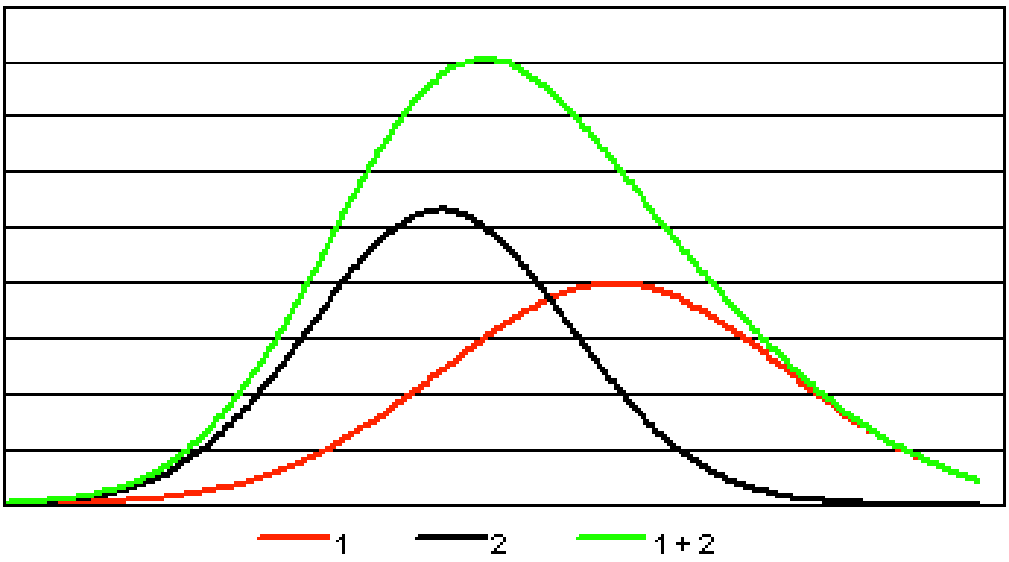
құрудың агломерациялық әдістері таралған-оларда бастапқы объектілерді дәйекті біріктіру және кластерлер санын тиісінше азайту жүргізіледі. Иерархиялық Алгоритмдер кластерлеудің салыстырмалы жоғары сапасын қамтамасыз етеді және кластерлер санын алдын ала тапсыруды талап етпейді. Олардың көпшілігі o(n2) күрделілігі бар.

Иерархиялық емес Алгоритмдер белгілі бір мағынада кластерлерге көптеген объектілердің оңтайлы бөлінуін анықтайтын кейбір мақсатты функцияны оңтайландыруға негізделген. Бұл топта k-орташа (k-means, fuzzy c-means, Густафсон-Кессель) тұқымдастарының алгоритмдері танымал, олар мақсатты функция ретінде объектілер координаттарының іздестірілетін кластерлер орталықтарынан өлшенген ауытқуларының квадраттарының сомасын пайдаланады. Кластерлер сфералық немесе эллипсоидтік пішін іздейді. Канондық іске асыруда функцияны азайту Лагранж көбейткіштерінің әдісі негізінде жүргізіледі және тек жақын жергілікті минимумды табуға мүмкіндік береді. Жаһандық іздеу әдістерін пайдалану (генетикалық Алгоритмдер) алгоритмнің есептеу күрделілігін едәуір арттырады (Сурет 2.1).



Сурет 2.1 – Екі ұрыстың гистограммасы

Алыстан негізделмеген иерархиялық емес Алгоритмдер арасында EM-алгоритмін (Expectation-Maximization) бөліп алу керек. Онда кластерлер орталықтарының орнына математикалық күту мен дисперсияның тиісті мәні бар әрбір кластер үшін ықтималдық тығыздығы функциясының болуы болжанады. Үлестіру қоспасында (Сурет 2.2) ақиқатқа ұқсас максимум принципі бойынша олардың параметрлерін (орташа және стандартты ауытқулар) іздеу жүргізіледі. Em алгоритмі осындай іздеудің бірі. Мәселе алгоритмді бастау алдында Жалпы деректер жиынтығында бағалау қиын үлестірілім түрі туралы гипотеза ұсынылады



Сурет 2.2 – Үлестіру және олардың қоспасы

Тағы бір мәселе объектінің атрибуттары аралас – бір бөлігі сандық түрі бар, ал екінші бөлігі санат. Мысалы, атрибуттары бар келесі нысандар арасындағы қашықтықты есептеу қажет болсын (Жасы, жынысы, білімі):

1) {23, күйеуі, жоғары};

2) {25, Әйел, орта}.

Бірінші атрибут сандық, қалғандары санаттылық болып табылады. Егер біз классикалық иерархиялық алгоритмді ұқсастықтың қандай да бір шарасы бар пайдаланғымыз келсе, біз "жас"атрибутын кемсітуге тура келеді. Мысалы, осылай:

1) {30 жасқа дейін, күйеуі, жоғары};

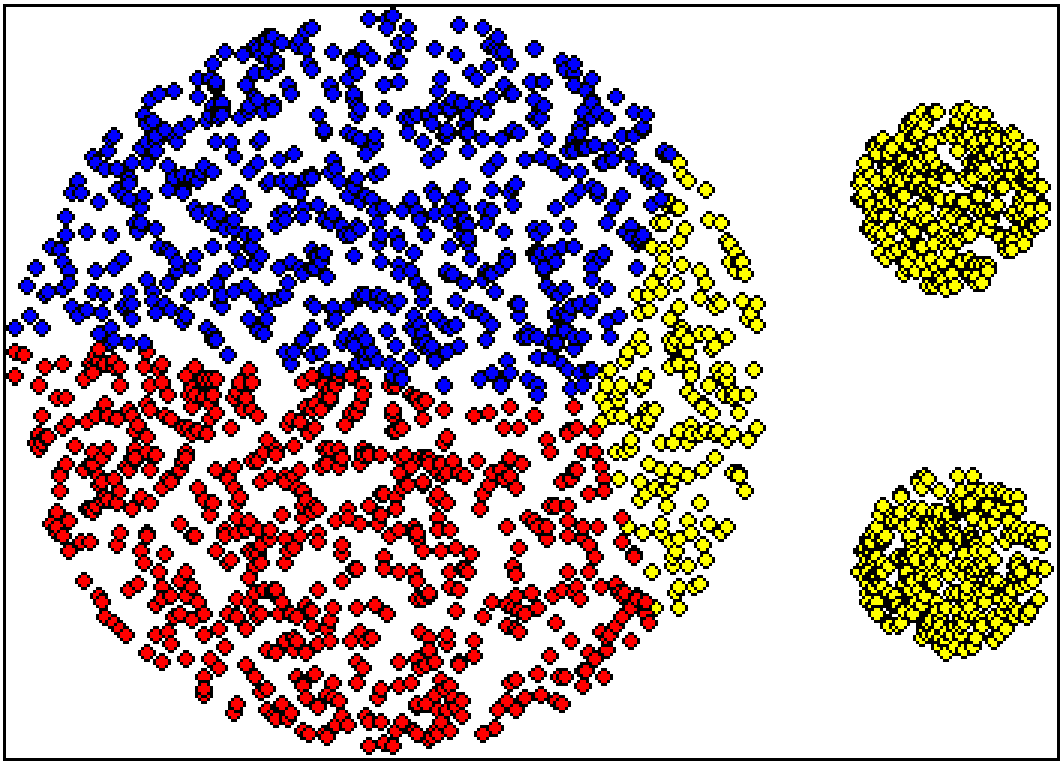
2) {30 жасқа дейін, әйел, орта}.

Бұл ретте ақпараттың бір бөлігі, біз сөзсіз жоғаламыз. Егер біз Евклид кеңістігіндегі қашықтықты анықтаймыз, онда категориялық атрибуттармен сұрақтар туындайды. "Жынысы күйеу" мен "жынысы әйел" арасындағы қашықтық 0-ге тең екені түсінікті, себебі бұл белгінің мәні атаулар шкаласында болады. Ал "Білім" атрибутын әр мағынаға белгілі бір балл бере отырып, атаулар шкаласы мен тәртіп шкаласы бойынша өлшеуге болады. Қандай таңдау керек? Егер категориялық атрибуттар сандық атрибуттардан аса маңызды болса, не істеу керек? Бұл мәселелерді шешу аналитиктің иығына жатады. Бұдан басқа, k-орташа және оған ұқсас алгоритмді пайдалану кезінде санаттық атрибуттарда кластерлер орталықтарын түсінуде, кластерлер санының априорлы тапсырмасында қиындықтар туындайды.

Қашықтықтарға негізделген иерархиялық емес алгоритмдерде мақсатты функцияны оңтайландыру алгоритмі итеративтік сипатқа ие және әрбір итерацияда объектілер арасындағы қашықтық матрицасын есептеу талап етіледі. Көптеген нысандар кезінде бұл тиімсіз және Елеулі есептеу ресурстарын талап етеді. K-means алгоритмінің 1ші итерациясының есептеу күрделілігі O(kmn) ретінде бағаланады, мұнда k,m,n – тиісінше кластерлер,

атрибуттар мен объектілер саны. Бірақ Итерация өте көп болуы мүмкін! Деректер жинағы бойынша көп өту керек.

K-means-те көптеген кемшіліктерге ие сфералық немесе эллипсоидтік формадағы кластерлерді іздеу идеясы. Кеңістіктегі деректер бір-бірінен жақсы ерекшеленетін шағын ұйыған кезде тәсіл жақсы жұмыс істейді. Ал егер деректер ішкі пішінде болса, онда k-means тобының бірде-бірі мұндай тапсырманы ешқашан жеңе алмайды. Сондай – ақ алгоритм бір кластер қалғандарын едәуір көп болған жағдайда нашар жұмыс істейді және олар бір-біріне жақын болған жағдайда-үлкен кластердің "ыдырау" әсері пайда болады (Сурет. 2.3).



Сурет 2.3 – Үлкен кластердің ыдырау әсері

Дегенмен, кластерлеу алгоритмдерін жетілдіру саласындағы зерттеулер үнемі жүріп келеді. Категориялық атрибуттармен (k-modes) және аралас атрибуттармен (k-prototypes) жұмыс істеу үшін k-means алгоритмінің қызықты кеңейтулері әзірленді. Мысалы, k-prototypes-те объектілер арасындағы қашықтықты есептеу атрибуттың түріне байланысты әр түрлі жүзеге асырылады.

Масштабталатын кластерлеу алгоритмдері нарығында күрес алгоритм жұмысы кезінде деректер жинау бойынша әрбір "қосымша" өту жолының төмендеуі үшін жүріп жатыр. K-means және EM (scalable k-means және scalable EM) масштабталатын аналогтары, масштабталатын агломерациялық әдістер (CURE, CACTUS) әзірленді. Бұл заманауи Алгоритмдер ақырғы кластерлеуді алғанға дейін Деректер базасын бірнеше (екіден онға дейін) сканерлеу қажет.

Масштабталатын алгоритмдерді алу жергілікті оңтайландыру функциясынан бас тарту идеясына негізделген. K-means алгоритмінде бір-бірімен объектілерді бумен салыстыру жергілікті оңтайландыру ретінде өзгеше емес, өйткені әрбір итерацияда кластер орталығынан әр объектіге дейінгі қашықтықты есептеу қажет. Бұл үлкен есептеу шығындарына әкеледі. Ол ескі мән, жаңа объект және кластерлік сипаттамалар (clusters features) негізінде есептеледі. Нақты кластерлік сипаттамалар қандай да бір алгоритмге байланысты. BIRCH, LargeItem, CLOPE және басқа да Алгоритмдер осылай пайда болды.

Осылайша, кластерлеу бірыңғай әмбебап алгоритмі жоқ. Кез келген алгоритмді пайдалану кезінде оның артықшылықтары мен кемшіліктерін түсіну, ол жақсы жұмыс істейтін мәліметтердің табиғатын және масштабталу қабілетін ескеру маңызды.

2.6 Классификациялау

Деректер-белгілі бір нысанда ұсынылған және одан әрі пайдалануға арналған қандай да бір жүйені, құбылысты, процесті немесе объектіні сипаттайтын мәліметтер.

Осы анықтамаға ақпарат пен деректер ұғымдары арасындағы арақатынасты түсіндіретін мынадай ескертулер жасау қажет:

Деректер-бұл ақпарат мазмұнын ұсынудың нақты нысаны (мысалы, қоршаған ортаның температурасын бақылау нәтижелері туралы ақпаратты сандық массив (кесте) түрінде ұсынуға болады, бірақ кесте түрінде де, кейбір тіл арқылы мәтіндік сипаттама түрінде де ұсынуға болады ·);

Табиғатта бар бағытталмаған (адресті емес, шашыраңқы) ақпаратқа қарағанда, бізге және оған деген қажеттіліктерімізге қарамастан, деректер тек тұтынушы үшін маңызы бар ақпарат деп аталады және демек, оны қандай да бір міндеттерді шешу үшін пайдалану көзделеді; басқаша айтқанда, табиғи ақпаратқа қарағанда, деректердің практикалық мәртебесі мен маңыздылығы жоғары.

Әрине, техникалық құрылғылардың көмегімен практикалық міндеттерді шешу кезінде ақпаратты ұсыну нысандары әрдайым нақты және ақпаратта біреу мүдделі, сондықтан "деректер" терминін қолдану әбден ақталған.

"Деректер" ұғымының мазмұны өте кең. Ол қандай да бір жеке шаманы, мысалы, адамның туған жылын немесе оның атын, сондай-ақ қандай да бір датчиктің көрсеткіштерін немесе фирманың өндірістік мәліметтерін қамтиды. Тұрмыстық деңгейде деректер мәліметтермен теңестіріледі, сондықтан кез келген ақпараттық массив деректер болып саналмайды. Мысалы, әдеби шығарманың немесе оқулықтың мәтіні, суретшінің суреті, фильм деректер ретінде қарастырылмайды,алайда олардың мазмұнындағы мәліметтер деректермен танылады. Компьютерлік жүйелерде мұндай айырмашылық жоқ және компьютер үшін рұқсат етілген нысанда ұсынылған кез келген ақпарат - мәтіндер, суреттер, музыка және т.б. - мәліметтер болып саналады. Сондай-ақ, ақпарат көзіне сыртқы тасымалдағыштарда сақталатын немесе компьютер

жадына жүктелген бағдарламалардың мәтіндері жатады. Бұл мазмұн бойынша кеңейтілген деректер терминінің түсіндірмесі келесі болып табылады.

Деректерге бірнеше жіктеу белгілері жазылады. Олардың ішіндегі ең маңыздысы деректер түрі болып табылады. Деректер түрін анықтайды:

- олардың рұқсат етілген мәндерінің жиынтығы;
- оларды өңдеу ережелері · түрлендіру);
- сақтау кезінде оларды ЖҚҚ және ЖҚҚ орналастыру тәртібі;
- оларға қол жеткізу тәртібі (яғни қажет болған жағдайда сақтау орнынан жүгіну және алу).

Деректер типтерінің рұқсат етілген жинағы және олардың ерекшеліктері бағдарламалық жүйемен немесе жүйе жазылған бағдарламалау тілімен анықталады. Бұл ретте тілдердің мүмкін болатын деректер түрлерінің әртүрлілігі, сондай-ақ жаңа үлгілердің құрылуы бойынша мүмкіндіктері өте қатты ерекшеленеді. Анық, неғұрлым кең және икемді көрсетіледі типтеу деректерінің бағдарламалық жүйесінде немесе тілінде көбірек мүмкіндіктер беріледі пайдаланушыға міндеттерін шешуде оңтайлы ұсыну, сақтау және қолдану. Деректерді типтеу ең орындалатын бағдарламаның ықшамдығына да әсер етеді. Мысалы, BASIC тілінде "жазбалар" деректер түрі жоқ; нәтижесінде деректер базасын құру және пайдалану үшін бірнеше массивтерді қатар өңдеуді ұйымдастыруға тура келеді [11].

Келесі белгі деректерді қарапайым (жалғыз, қарапайым) және құрылымдалған (күрделі) болып бөлу болып табылады.

Қарапайым деректерге символдар, сандар (бүтін және заттай) және логикалық деректер жатады. Жеке мәліметтердің жалпы және міндетті ерекшелігі-олардың әрқайсысы бір мағынаға ие және өз атының. Мән-бұл жад ұяшықтарының мазмұны. Аты (оны идентификатор деп те атайды) - бұл бағдарлама мәтінінде берілген белгі. Қарапайым деректердің идентификаторын құру ережелері жазылған бағдарламаны бағдарламалау тілімен анықталады.

Қарапайым деректер біріктіру жолымен күрделі деректер салынатын "кірпіштер" болып табылады. Біріктіру нұсқалары көп - бұл деректер құрылымының көптеген түрлерінің пайда болуына әкеледі.

Деректер мен байланысты (қатынастарды) біріктіретін ақпараттық массив құрылымдық деректер деп аталады.

Біріктірілген жеке деректер тізбесі, олардың сипаттамалары, сондай-ақ олардың арасындағы байланыс ерекшеліктері деректер құрылымын құрайды.

Құрылымдалған мәліметтердің мысалдары оқушылардың тегі, сабақ күндері мен белгілері, Телефон анықтамалығы, мекеменің ұйымдық құрылымы және т. б. бар сынып журналының беті болып табылады.

Деректер құрылымының рұқсат етілген тізбесі бағдарламалау тілімен немесе қолданбалы бағдарламамен анықталады. Ол BASIC тілінде немесе бағдарламалаудың кіріктірме мүмкіндіктерінсіз қолданбалы бағдарламаларда сияқты бекітілген (кеңейтілмейтін) болуы мүмкін. Программалаудың дамыған тілдерінде (PASCAL, C және т. б.) және бірқатар қолданбалы жүйелерде

деректер құрылымының резервтелген үлгілерімен қатар жаңа типтерді құруға жол беріледі, бұл ретте, құрылым элементтері күрделі деректер болуы мүмкін, мысалы, жазбалар массиві.

Күрделі деректер, элементар сияқты, мәндер мен идентификаторлар бар. Мәндер белгілі бір схемалар бойынша ОЗУ ұяшықтарында орналастырылады (6-тармақты қараңыз.Ескерту.Ескерту.). Идентификаторларды құру ережелері бағдарламалау тілімен немесе бағдарламалық жүйемен белгіленеді. Файл аттарын қалыптастыру ережелері - олар операциялық жүйемен беріледі және онда жұмыс істейтін барлық бағдарламалар мен тілдерді сақтауы тиіс. Мысалы, MS-DOS-та Файл атауы ретінде латын әріптерінен, цифрлардан және жалпы ұзындығы 8 белгіден аспайтын кейбір арнайы символдардан комбинациялар рұқсат етіледі; 32 биттік файлдық жүйесі бар Windows 95 (98) жүйесінде қолданылатын таңбалар жиынтығын шектемей, ұзындығы 255 белгіге дейінгі атаулар рұқсат етілген.

Жалпы өңдеу барысында деректер мәндерін (қарапайым және құрылымдалған) өзгерту мүмкіндігі бойынша оларды ауыспалы және тұрақты (тұрақты) деп бөледі. Атаудан әлбетте, айнымалылар бағдарламаны орындау барысында өз мәнін өзгертуі мүмкін, ал константтар-жоқ. Операциялық жүйе деңгейінде ауыспалы және тұрақты шамалар арасындағы айырмашылық жоқ, сондықтан оларда ОЗУ-да орналастырудың және оларға қол жеткізудің бірдей тәртібі бар. Бөлу бағдарламалау тілінде және оның көмегімен жасалған қолданбалы бағдарламада жүргізілуі мүмкін; мұндай бөлу бағдарламаның корректілігін синтаксистік бақылаудың қосымша шарасы болып табылады.

Деректер өңдеудің қай кезеңінде қолданылатындығына байланысты олар бастапқы (кіру), аралық және демалыс болып бөлінеді. Бастапқы деректерге бағдарламаны орындау үшін қажетті және оған жұмыс басталғанға дейін немесе жұмыс процесінде енгізілетін деректер жатады. Бастапқы деректер алдын ала кейбір тасымалдауышта жазылуы және одан енгізілуі, байланыс желілері бойынша қандай да бір датчиктерден немесе басқа компьютерлерден түсуі, бағдарлама пайдаланушысы енгізу құрылғылары арқылы енгізілуі мүмкін. Аралық деректер бағдарламаны орындау барысында қалыптасады және көбінесе пайдаланушыға қолжетімсіз; олар шығару құрылғыларында көрсетілмейді, бірақ ОЗУ немесе ВЗУ бар. Аралық деректерге сәйкестендіргіштер бағдарламаны әзірлеуші береді немесе оған енгізілген ережелер бойынша бағдарламаның өзі анықтайды. Шығыс деректері бағдарлама жұмысының нәтижесі болып табылады-олар үшін және кіруді өңдеу жүргізіледі. Адамға арналған шығыс деректері оған талап етілетін нысанда (мәтіндер, суреттер, дыбыстар) ұсынылады; шығыс деректерін тасығыштарда сақтау немесе желілер арқылы беру кезінде оларды ұсынудың екілік компьютерлік форматы сақталады. Осылайша, бағдарлама жұмысын кіріс деректерін демалыс күндері осы үшін қажетті аралық арқылы түрлендіру іс-әрекеттері ретінде қарастыруға болады. Бағдарламаның өзі тұрғысынан бұл түрлердің барлығы тең, яғни функционалдық мақсатына немесе кезеңіне емес, олардың түріне сәйкес ғана өңделеді.

Деректерді сақтау және өңдеу кезінде ұсыну үш негізгі міндеттерді шешуді талап етеді:

- қарапайым (қарапайым) деректерді ұсыну тәсілдерін анықтау;
- деректерді құрылымға біріктіру тәсілдерін анықтау;
- материалдық тасығышта ақпаратты орналастыру тәсілдерін орнату ·

Деректерді ұсынудың үш деңгейін атап көрсетеді - тұжырымдамалық, логикалық және физикалық. Тұжырымдамалық деңгейде ақпараттық массивтің жалпы құрылымы анықталады - ол деректер үлгісі деп аталады. Бірнеше деректер үлгілері белгілі және пайдаланылады: иерархиялық, желілік, реляциялық, объектілі-бағытталған. Таңдалған деректер моделіне сәйкес деректер сақталатын ақпараттық жүйе, сондай-ақ оларды өңдеуді жүргізетін бағдарламалар (деректерді манипуляциялау) құрылады. Логикалық деңгей Элементарлық деректерді ұсыну тәсілдерін, олардың құрылымға біріктіру кезіндегі тізбесін, сондай-ақ таңдалған деректер моделі шеңберінде олардың арасындағы байланыстардың сипатын анықтайды. Физикалық деңгей ақпараттың сыртқы тасымалдауышында (магниттік немесе оптикалық дискілерде, қағазда, компьютер жадында) құрылған деректердің логикалық құрылымын орналастыру форматтарын анықтайды. Деректерді ұсыну сақтау кезінде ақпаратты жазу және оларды пайдалану кезінде қажетті деректерге жылдам қол жеткізуді қамтамасыз ететін маңызды фактор болып табылады. Бұдан әрі компьютерлік жүйелерде аталған есептерді шешу нұсқалары қарастырылады.

3. Жүйені әзірлеу

3.1 Жүйені әзірлеу үшін қолданылған технологиялар

Тональды талдау жүйесін әзірлеу үшін Python бағдарламалау тілі қолданылды. Бұл бизнестің басымдықты міндеттерін шешу үшін, сондай-ақ, ақпараттың үлкен көлемін талдау саласындағы көшбасшылардың бірі болып табылатын ең танымал және қолданылатын тілдерінің бірі болып табылады.

Python-бұл интерпретацияланатын, объектілі-бағытталған жоғары деңгейдегі динамикалық типтеу, жадыны автоматты басқару, тізімдер, кортеждер, сөздіктер сияқты ыңғайлы жоғары деңгейлі деректер құрылымы бар бағдарламалау тілі. Кластарды, модульдерді, ерекшеліктерді өңдеуді, сондай-ақ көп нүктелі есептеулерді қолдайды. Тілдің артықшылығы оның қарапайым және мәнерлі синтаксисі бар, құрылымдық, объектілі-бағытталған, функционалдық және аспектілі-бағытталған бағдарламалау қолдайды. Едәуір қарапайым синтаксисі оңай әрі қарапайым, кез келген уақытта бағдарламаны жаңартып тексерген уақыт жағынан өте тиімді.

Python бағдарламаны әзірлеу стильдерінің кең тізімін қолдай алады, соның ішінде, ООП және функционалдық бағдарламалау жұмыс істеу үшін өте ыңғайлы.

Тілдің ең танымал интерпретаторларының бірі Си жазылған Python. Бұл даму ортасы еркін лицензия бойынша тегін таратылады. Интерпретатор ең танымал платформаларды қолдайды.

Питон белсенді дамып келеді. 2 жылда бір рет жаңартулар шығады. Тілдің маңызды ерекшелігі ANSI, ISO және басқа да кодтау стандарттарының болмауы болып табылады, олар интерпретатордың арқасында жұмыс істейді.

Питон-ең "жас" бағдарламалау тілі емес, бірақ тым ескі емес. Оны құру кезінде Паскаль немесе Си сияқты "монстралар" болды. Сондықтан ЯП құру кезінде авторлар әзірлеушілерге арналған түрлі платформалар ішінен ең үздік алуға тырысты. Шын мәнінде, Python 8 түрлі тілдерден астам табысты шешімдердің өзіндік "джем" болып табылады. Мысалы, байт компиляция Питон құруға дейін пайда болды, бірақ оған өте сәтті болды.

Питон барлық кең таралған операциялық жүйелерді қолдайды. Ол қалта компьютерлерінде және үлкен серверлерде жақсы жұмыс істей алады. Платформа айтарлықтай ескірген жағдайда, ол ядроның қолдауынан шығарылады. Мысалы, 2.6 бастап тіл нұсқалары Windows 95, 98 және ME платформаларымен жұмыс істемейді. Қажет болған жағдайда тілдің қазіргі құралдарын қолданудан бас тарта отырып, ескі нұсқаларды пайдалануға болады. Содан кейін бағдарлама осы OS қоса жұмыс істейтін болады. Ескі нұсқалар үшін мезгіл-мезгіл патчи шығады. Тіл Java виртуалды машинасымен жұмыс істей алады.

ЯП нақты құрылымдалған семантикалық ядро және қарапайым синтаксис бар. Бұл тілде жазылған барлық нәрсе әрдайым оңай оқылады. Егер аргументтерді жеткізу қажет болса, тіл call-by-sharing функциясын пайдаланады.

Тілде операторларды теру өте стандартты. Синтаксисте блоктың басы мен ұшын білдіретін фигуралық немесе операторлық жақшалар жоқ. Мұндай шешім бағдарлама денесінің жолдарының санын едәуір қысқартады және программист кодты жазу кезінде жақсы стиль мен ұқыптылықты сақтауға үйретеді.

Осы платформадағы барлық кітапханалар модуль ретінде жазылады. Мұндай тұжырымдаманың артықшылығы пакетке бірнеше модульдерді жинау мүмкіндігі болып табылады.

ЯП интерпретаторы көптеген пайдалы функцияларды қолдайды. Жөндеу режимінде жұмыс істеу кезінде, онымен интерактивті жұмыс істеуге болады. Бағдарлама мәтіні пернетақтадан енгізіледі, содан кейін оны бірден орындауға және дисплей нәтижесін алуға болады. Мұндай тәсіл бағдарламаның жеке модульдерін немесе бөліктерін жылдам тексеруге мүмкіндік береді.

Питон тамаша және өте егжей-тегжейлі құжаттамамен ерекшеленеді, соның ішінде орнатылған анықтамалық жүйе түрінде. Функциялары мен кітапханалардың барлық атауларын басыда ұстау мүмкін емес. Мұнда pydoc анықтама стандартты кітапханасы көмекке келеді.

Оған қол жеткізу үшін help функциясын шақыру жеткілікті, одан кейін жақшада сізді қызықтыратынын дәлел ретінде көрсету керек.

Мысалы, help (os) OS кітапханасы бойынша көмек шақырады. Егер модульдің атауын есіңізге түсіре алмасаңыз, help ('modules') конструкциясын көрсетіңіз және сіз экранға барлық стандартты модульдердің тізімін аласыз.

Python көмегімен әзірленген текстты тазалау және тілді анықтау функциясы кіріс деректерді сүзуге мүмкіндік береді. Сүзгілеуден және талдаудан кейін барлық деректер дерекқорға жазылады.

Тілді анықтау үшін langdetect python кітапханасы арқылы жүзеге асырылды. Langdetect кітапханасын google әзірлеген, сондықтан тілді тану сапасы жоғары деңгейді көрсетеді. Кітапхана 55 тілді қолдайды, соның ішінде бізге керек орыс тілі бар.

Деректер базасымен жұмыс істеу үшін sqlite3 кітапханасы таңдалды, ол деректер базасын құруға және базаға әртүрлі сұраныстар жасауға және оны толтыруға мүмкіндік береді. SQLite3 оңтайлы опция, егер үлкен кесте құру жоспарланбаған болса. Python SQLite3-нақты кітапхана емес, бұл жеке модульдер бағынатын белгілі бір ережелер жиынтығы. Мұндай иерархияның арқасында әртүрлі деректер базаларымен жұмыс жүргізуге мүмкіндік бар.

SQLite-бұл SQL 92 стандартының көп бөлігін іске асыратын қ кітапхана. Оның танымалдылығына деген талабы база қозғағышының өзі, сондай-ақ оның интерфейсі (дәлірек айтқанда оның қозғағыштары) бір кітапхана шегінде, сондай-ақ барлық деректерді бір файлда сақтау мүмкіндігі болып табылады. MySQL және PostgreSQL арасында бір жерде SQLite функционалдылық позициясы. Дегенмен, іс жүзінде, SQLite жиі 2-3 есе (тіпті одан да көп) жылдам. Бұл жоғары ретсіз ішкі архитектураның және "сервер-клиент" және "клиент-сервер" типті қосылыстардың қажеттілігін жоюдың

арқасында мүмкін. Бір пакетке жиналған осының барлығы MySQL кітапханасының клиенттік бөлігінің көлемі бойынша аз ғана көп, толық мәліметтер базасы үшін әсерлі жетістік болып табылады. Жоғары тиімді инфрақұрылымды пайдалана отырып, SQLite кез келген басқа ДБ жүйелерінен әлдеқайда аз, оған бөлінген жадтың кішкентай көлемінде жұмыс істей алады. Бұл SQLite деректер базасына жүктелетін кез келген міндеттерде іс жүзінде пайдалану мүмкіндігімен өте ыңғайлы құрал жасайды.

SQLite бағдарламасының құрамдас бөлігі болып табылатын, ДБ қозғалтқышы кітапхана болып табылады. Барлық ДБ бағдарлама орындалатын машинада жалғыз стандартты файлда сақталады.

Бірнеше процестер немесе ағындар бір базадан деректерді бір уақытта қандай да бір проблемаларсыз оқи алады. Базаға жазуды қазіргі уақытта ешқандай басқа сұрау салуларға қызмет көрсетпеген жағдайда ғана жүзеге асыруға болады; олай болмаған жағдайда жазбаға әрекет сәтсіздікпен аяқталады және бағдарламаға қате коды қайтарылады. Оқиғалардың дамуының басқа нұсқасы берілген уақыт аралығы ішінде жазу әрекеттерін автоматты түрде қайталау болып табылады.

Деректерді талдау алдында жиналған деректерді қажетсіз ақпараттан тазарту қажет. Деректерді тазарту үшін тұрақты өрнектер қолданылды. Python бағдарламалау тілі тұрақты өрнектерді қолтайды. Тұрақты өрнектер (ағылш. regular expressions, жарг. регэкспа немесе регекса) - іздеу үшін үлгілерді жазу арнайы жүйесіне негізделген мәтінді өңдеу жүйесі. Үлгі (ағылш . pattern), іздеу ережесін беретін, кейде "шаблон", "маска"деп аталады. Тұрақты өрнектер мәтін фрагменттерін іздеу жүйесі болып табылады. соңғы автоматтардың әрекеті. Тұрақты өрнектер тақырыптарды тікелей іздеудің басқа әдістерінен ерекшеленеді, іздеу критерийлерін дұрыс қоюға мүмкіндік береді. В қазіргі уақытта тұрақты сөздер көптеген салаларда қолданылады: қолданбалы Лингвистикада морфологиялық талдау және стемминг, талдау және салыстыру үшін биоинформатика анықтау жүйелерінде ДНҚ молекуласының вирустар мен спамды сүзуге арналған желілік шабуылдар.

3.2 Деректерді жинау және талдау схемасы

Деректерді жинау көзі ретінде вконтакте әлеуметтік желісі таңдалды. Жастардың көпшілігі осы әлеуметтік желіде көп уақыт өткізеді.

Деректерді жинау бағдарламаның ең маңызды бөлігінін бірі болып табылады. Деректерді жинау үшін көптеген кітапханалар әзірленді: urllib2, requests, BeautifulSoup, Vkapi.

Urrlib2 url адресімен жұмыс істеу үшін пайдаланылатын Python модулі. URL модулі - basic және digest аутентификация, cookie және т.б. жұмыс істеуге көмектесетін өз функциялары мен класстары бар. Urrlib2 сұрауына тақырыптарды аргумент ретінде қосу үшін Request object қабылдай алады. Requests талғампаз және жай HTTP сұрауларды орындауға мүмкіндік береді. Осы кітапханада барлық HTTP сұраныстарын пайдалануға, авторизациядан өтуге және бізге қажетті мәліметтерді алуға болады.

Requests пен urrlib2 ұқсастығына қарамастан, олардың арасында елеулі айырмашылықтар бар. Біріншіден, requests-те барлық жауаптар автоматты түрде Unicode – әлемнің барлық жазбаша тілдерін қамтитын символдарды кодтау стандартына кодталады. Екіншіден, requests кітапханасының көмегімен алынған барлық жауаптар мазмұнын автоматты түрде сақтайды. Сондықтан, бұл нысандарға бірнеше рет жүгіну мүмкіндігі бар, бұл өз кезегінде неғұрлым икемді код жазуға мүмкіндік береді.

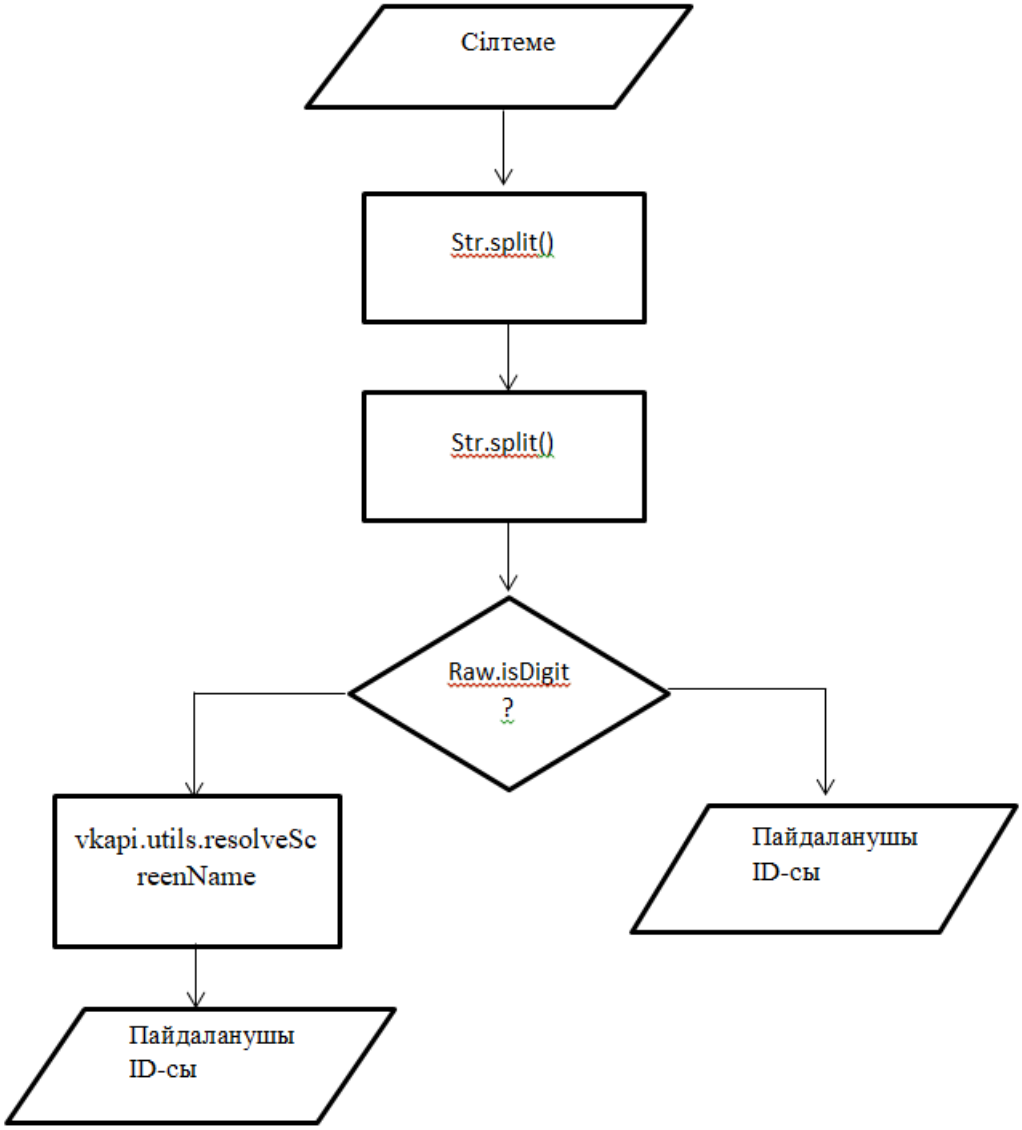
BeautifulSoup3 HTML / XML файлдарын синтаксистік талдау үшін парсер модулі болып табылады. Бұл модуль әр түрлі шарлау, түрлендіру және синтаксистік талдау ағашын іздеу тәсілдерін қолдайды. Beautiful Soup конструкторына жол түрінде XML немесе HTML құжаты қажет (немесе ашық файлға ұқсас нысан). Ол синтаксистік талдау жасайды және құжатқа сәйкес деректер құрылымын жадына жасайды. Beautiful Soup көмегімен Жақсы ресімделген құжатты өңдесеңіз, бөлшектелген құрылым бастапқы құжат сияқты көрінеді. Бірақ егер оның белгісі қателерді қамтитын болса, онда Beautiful Soup ең қолайлы деректер құрылымын құру үшін эвристикалық әдістерді қолданады. BeautifulSoup класы веб-шолғыштарда толығымен қолданылатын эвристиктерден тұрады, Бұл HTML файлдарының авторларының ой-пікірлері туралы болжам жасауға мүмкіндік береді.

VKAPI кітапханасы Вконтакте әлеуметтік желісінде ақпарат жинауды жүзеге асыруға мүмкіндік береді, деректерді жинау үшін әртүрлі функциялардың кең саны бар. Осындай жинау үшін арналған кітапханалар ішінен ол өзінің қарапайымдылығымен және көп функциялығымен ерекшеленеді. Сондай-ақ, VKontakte әлеуметтік желісінің API-нің барлық ресми әдістерін пайдалану мүмкіндігі үлкен артықшылығы болып табылады. Дәлелдер атауы=мән түрінде беріледі, бұл кодты түсінікті және қарапайым етеді. Сондай-ақ, өз standalone қолданбасын жасау арқылы белгіні алу керек.

Осы кітапхананың көмегімен жиналған барлық деректер json форматында сақталады. JSON-мәтін, жұп-кілт жиынтығы: мән. Әр түрлі тілдерде бұл нысан, жазба, құрылым, сөздік, хеш-кесте, кілтпен тізім немесе ассоциативті массив ретінде жүзеге асырылған. Кілт тек жол болуы мүмкін (тіркелімге тәуелді: әр түрлі тіркелімдегі әріптермен есімдер әртүрлі болып саналады), мәні – кез келген нысан. Кілттер арқылы бізге қажетті деректерді алуға болады. BeautifulSoup3 және requests кітапханаларының кең мүмкіндіктеріне қарамастан, таңдау vkapi-ге түсті. Бұл модуль ВКонакте әлеуметтік желісімен жұмыс істеуге арналған және барлық ресми әдістерді қолданады.

Бағдарламаға пайдаланушы профиліне сілтеме беріледі. Содан кейін әзірленген функцияның көмегімен VKapi деректерді жинау үшін қажетті пайдаланушының ID-сы алынады. Сілтеменің өзінде ID болмауы мүмкін, сол себептен ең алдымен сілтемеде ID бар ма деген сұраққа жауап табу керек. Python тілінің split функциясы көмегімен, бөлгішті пайдалана отырып, жолды бөліктерге бөліп тізіммен қайтаруға болады. Бөлгіш символ ретінде “/” символы қолданылады және тізімнің соңғы элементін тексеру қажет, себебі

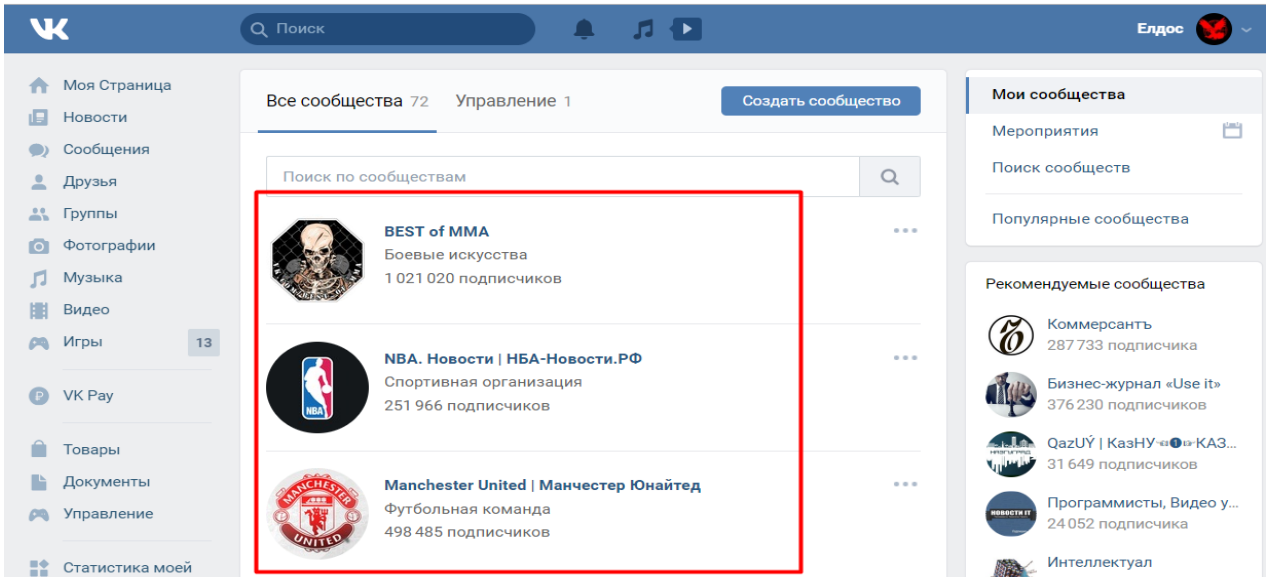
сілтемеде ID жолдың ең соңында орналасады, мысалы vk.com/id196199167. Келесі кезеңде сүзіп алынған элементтен “id” символдарын replace функциясымен бос жолға алмастырып, isdigit функциясының көмегімен жол тек сандадан тұратынын анықтаймыз. Егер жол тек сандардан тұратын болса, онда ол id екендігін білдіреді. Ал жол тек сандардан тұрмаса vkapi.utils.resolveScreenName функциясы көмегімен id-ны анықтаймыз. Сипатталған модельдің блок сызбасы (Сурет 3.1).



Сурет 3.1 – ID-ны өңдеу модулінің блок сызбасы

Бағдарлама жұмысының келесі кезеңінде, пайдаланушы ID-сының көмегімен деректерді іздеу жүзеге асырылады. Алдымен vk api нысанын пайдалана отырып, id standalone қосымшаларының көмегімен авторландырылған сессияны орнатып деректерді жинау үшін пайдаланушының логині мен паролін еңгізу қажет. Жаңадан ашылған пайдаланушы парағын қолдана беруге болады.

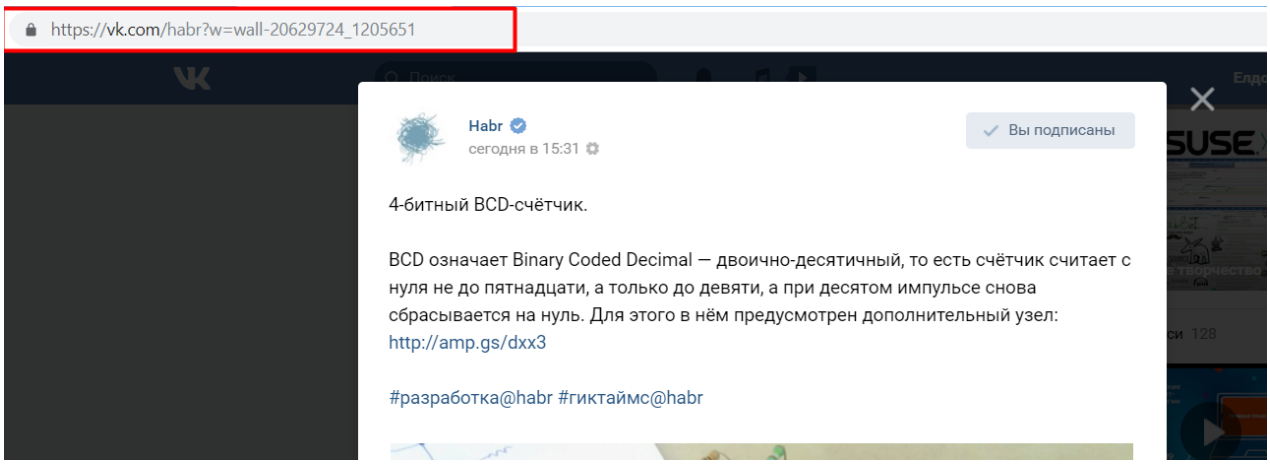
Пайдаланушы тұратын топтардың тізімін жинау үшін Get_groups функциясы әзірленді. Вконтакте әлеуметтік парақшасында мәліметтің 75% астамы топтардан жарияланады. Сайттың топтары (Сурет 3.2).



Сурет 3.2 – Топтардың тізімі

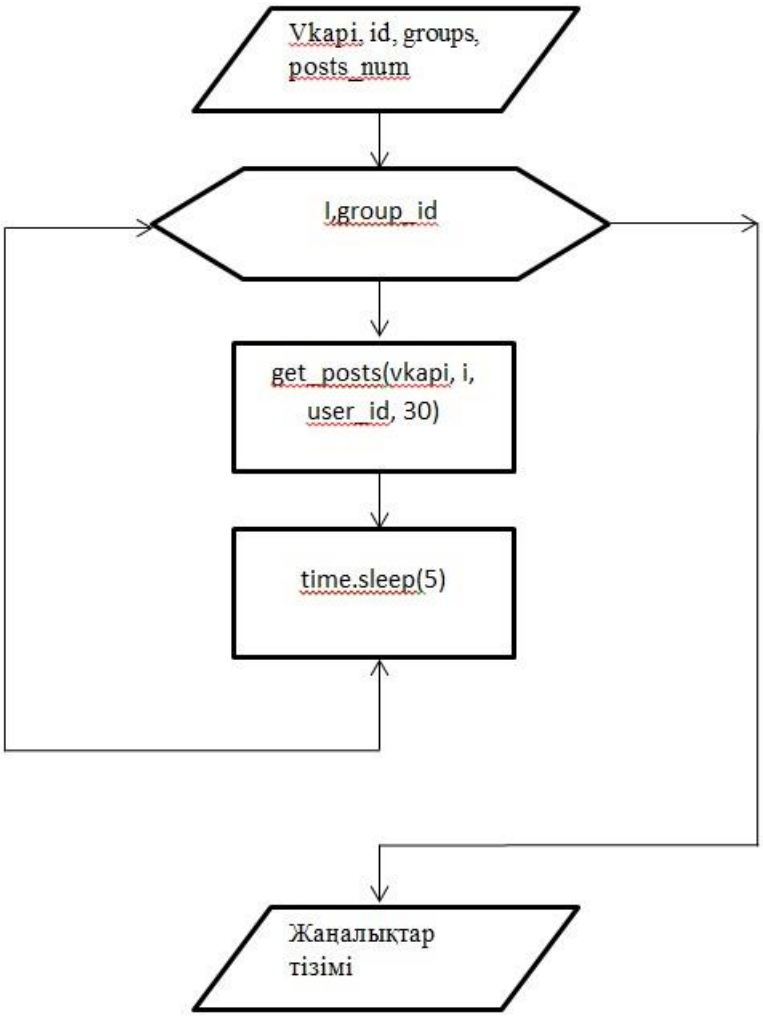
Get_groups функциясы, пайдаланушының id және vkapi нысанын қабылдайды. Бұл функция API groups әдісі арқылы топтар туралы барлық ақпаратты алады. Items кілтінің көмегімен қажетті топтардың барлық id-лерін жинайды және деректерді одан әрі жинау үшін массивке қосады. Жиналған массив көмегімен толық немесе тез анализ жүргізуге болады. Толық анализ режиминде пайдаланушының барлық топтары қарастырылады, ал жылдам режиминде ең танымал бес топ қарастырылады.

Жиналған топтардың тізімдері бойынша сол әлеуметтік парақшаларда жарияланған жаңалықтарды жинап анализдеу қажет. Әлеуметтік желіде жаңалықтарға сілтемелер топтың id-сын және жаңалық id-сын қамтиды (Сурет 3.3).



Сурет 3.3 – Жаңалыққа сілтеме

Жаңалақтарды жинау үшін get_posts функциясы әзірленді. Get_posts функциясы тізім бойынша топтарда қажетті жазба санын алуға мүмкіндік береді. Кіруде Vk api нысанын, пайдаланушының id, топтар тізімін және жинау қажет жазбалар санын қабылдайды. Wall әдісі get функциясының көмегімен барлық жазбаларды жинайды. Name кілті бойынша мәтіндік деректер іріктеледі. Әр топтың жаңалақтарын жинаған кейін 5 секундтық таймаут қажет, себебі көп мәліметті ретсіз жинау жүзеге асырылса, Вконтакте әлеуметтік парақшасының әкімшілері күдікті әрекеттерді байқап, парақшаны блоктауы мүмкін. Функцияның блок схемасы (Сурет 3.4).



Сурет 3.4 – Жаңалықтарды жинау блок сызбасы

Талдау кезінде ең маңызды кезеңдердің бірі-деректерді тазалау. Талдау алдында деректерді жеткіліксіз өңдеу кезінде мәтіннің тоналдығы аз дәлдікпен анықталуы мүмкін. Деректерді тазалау үшін langdetect кітапханасы және Python стандартты функциялары пайдаланылды. Тазалау үшін мәтіндік деректер json деректерінен алынады. Осыдан кейін мәтіннің орыс тіліне тиесілілігін langdetect кітапханасының көмегімен тексеру жүргізіледі. Бұл кітапхана тілді анықтау үшін google әдістерін пайдаланады. Тілді анықтау алгоритмі детерминацияланбаған болып табылады. Бұл ретте, егер сіз оны

тым қысқа немесе бір мәнді мәтінде іске қосуға тырыссаңыз, оны әрбір іске қосу кезінде әр түрлі нәтиже алуға болады.

Кітапхана 53 тілді қолдайды, бірақ онда қазақ тілін анықтау мүмкіндігі жоқ. Орыс және қазақ тілінің ұқсастығынан, кейде модуль екі тілді бірдей деп санайды. Мәтінде қазақ рәміздерінің болуын тексеру үшін жүйелі түрде тазалау функциясы жасалды. Тұрақты өрнектер (ағылш. regular expressions) - метасимволдарды (Джокер символдары, ағылш. wildcard characters). Іздеу үшін үлгі жолы қолданылады (ағылш . pattern, оны жиі "шаблон", "маска" деп атайды), символдар мен метасимволдардан тұратын және іздеу ережесін беретін. Мәтінмен манипуляция үшін қосымша ауыстыру жолы қойылады, ол да арнайы таңбаларды қамтуы мүмкін.

Егер тазалау функциясы қазақ символдарын тапса, деректер одан әрі талдау үшін базаға жазылды, ал қалған деректер тазартудың келесі кезеңіне өтті. Жаңалықтарды жазу тілі бойынша сүзу қазақ бекеттерінің санын айтарлықтай төмендетті. Келесі кезеңде мәтін lower жол функциясы арқылы төменгі регистрге аударылады. Тазалау кезінде барлық сілтемелер, смайликтер және пайдаланушылар аттары жойылады. Осылайша, тазалау процесінде ұсыныстарда тек орыс әріптері қалдырылады.

Деректер үнділігін талдау үшін табиғи тілді (Natural Language Processig, NLP) өңдеу үшін қолданылатын Polyglot кітапханасы пайдаланылды. Модульдің құрамына Word2Vec кіреді, ол барлық лингвистикалық заңдылықтарды векторларға бөлінген сөздер арасындағы косинустық қашықтықты бөлу арқылы қамтуға мүмкіндік береді. Осы манипуляциялардың арқасында ұқсас контекстерде кездесетін сөздер оңай. Осының арқасында олар синонимдер болуы мүмкін деген қорытынды жасауға болады.

Word2Vec жұмыс алгоритмі:

1) бірінші кезеңде корпуста әр сөздің кездесуі есептеледі;

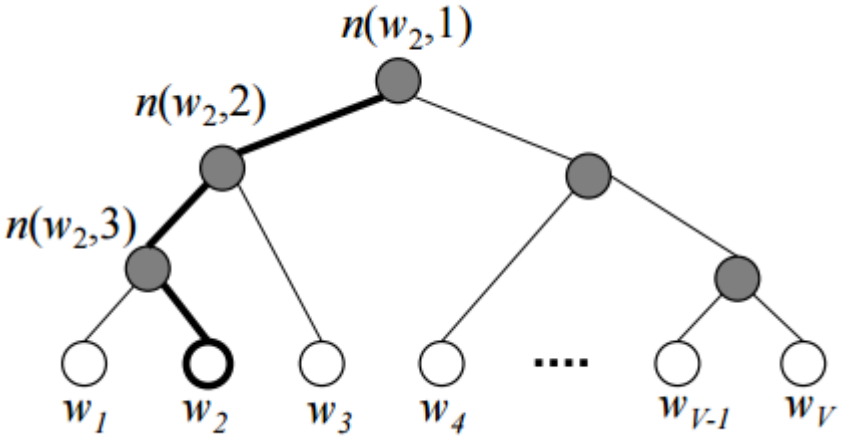
2) алдымен барлық сөздер массивке айналады және хэш-кестеде сақталады. Содан кейін олар жиілікте сұрыпталады және массивтегі ең сирек сөздер жойылады.

3) келесі кезеңде сөздікті кодтау үшін Хаффман ағашы пайдаланылады. Ол 3.5-ші суретте көрсетілген;

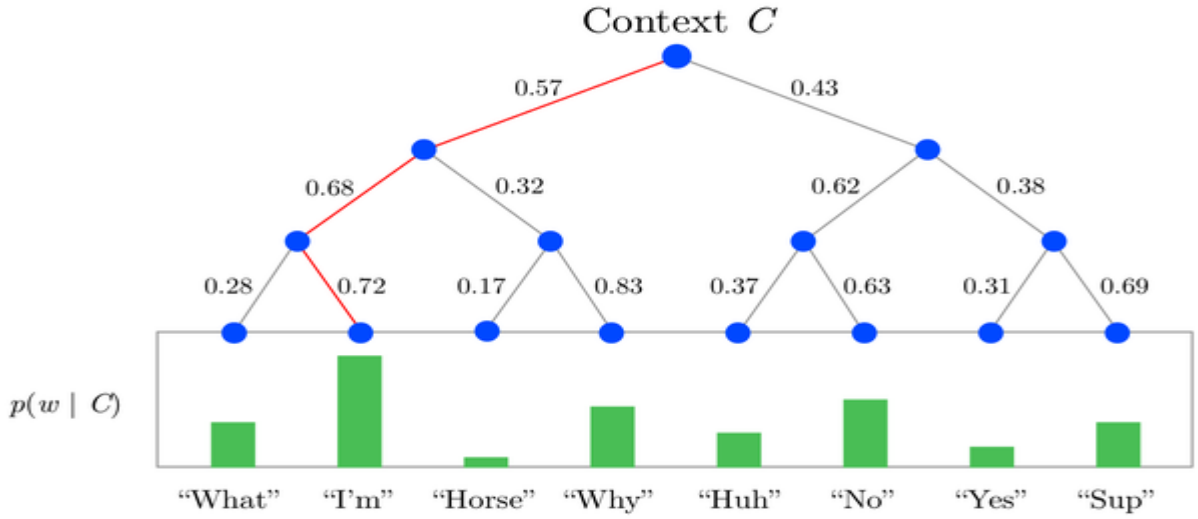
4) корпуста қосалқы ұсыныстар бар және жиі кездесетін сөздердің қосалқы эмпликациясы пайдаланылады. Қосалқы ұсыныс корпустың негізгі элементі болып табылады. Көбінесе бұл ұсыныс, бірақ кейде сөйлем-Бұл мақала. Қосалқы эмплирлеу көмегімен жиіліктік сөздер алынып, талданады, бұл өз кезегінде нейрондық желіні оқыту сапасын арттырады;

5) таңдалған ұсыныс бойынша өту терезесімен. Терезе өлшемі параметр ретінде алгоритмге қойылады. Терезе-бұл ағымдағы және сөйлемдегі сөзді болжау арасындағы ең үлкен қашықтық;

6) соңғы кезеңде hierarchical softmax функциясы бар тікелей таралу нейросеть қолданылады (Сурет 3.6).



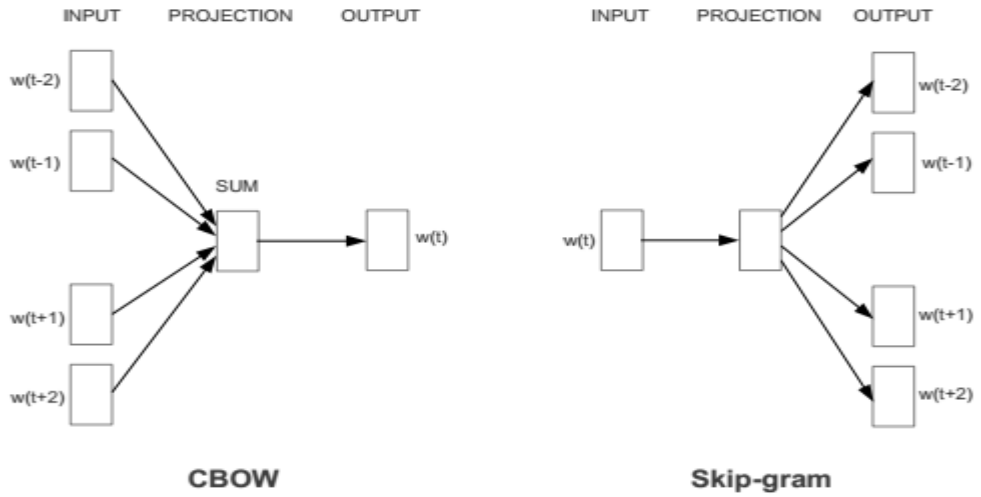
Сурет 3.5 – Хаффман ағашы



Сурет 3.6 – Hierachical softmax

Word2vec екі негізгі архитектураны жүзеге асырады – Continuous Bag of Words (C BOW) және Skip-gram. Кіріске мәтін корпусы беріледі, ал шығыста сөз векторлары алынады.

CBOW әдісі жақын сөздер негізінде сөзді болжайды. Skip-gram әдісі өз кезегінде кері әдісті қолданады – жақын сөз жиынтығының бір сөзі негізінде болжау. Бұл екі әдіс жіктеу алгоритмі ретінде нейрондық желілерді пайдаланады. Бірінші кезеңде сөздікте кез келген сөз кездейсоқ N-өлшемді вектор болып табылады. Оқыту барысында әрбір сөз үшін оңтайлы вектор қалыптасады.CBOW және WORD2VEC архитектуралары (Сурет 3.7).



Сурет 3.7 – CBOW және Skip-gram архитектурасы

Бірақ бұл синонимділік Word2Vec оқыту жүргізілген тақырып бойынша ұқсас деректер негізінде ғана өзекті екенін есте сақтау керек. Егер оқыту үшін баскетбол жаңалықтарын пайдаланып, сәндегі трендтер туралы мақалаларда оқытылған модельді қолданса, талдаудың дәлдігі қатты төмендейді. Бұл мәселе polyglot модулінде ішінара шешілді, нейрондық желі деректерді үлкен таңдауда оқытылған және тақырып бойынша әр түрлі мәтіндердің тоналдығын дәл анықтай алады.

Үнсіздікті талдау үшін тек жеке сөздің ғана емес, бүкіл мәтіннің тоналдығын анықтауға мүмкіндік беретін функция жазылған. Сондай-ақ, мазмұнды теріс және позитивті бөлуге арналған polarity=0,27 коэффициенті таңдалды. Әр түрлі деректермен тест кезінде бұл коэффициент нақтылықты анықтаудағы ең жоғары дәлдікті көрсетті.

Қалаусыз мазмұнды анықтағаннан кейін барлық деректерді деректер қорына жазу мәселесі пайда болды. Деректер базасымен жұмыс істеу үшін sqlite3 кітапханасы таңдалды, ол деректер базасын құруға және базаға әртүрлі сұраныстар жасауға және оны толтыруға мүмкіндік береді. SQLite3 оңтайлы опция, егер үлкен кесте құру жоспарланбаған болса.

Python SQLite3-нақты кітапхана емес, бұл жеке Модульдер бағынатын белгілі бір ережелер жиынтығы. Мұндай иерархияның арқасында әртүрлі деректер базаларымен жұмыс жүргізілуде.

Барлық деректерді сақтау үшін User_id - пайдаланушы идентификаторын, group_id - топтар идентификаторын және талдау нәтижесі бар мәтінді қамтитын кесте жасалды. Әзірленген деректер базасы тұрақты жұмыс істейді, себебі салмағы өте жеңіл. Жиналған дерекектердің кестесі (Сурет 3.8).

Таблица: user

	user_id	group_id	text	content
	Фильтр	Фильтр	Фильтр	Фильтр
1	328871130	55692777	Ребята на работе сильно нравится одна ...	positive
2	328871130	55692777	Почему чужие парни, притягивают больш...	positive
3	328871130	55692777	Что делать думаю о ней каждый день э...	negative
4	328871130	55692777	Опубликуйте сегодня по возможности!У ...	negative
5	328871130	55692777	Хочу купить хромакей откуда заказыват...	negative
6	328871130	55692777	Кто учился в Китае с программой mychin...	positive
7	328871130	55692777	Кто в курсе люди добрые.А правда ли чт...	positive
8	328871130	55692777	Всем привет , тут такое дело что мне уж...	negative
9	328871130	55692777	I have a question Здравствуйте!	positive

Сурет 3.8 – Деректер базасы

Деректер базасын толтыру сол циклда жүргізілді, онда деректердің теріс контентке тиістілігі талданды. Деректерді толтыру үшін execute script командасы пайдаланылды. Format көмегімен скриптті орындау үшін жол жиналды.

3.3 Пайдаланушы интерфейсінің элементтері

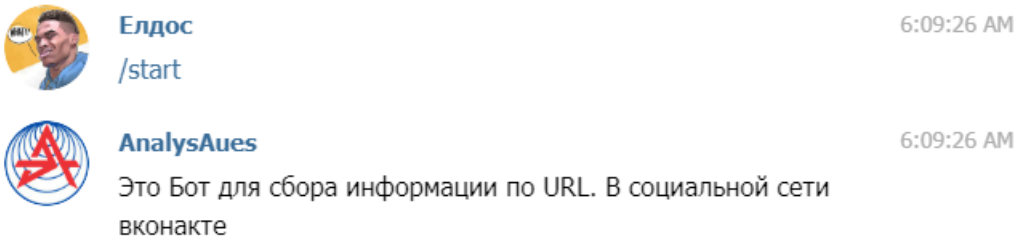
PyTelegramBotAPI кітапханасының көмегімен бағдарлама мен пайдаланушы арасындағы интерфейс үшін бот әзірленді. Бот telegram интерфейсі арқылы кез келген бағдарлама шақыруы мүмкін . Бұл сіздің телефоныңыздағы чат пен компьютердегі немесе сервердегі код арасындағы сұхбат. Нақты сұрақ туындауы мүмкін: неге сайт емес? Біріншіден, барлық адамдар жеке тәсілді жақсы көреді. Бот пайдаланушыға жеке қызмет көрсетеді деген әсер жасайды, және ол тіпті сол жағында ақылды бағдарлама бар деп болжай алмайды. Екіншіден, бот жауаптарының икемділігі мен жылдамдығы кез келген уақытта қайта бағдарламалауға және қайта іске қосуға болады. Үшіншіден, ботты пайдалану үшін клиентке ешқандай тіркеу және қосымша қосымшаларды жүктеу қажет емес, тек оның жеке мессенджері ғана жеткілікті.

Кітапхананың өзі ботты әтрүрлі тәсілдермен бағдарламалауға мүмкіндік береді. Ыңғайлы интерфейс үшін түймелер жасау қолдауы бар. Пайдаланушы хабарламаларына әрекет ету үшін өңдеушілер пайдаланылады. Өңдеуші боттан пайдаланушыға не қажет екенін анықтауға мүмкіндік береді.

Өңдеуші арқылы командаларға жауап ретінде артүрлі әрекеттерді бағдарламалауға болады. Әзірленге ботпен жұмыс істеу үшін оған start командасын жіберу керек. Message_handler бұл пәрменді танып, жауап ретінде боте туралы ақпарат бар хабар жібереді. Содан кейін пайдаланушы профиліне сілтеме жіберу қажет. Handle_message функциясы сілтемені қабылдайды және оны URL айнымалысына жазады.

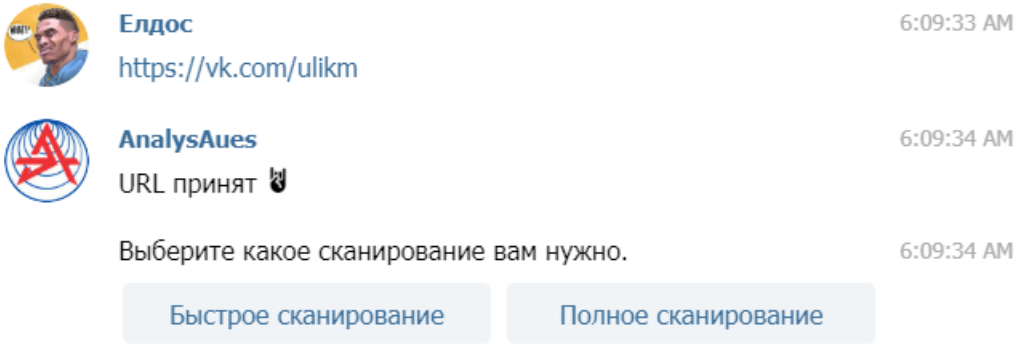
Кейде сервер немесе интернет тұрақсыз жұмыс істей алады және бот пакеттерін аздаған жоғалтса, өте қиын болады. Ол non_stop=true параметрімен Long Polling іске қосады. Long Polling-бұл "ұзын сұраныстар"арқылы жаңа оқиғалар туралы деректерді алуға мүмкіндік беретін технология. Сервер сұрау алады, оған жауапты бірден жібермейді, тек қандай да бір оқиға болған кезде (мысалы, жаңа хабар келгенде) немесе күту уақыты аяқталғанда ғана.

Ботпен жұмысты бастау үшін старт командасын жазу қажет. Бұл командаға жауап ретінде бот өзі туралы ақпаратты жібереді (3.9 Сурет).



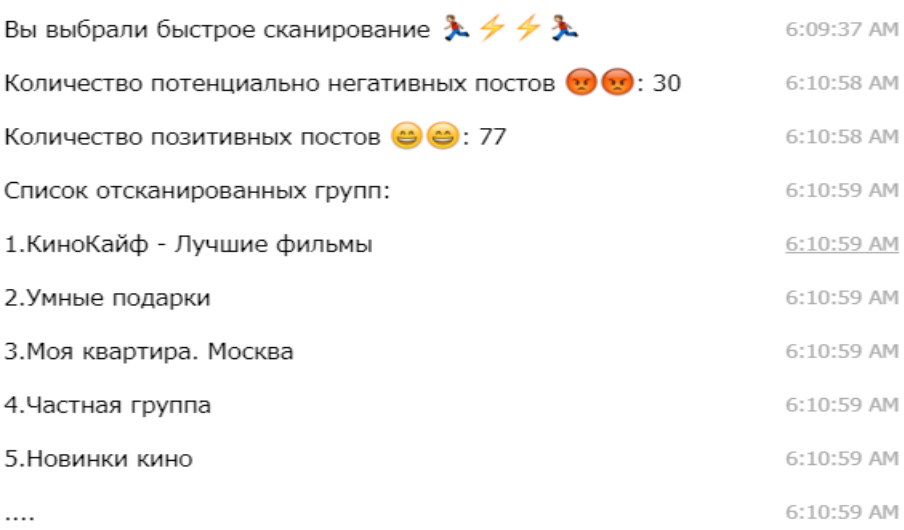
3.9 Сурет – Ботпен әрекет ету

Одан кейін пайдаланушы профиліне сілтеме жіберу керек (3.10 Сурет).



Сурет 3.10 – Ботқа сілтеме жіберу.

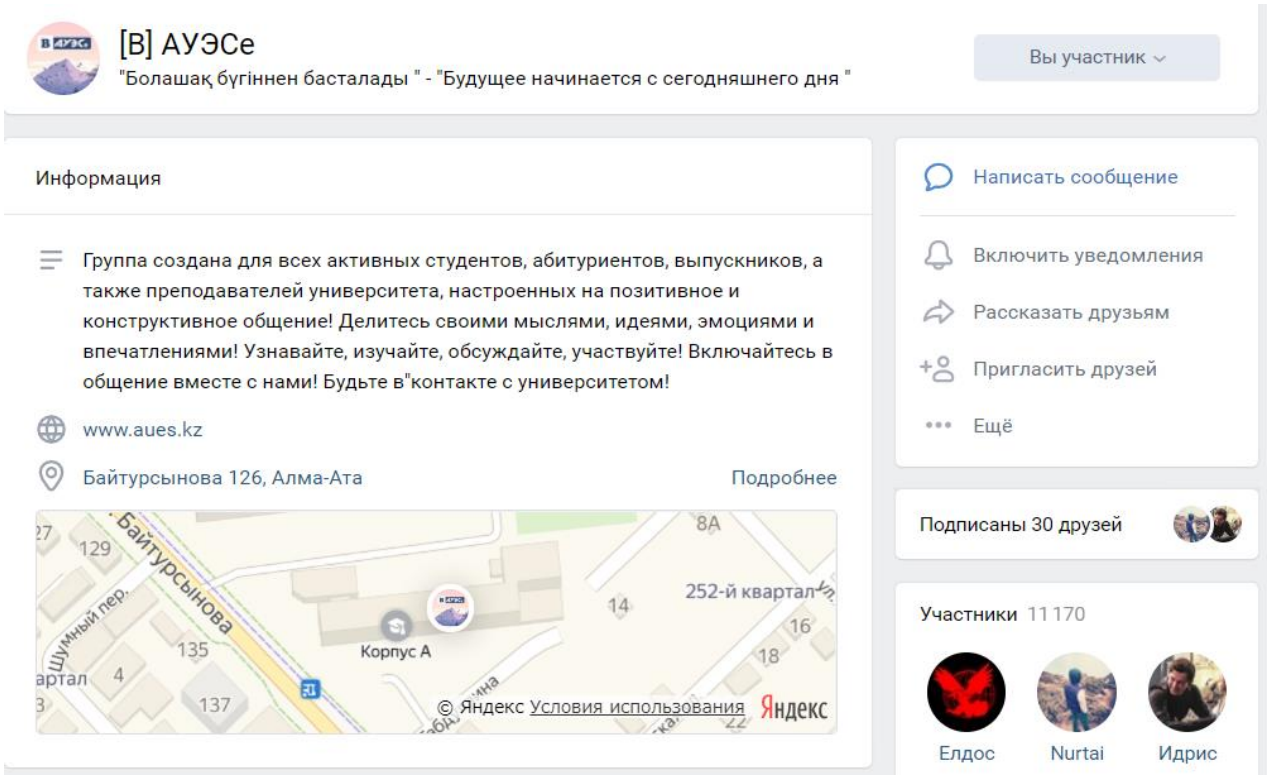
Екі сканерлеу режимі бар, бот режимдерінің бірін таңдағаннан кейін деректерді жинау функциясына сканерленетін пайдаланушының профиліне сілтеме жібереді. Бот әзірленген функциялырдан көмегімен тұрақты және жылдам жұмыс істейді. Ddos секілді шабуылдарға қарсы, кезектер модулі арқылы жіберілген сұраныстарды стек арқылы жүйелеп қалыпты жұмыс істеу мүмкіндігін алады. Боттың интерфейсі көрсетілген (Сурет 3.11).



3.11 Сурет – Анализдалған деректерді шығару

3.4 Эксперимент нәтижелері

Әзірленген жүйені тестілеу үшін АУЭС-тің әлеуметтік тобы қолданылды. Тест өткізу үшін мың пайдаланушы парақтары тексерілді. Осы топты полигон ретінде таңдаудың себептері оқырмандардың жеткілікті саны, негізгі комьюнити жастар болып табылуы. Топ барлық белсенді студенттер, талапкерлер, түлектер, сондай-ақ университет оқытушылары үшін құрылған. Топта үнемі жаңалықтар жазылып, бет үнемі жаңартылып отырады. Әлеуметтік парақшаның негізгі беті көрсетілген (Сурет 3.12).

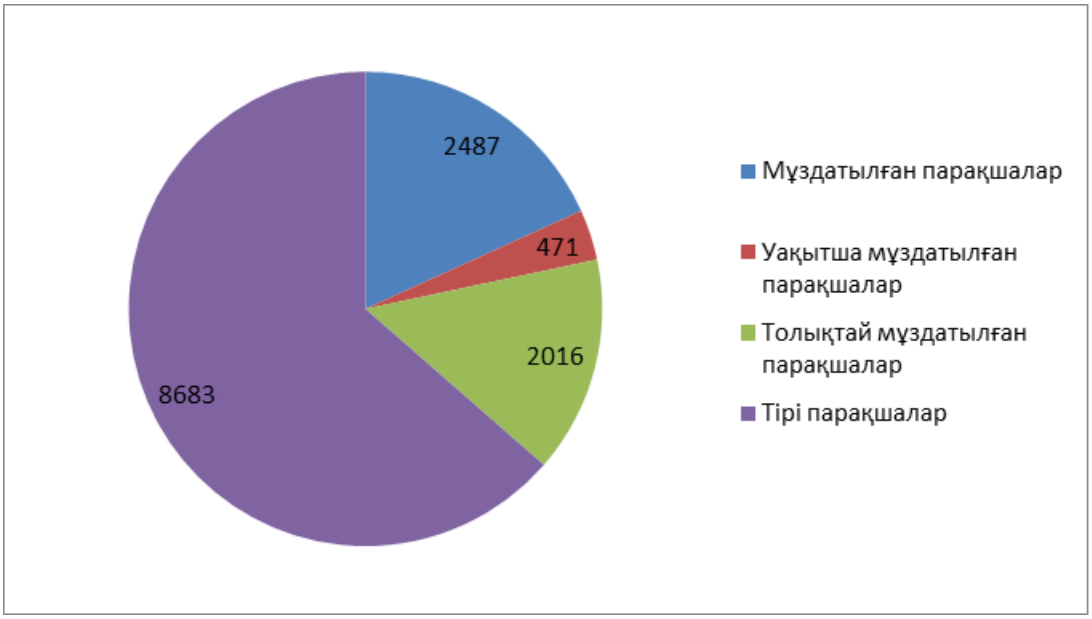


3.12 Сурет – [B] АУЭСе әлеуметтік тобы

Талдау алдында топ оқырмандары арасында мұздатылған беттердің болуын тексеру қажет болды. Memedia сканерының көмегімен тексеріс жүргізілді. Толық статистика (Кесте 3.1). Оқырмандар диаграммасы (Сурет 3.13).

3.1-Кесте. Оқырмандар статистикасы

Оқырмандар саны	11170
Мұздатылған парақшалар	2487
Уақытша мұздатылған парақшалар	471
Толықтай мұздатылған парақшалар	2016
Тірі парақашалар	8683



3.13 Сурет – Оқырмандар статистикасының диаграммасы

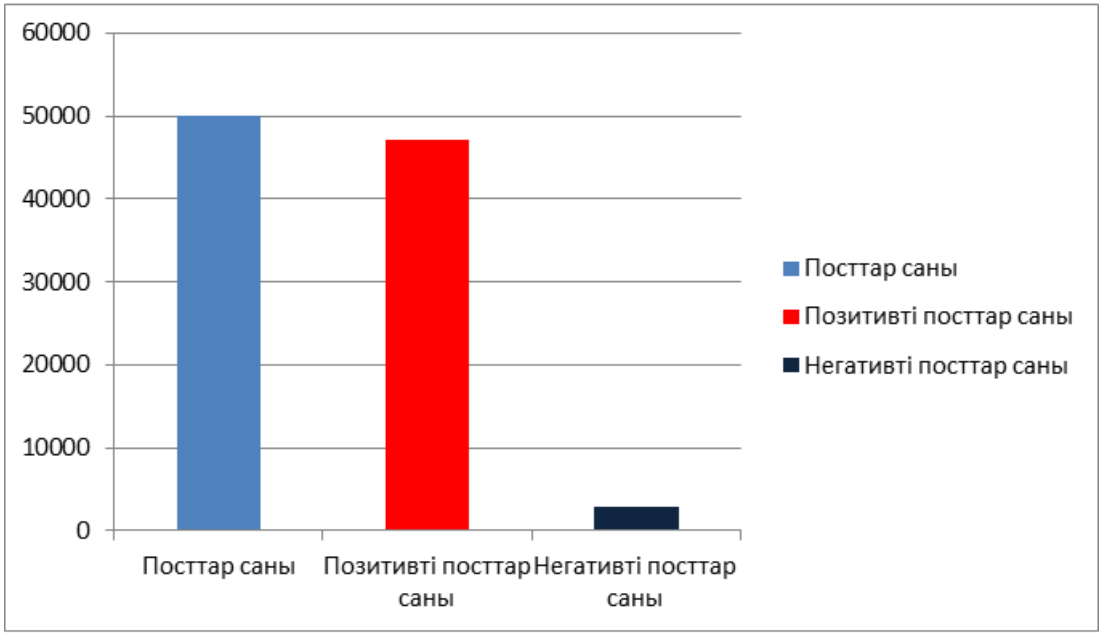
Тестілеу үшін тірі парақшалардың ішінен мың парақша таңдалды. Әр парақшаның 5 тобынан 10 жаңалықтан жиналып анализденді. Әзірленген жүйе үшін көптеген ақпарат көлемі қиындық тұдырмады. Барлық мәлімметтерді жинап талдау үшін он бес минут уақыт қажет болды. Орындау уақыты интернет жылдамдығына және компьютердің сипаттамасына байланысты қатты өзгеруі мүмкін. Біздің сынақта компьютердің орташа конфигурациясы мен жылдам емес интернет қолданылды. Сол себептен жүйенің жинап талдау модулі жеткілікті жылдам деп санауға болады.

Хабарламаның үнсіздігін анықтау дәлдігіне келетін болсақ, ол 73 пайызды құрады. Жалпы 50мың посттар көлемінде контент талданды. Жиналған посттардың 47117-сі позитивті бағаға, ал қалған 2887-сі негативті бағаға ие болды.

Деректер аймағын 50000 жазбадан 2887-ге дейін тарылтып, әзірленген жүйе арқылы тыйым салынған контентті анықтауды 17 есе оңтайландыруға қол жеткізілді. Толық статистика (Кесте 3.2). Посттар статистикасының диаграммасы (Сурет 3.14).

3.2-Кесте. Оқырмандар статистикасы

Посттар саны	50000
Позитивті посттар саны	47117
Негативті посттар саны	2887



3.14 Сурет – Посттар статистикасының диаграммасы

4 Техникалық-экономикалық негіздеме

Бұл дипломдық жобаның мақсаты тыйым салынған контентті анықтау үшін Интернет желісіндегі ақпараттық объектілерді зияткерлік талдау жүйесін әзірлеу болып табылады. Жүйе топтардың, әлеуметтік желілердегі пайдаланушылардың контентін тексеруге мүмкіндік береді.

Бағдарламалық қамтамасыз етуді әзірлеуге бас әзірлеушіден және әзірлеуші бағдарламашыдан тұратын мамандар тобы қатысады. Бас әзірлеушінің міндетіне жүйенің архитектурасын сақтау және әзірлеу, кодты оңтайландыру және кестені құру кіреді. Программист-әзірлеуші міндетіне техникалық негіздеме әзірлеу, бағдарламалық қамтамасыз етуді әзірлеу, оны тестілеу және сүйемелдеу кіреді. Техникалық-экономикалық негіздеме мынадай тармақтардан тұрады:

- бағдарламалық қамтамасыз етуді әзірлеудің еңбек сыйымдылығын анықтау;
- БҚ әзірлеуге арналған шығындарды есептеу;
- дайын өнімнің құндылығын анықтау;

4.1 Әзірлеу күрделілігін анықтау

Бағдарламалық жасақтаманы әзірлеудің күрделілігін дәл анықтау үшін барлық тапсырманы қарапайым кезеңдерге бөлу қажет. Бұл күрделі міндетті неғұрлым қарапайым тапсырыстарға бөлу есебінен бағдарламалық қамтамасызтандыруды әзірлеу прогресін тиімді бақылауға мүмкіндік береді. Менің көзқарасым бойынша мұндай тәсіл неғұрлым тиімді болып саналады және нәтижелі, тез табыс табуға мүмкіндік береді. Моделі бөлу күрделілігі әзірлеу және игеру сатысындағы 5.1-кестеде көрсетілген.

4.1-кесте – Бағдарламалық қамтамассыздандырудың әзірлеу кезеңдері

Әзірлеу кезеңдері	Жұмыс түрі	Еңбек сыйымдылығы, адам сағ..
Кезең 1	Тапсырмалар қою	8
Кезең 2	БҚ әзірлеуге ТТ әзірлеу және бекіту	8
Кезең 3	Мұндай бағдарламаларды іздеу және зерттеу	32
Кезең 4	Ілеспе әдебиеттерді іздеу және зерттеу	16
Кезең 5	БҚ бойынша талдау кестелерін құру	8
Кезең 6	Дипломдық жұмыстың теориялық бөлімін рәсімдеу	24
Кезең 7	Дипломдық жобаның тәжірибелік бөлігін әзірлеу	32
Кезең 8	Жобаны іске асыру	32

Кезең 9	Ақауларды жөндеу және жою	24
Кезең 10	Тестілеу	16
Кезең 11	БҚ әзірлеу бойынша қорытынды шығару	8
Кезең 12	Енгізу	24
Барлығы: жобаны орындаудың еңбек сыйымдылығы		232

Жұмыс күнінің ұзақтығы 8 сағатқа тең. Нәтижесінде бағдарламалық қамтамасыздандыруды іске асыру үшін 29 жұмыс күні қажет.

4.2 БҚ әзірлеуге арналған шығындарды есептеу

Бағдарламалық қамтамасыз етуді әзірлеу үшін қажетті шығындарды анықтау қолда бар смета негізінде жүргізіледі, ол мынадай элементтерді қамтиды:

- материалдық шығындар;
- еңбекақы төлеу шығындары;
- әлеуметтік салық;
- негізгі қорлардың амортизациясы;
- өзге де шығындар.

Материалдық шығындар негізгі және қосалқы шығындарға, материалдарға, энергияға және БҚ әзірлеу үшін қажетті басқа да шығындарға бөлінеді. Материалдық шығындарды есепте берілген нысан бойынша жүргізіледі (Кесте 4.2).

Кесте 4.2 – Материалдық ресурстарға шығындар

Материал аты	Маркасы	Өлшем бірлігі	Саны	Бір дана үшін бағасы	Сома-сы, теңге
Кеңсе қағазы	Svetocopy	Қорап	2	1400,00	2800,00
Блокнот	Index	Дана	2	590,00	1180,00
Қаламдар	Celllo	Дана	2	140,00	280,00
Маркерлер	Koi	Дана	2	350,00	700,00
Компьютерлік тышқан	Qcyber	Дана	2	5690,00	11380,00
Барлығы:					16620,00

Бағдарламалық қамтамасыз етуді әзірлеу үшін Asus X507UB-EJ501T ноутбук қолданылады. Ноутбук қуаты қойылған міндеттерді орындау үшін жеткілікті. Себебі ноутбукта 4 ядро, микропроцессор CoreI7-8200M және құрамында белгіленген операциялық жүйесі Windows 10 x64 және үшін қажетті бағдарламалық қамтамасыз етуді әзірлеу және өндіруге қосымша шығындар қажеттілігі жоқ ОС бар.

Материалдық құралдарға ($З_м$) қажетті жалпы соманы мынадай формула бойынша есептеуге болады:

$$З_м = \sum P_i * Ц_i, \quad (4.3)$$

мұнда P_i - материалдық ресурстың і түрінің шығысы, заттай бірліктер;

$Ц_i$ - материалдық ресурстың і түрінің бірлігінің бағасы, тг;

і - материалдық Ресурстың түрі;

n - материалдық ресурстар түрлерінің саны.

Қажетті жабдықтар мен бағдарламалық қамтамасыз ету шығындарын есептеу 4.3-кестеде келтірілген нысан бойынша жүргізіледі.

Кесте 4.3 – Жоба үшін қажетті жабдық пен БҚ шығындарын есептеу

Материалдың атауы	Маркасы	Өлшем бірлігі	Саны	Бір дана үшін бағасы	Сомасы, теңге
Ноутбук	Asus X507UB-EJ501T	Дана	2	200 000,00	400 000,00
Принтер	Epson L-120	Дана	1	43 550,00	43 550,00
Хостинг	PS.kz	Дана	2	1 800,00	3 600,00
Модем	Ericsson T073G	Дана	1	14 000,00	14 000,00
ОЖ	Windows 10	Дана	2	-	-
Домен	PS.kz	Дана	1	3 338,00	3 338,00
Итого:					464 488,00

Зм = 16620,00+ 464 488,00= 481 108,00 (тг)

Бағдарламалық қамтамасыз етуді іске асыру үшін 481 108 теңгеге материалдар қажет.

4.3 Электр энергиясына шығындарды есептеу

Электр энергиясын тұтынбай бағдарламалық қамтамасыз етуді әзірлеу мүмкін емес болғандықтан электр энергиясына жұмсалатын шығындарды есептеу қажет.

4.1 кестеге сүйене отырып, бағдарламалық қамтамасыз етуді әзірлеу үшін шамамен 232 сағат қажет, енді 232 сағат ішінде жұмсалатын электр энергиясының құнын есептеу қажет. Принтер үшін есептеу 24 сағат кезеңі үшін жүргізіледі, себебі принтерді үнемі пайдалану қажет емес.

$$Э = Ш_{эл.эн.құрал.} + Ш_{қос.шығ.}, \quad (4.2)$$

мұндағы $Ш_{эл.эн.құрал.}$ - жабдықтың электр энергиясына арналған шығындар;

$Ш_{қос.шығ.}$ - қосымша мұқтаждықтарға электр энергиясының шығындары.

Жабдық үшін қажетті электр энергиясын есептеу мынадай формула бойынша анықталады:

$$З_{эл.эн.обор.} = \sum W * К_{исц} * S * T, \quad (4.3)$$

мұндағы W-тұтынылатын қуат, Вт;

$К_{исц}$ - пайдалану коэффициенті ($К_{исц}$ = 0,7..0,9);

Т-жұмыс уақыты;

S-тариф (1кВт / сағ = 18,32 тг).

Электр энергиясының құнын есептеу бойынша қорытынды (Кесте 4.4).

4.4 кесте - Электр энергиясына шығындар

Құрал атауы	Төлқұжат қуаты, кВт	Қуаттылық коэффициенті	Құрал жұмыс істеу уақыты, сағ	Баға ЭЭ тг/кВтч	Сома, тг.
Ноутбук	0,6	0,7	231	18,32	1777,41
Модем	0,08	0,8	231	18,32	270,84
Принтер	0,5	0,9	32	18,32	263,81
Кондиционер	0,8	0,8	180	18,32	2110,46
Жарықтандыру	0,3	0,7	231	18,32	888,71

$$З_{эл.эн.құрал} = 5311,23(\text{тенге})$$

Қосымша қажеттіліктерге шығыстар электр энергиясына арналған шығыстардың 5% көлемінде жоғары көрсеткіш негізінде есептеледі:

$$З_{қос.шығ} = 5\% * З_{эл.эн.құрал.}, \quad (4.4)$$

Формулаға (4.4) сәйкес қосымша қажеттіліктерге жұмсалатын шығындарды анықтаймыз:

$$З_{қос.шығ.} = 0,05 * 5311,23 = 265,56(\text{тенге})$$

Барлық есептеулерге сүйене отырып, электр энергиясына толық шығындар құрайды:

$$Э = 256,56 + 5311,23 = 5567,79(\text{тенге})$$

4.4 Еңбекақы төлеу шығындарын есептеу

Бағдарламалық қамтамасыз етуді әзірлеу үшін бұрын көрсетілгендей, екі қызметкер қажет:

- жоба жетекшісі – жұмыс уақытын басқару, жұмыс процестерін түзету, үйлестіру, пәндік облысты зерттеу;
- әзірлеуші – БҚ әзірлеу, тестілеу және сүйемелдеу.

Еңбекақы төлеу шығындарының сомасын келесі формула бойынша есептеуге болады:

$$З_{тр} = \sum ЧС_i * T_i \quad (4.5)$$

мұндағы $СМ_i$ - і қызметкердің сағаттық мөлшерлемесі, тг;

T_i - модельді әзірлеудің еңбек сыйымдылығы, адам*сағ; і-қызметкердің санаты;

n – бағдарламалық продукт әзірлеумен айналысатын қызметкерлердің саны.

Жұмыс уақыты әр түрлі, сондықтан әрбір қызметкердің сағаттық ставкасын және жалпы жалақы көлемін белгілеу қажет.

Қызметкердің сағаттық мөлщерін келесі формула бойынша есептеуге болады:

$$СМ_i = \frac{Ж_i}{ЖУҚ_i} \quad (4.6)$$

мұндағы $Ж_i$ - і-ші қызметкердің айлық жалақысы, тг;

$ЖУҚ_i$ - і жұмыс уақытының айлық қоры, сағат.

Жетекшінің айлық жалақысы 250 000 теңгеге тең және әзірлеушінің айлық жалақысы 160 000 теңгеге тең. Әр қызметкердің сағаттық мөлшерін (4.6) формулаға сәйкес есептейміз:

$$СМ_{жетекші} = \frac{250000}{22 * 8} = 1420{,}45 \text{ тг/сағ}$$

$$СМ_{әзірлеуші} = \frac{160000}{22 * 8} = 909{,}09 \text{ тг/сағ}$$

Жетекшінің сағаттық мөлшері 1420,45 (тг/сағ) құрайды, еңбек сыйымдылығы 150 сағатқа тең. Әзірлеушінің сағаттық мөлшерлемесі 909,09 (тг/сағ), әзірлеудің еңбек сыйымдылығы 231 сағатқа тең. (4.5) формулаға сәйкес қызметкерлердің еңбекақысына арналған шығындар сомасын есептеуге болады:

$$З_{тр} = 1420{,}45 * 150 + 909{,}09 * 231 = 213067{,}5 + 209999{,}79 = 423067{,}29$$

Еңбек ақы төлеу бойынша шығындарды есептеу көрсетілген (Кесте 4.5).

Кесте 4.5 – Жалақыны есептеу

Қызметкердің санаты	Квалификация	Еңбек сыйымдылығы БП, сағ.	Сағаттық мөлшер, тг/сағ	Сумма, тг.
Жетекші	Проект жетекшісі	150	1420,45	213067,50
Разработчик	Бағдарламашы	231	909,00	209999,79
Итого:				423067,29

4.5 Әлеуметтік салық бойынша шығындарды есептеу

Қазақстан Республикасының Салық Кодексіне сәйкес әлеуметтік салық еңбекақы төлеу қорының 9,5% - ын құрайды. Әлеуметтік салықты келесі формула бойынша есептеуге болады:

$$Ә_с = (ЕТҚ - ЗА) * 0,095, \quad (4.7)$$

бұл жерде ЗА - зейнетақы қорына аударымдар ЕТҚ-ның 10% құрайды.

$$ЗА = 423067,29 * 0,1 = 42306,73 \text{тенге}$$

$$Ә_С = (423067,29 - 42306,73) * 0,095 = 36172,18 \text{тенге}$$

Есептеу нәтижелері кестеде берілген (4,6):

Кесте 5.6 - Әлеуметтік салықты есептеу

Қызметкер санаты	Адам саны	Айлық мөлшері, тг	Зейнетақы аударымы, тг	Әлеуметтік салық, тг
Жетекші	1	213067,50	21 306,00	18217,20
Әзірлеуші	1	209999,79	20 999,98	17954,98
Барлығы:				36172,18

4.6 Негізгі қорлардың амортизациясы және өзге де шығындар

Амортизация нормаларын НҚ салық кодексіне сәйкес анықтау қажет. НҚ амортизациясын келесі формула бойынша анықтауға болады:

$$A_г = \frac{Ж_{құны} * Н_а}{100} (5.8)$$

мұндағы, С_об – жабдықтың құны;

$Н_а$ – амортизация нормасы (амортизация нормасы = 25);

Формула (5.8) ноутбук үшін бір жыл ішінде амортизациялық аударымдар үшін қажетті соманы есептеуге мүмкіндік береді:

$$A_г = \frac{400000 * 25}{100} = 100000 \text{тенге}$$

Енді әзірлеу кезеңі үшін амортизация нормасын есептеу қажет:

$$A_{\text{г}} = \frac{100000 * 29}{365} = 7945{,}21\text{тенге}$$

Осылайша, барлық жабдық үшін амортизация нормасын есептеу қажет. Есептеу нәтижелері кестеде келтірілген (Сурет 4.7).

Кесте 4.7 – БҚ аммортизациясы

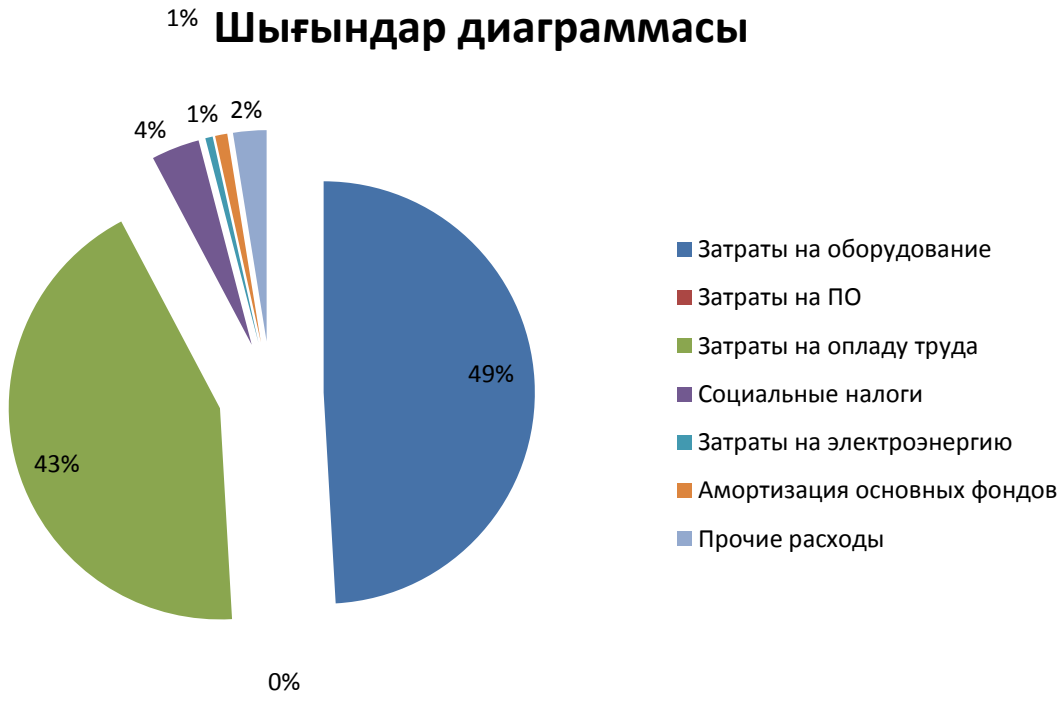
Жабдық атауы және БҚ	Жабдықтар мен БҚ құны, тг	Жылдық амортизациясы, %	Жылдық амортизациясы сомасы, тг	Әзірлеу кезіндегі амортизация сомасы, тг
Ноутбук	400 000	25	100000,00	7945,21
Принтер	43 550	25	10 887,50	865,03
Модем	14 000	20	2 800,00	222,46
Хостинг	3 600	20	720,00	57,21
Домен	3 338	15	500,70	39,78
Итого:			114908,20	9129,69

БҚ әзірлеуге арналған шығыстар сметасы.

Барлық берілген есеп-қисаптардың негізінде (5.8) кестеде келтірілген нысан бойынша әзірлеуге арналған шығыстар сметасын ресімдеу қажет. Жұмыс шығыстарының диаграммасы көрсетілген (Сурет 4.1).

Кесте 5.8 – БҚ әзірлеуге арналған шығындар сметасы

Шығындар	Сумма, тг
Құралға кеткен шығындар	481 108,00
Айлыққа кеткен шығындар	423067,29
Әлеуметтік салық	36172,18
Затраты на электроэнергию	5567,79
Негізгі қор амортизациясы	9129,69
Смета бойынша жиыны:	955 045,00



Сурет 4.1 – Шығындар диаграммасы

4.7 Ықтимал бағаны анықтау.

Бағдарламалық қамтамасыз етудің құны әзірленген өнімнің сапасы, оны әзірлеу мерзімі және өнімнің өнімділігі негізінде анықталады. Бағдарламалық қамтамасыз ету үшін $Б_{ш}$ құнын мына формула бойынша есептеуге болады:

$$Б_{ш} = Ш_{толық}\left(1 + \frac{P}{100}\right), (5.9)$$

где $З_{нир}$ – затраты на разработку программного обеспечения, тг;

Р – средний уровень рентабельности ПО, (%). Данный параметр принят равным 25%.

$$Б_{Ш} = 955045\left(1 + \frac{25}{100}\right) = 955045 + 238761{,}25 = 1193806{,}25\text{тенге}$$

Далее необходимо определить стоимость реализации с учетом НДС, ставка НДС устанавливается законодательством РК. На 2019 года ставка НДС составляет 12%. Стоимость реализации учитывая НДС можно рассчитать по следующей формуле:

$$Б_{р} = Б_{ш} + Б_{ш} * ҚҚС, (4.10)$$

$$Б_{р} = 1193806{,}25 + 1193806{,}25 * 0{,}12 = 1337063{,}00\text{тенге}$$

5 Өмір тіршілік қауіпсіздігі

5.1 Электрмагниттік өрісінің қауіпі және зиянды факторлары

Электрмагниттік өріс – электр заряды не магниттік моменті бар бөлшектердің өзара әсерін жеткізетін (тарататын) физикалық өрістің бір түрі. Ол координаттардың екі векторлық функциясы – электр өрісінің кернеулігі (Е) мен магнит өрісінің кернеулігі (Н) кейде магнит индукциясы (В) арқылы сипатталады. Электрмагниттік өрістің дербес түріне қозғалмайтын электр зарядының жасайтын таза электр өрісі және тұрақты тогы бар қозғалмайтын өткізгіштердің не тұрақты магниттің жасайтын таза магнит өрісі жатады. Алайда бұл өрістердің әрқайсысы басқа бір инерц. санақ жүйесімен салыстырғанда таза электр өрісі не таза магнит өрісі бола алмайды. Электрмагниттік өріс электр және магнит өрісі деп шартты түрде ғана бөлінеді. Бір-біріне қарағанда әр түрлі инерц. санақ жүйесінде қозғалатын Электрмагниттік өрістің белгілі бір нүктесіндегі Е мен Н векторларының мәні әр түрлі болады. Айнымалы электр және магнит өрістері бір-біріне тығыз байланысты, олар біріге отырып айнымалы Электрмагниттік өрісті құрайды. Қозғалмайтын ортадағы Электрмагниттік өрістің заңдары Максвелл теңдеуімен сипатталады. Электрмагниттік өріс кеңістіктің барлық бағытында $3*10^8$ м/с жылдамдықпен электромагниттік толқын түрінде тарайды.

Электромагниттік толқындар кез келген үйде, мекемеде жалпы адам өмір сүретін барлық ортада бар.Электромагниттік толқындарды тұрғын үйдің теледидарында, өтегінде, мұздатқыштарында, микротолқынды пеште, шаңсорғыштарда, компьютерде ұялы телефондарда болады.

Электромагниттік толқындардың 1000 мГц таралу керек болса, электро техникалық құрылғыларда кейде одан көп асып кетеді. Мысалы: өтекті қосқан кезде 25см қашықтықта одан 0,2мкТл электромагниттік толқындар бөлінеді. «Tefal» шәйнегінде 20см-0,6мкТл. Кір жуғыш машина 50Гц, ал теледидардың пульті 1 метрде 1мкТл болады. Микротолқынды пеште 30см қашықтықта 50Гц магниттік өрісі 10,3-8мкТл болады. Ал ер адамдар электр қыздырғыш қолданған кезде өздерін әдемілей отырып, беттерінен ток жүргізеді себебі: Электр қырынғыш 100мкТл (микро Тесла) есептеледі екен. Компьютерде 60Гц болады.Бірінші күн радиациясы жайлы айтып өту керек.

Электрмагниттік радиация электрмагниттік толқындар түрінде, 300000 км/с жылдамдықпен тарап, жер атмосферасына енеді. Жер бетіне дейін тура (жерге бұлтсыз ашық жағдайда атмосферадан көктей өтіп жететін Күн сәулелері) және шашыранды (атмосферадағы шаң-тозаңнан, бұлттан шашыраған Күн сәулесі) радиация түрінде жетеді.Радиацияның ағзаға беретін энергия мөлшері сәулелену дозасы деп аталады.«Күн өтіпті »деген халық диагностикасы мен «сәулелік ауру» деген қазіргі медицина диагностикасы арасында тура байланыс бар.Жаздың ыстық күндерінде білмеген адамға ерсі көрінгенмен,өзбек пен тәжіктің ала шапан киюінде,қырғыз бен түрікменнің ақ киіз қалпағы мен елтірі бөрігін,дала қазағының түйежүн шекпенін тастамауында,халықтың радиациядан қорғануының ғасырлық тәжірібесі

жатыр.Күні шуақты елдердегі әйелдердің бетін,денесін бүркеп жүруінің де бір сыры осында жатыр.

Табиғи және жасанды радиоактивті изотоптарда ядролардың өздігінен ыдырау процесі үздіксіз жүріп жатады. Демек, олар сыртқы ортаға туынды белшектерді, гамма кванттарын үнемі атқылаумен болады. Радиоактивті сәулелер кейде радиация немесе иондағыш сәулелер деп аталады. Олардың кинетикалық және электромагниттік энергиялары үлкен шама құрайды. Сондықтан ондай бөлшектер жолындағы денелердің атомдары мен молекулаларының химиялық-физикалық касиеттерін өзгертіп иондайды, олардың араларындағы қалыпты байланыстарды үзеді. Сөйтіп, биологиялық денелер де, басқа табиғи денелер де өзгеріске ұшырайды. Әсіресе тірі табиғат: адам мен жан-жануарлар, өсімдіктер мен басқа да тіршілік иелері зор зардап шегеді.[1]

Атом бомбалары мен уран кеніштерін айтпағанның өзінде, атомдық реакторлар мен атомдық электр станциялары да радиацияның көзі болып табылады. Сондай-ақ Күн радиациясының, ғарыштан келетін басқа да бөлшектердің зиянды әсерін де білуіміз қажет. Ол үшін изотоптардың сәуле атқылау белсенділігін, сондай-ақ радиацияға душар болған денелердің алған сәулелерінің мөлшер-дозасын нақты білу қажет. Қандай доза шегінде жұмыс істеуге болады, қандай доза денсаулыққа зиян немесе адам өміріне қауіпті деген сұрақтарға да жауап беруіміз керек.

Иондағыш сәулелерден қорғана білу үшін олардың өтімділік қасиеттерін білген жөн. Радиоактивті изотоптармен жұмыс істегенде, олардың өтімділігіне орай тиісті қауіпсіздік ережесін бұлжытпай орындау керек.

Альфа-бөлшек парақ қағазға тұтылып, одан өте алмайды. Алайда адам терісінде қалып қойса немесе ішкі органдарына тыныс жолымен, яғни жеген тағамы арқылы етіп кетсе, өте қауіпті.

Бета-бөлшектердің өтімділік қабілеті үлкен. Олар адам ағзасына 1–2 см тереңдеп ене алады. Алайда бірнеше миллиметр алюминий қаңылтыры оны толық жұтып алады.

Гамма-сәуленің өтімділік қабілеті аса күшті. Сондықтан одан қорғану үшін корғасынның немесе бетон плиталардың калың қабаты пайдаланылады.

Ал қоғамдағы, күнделікті біздің өміріміздегі электромагниттік сәулелену дегеніміз ол күні бойына көз алдымызда болатын тұрмыстық техникамен құралдар, қолымыздан түспейтін ұялы телефон және т.б. Сырт көзге байқалмаса да бұлар адам ағзасына белгілі бір мөлшерде зиян келтіруде.

Жердегі кез-келген ағза сияқты адамның денесінде де өзіндік электромагниттік өріс болады, ол ағзаның барлық жасушаларының үйлесімді жұмыс істеуін қамтамасыз етеді. Адамның электромагниттік сәулесін биоөріс (көрінетін бөлігі – аура) деп атайды. Бұл өріс ағзаны кез-келген негативті әсерлерден қорғайтын негізгі қабықша болып табылады.

Адамның биоөрісі бұзылған жағдайда ағзамыздың мүшелері мен жүйелері кез-келген ауру факторының тууына жол береді. Егер бізге электромагниттік өрістен басқа сәулелену көздері әсер етсе, яғни біздің

биоөрісіміздің қарқындырақ, онда ағзамыз бейберекет күйге тап болады. Бұл дегеніміз денсаулықтың түбегейлі нашарлауының бастамасы.

Бірқатар елдерде жүргізілген клиникалық зерттеулер электромагнитік өріспен ұзақ уақыт ұштасу «радиотолқынды ауру» деген атқа ие аурудың дамуына алып келетінін көрсетті. Бұл аурудың клиникалық сипаты, ең алдымен жүйке және жүрек-қантамырлары жүйесінің қызметі нашарлап, бастапқы күйінен ауытқиды. Ұзақ уақыт аралығында сәулелену аймағында болған адамдарда келесідей белгілер байқалады:

- Әлсіздік;
- ашуланғыштық;
- тез шаршау;
- есте сақтау қабілетінің төмендеуі;
- ұйқының бұзылуы;
- жүйке жүйесінің вегетативті қызметінің бұзылуы;
- гипотония;
- жүректің ауруы;
- тамыр соғысының бұзылуы;
- мазасыздану;
- есте сақтауы және зейін қоюы бұзылады [2].

Жүйке жүйесіне әсері тіпті жылулық әсер байқалмайтын электромагнитті сәулелену деңгейі ағзаның ең маңызды деген қызметтік жүйелеріне әсер етеді. Осы саладағы мамандардың көпшілігінің ойынша жүйке жүйесі ең осал ағза болып табылады. Әсер ету механизмі өте қарапайым – анықталған, электромагнитті өріс кальций иондары үшін торлы мембрана өткізгіштігін бұзады. Нәтижесінде жүйке жүйесі қалыпты қызметінен ауытқи бастайды. Аталған процесстер барысында туындайтын ауытқулар ауқымы кең – жүргізілген тәжірибелер барысында есте сақтау қабілетінің төмендеуі, реакцияның төмендеуі, депрессиялық өзгерітер және т.б. сынды құбылыстар тіркелген.

Имундық жүйеге әсері Имундық жүйеде әсер ету аймағына ұшырайды. Бұл бағыттағы тәжірибелік зерттеулер ЭМС сәулелендірілген жануарларда инфекциялық процесс сипаты өзгеретінін көрсеткен (инфекциялық процесстің жүруі ауырлайды).

ЭМС әсер еткен кезде соңы жойылуға әкеп соғатын иммуногенез процессі бұзылады. Бұл процессті аутоиммунитет туандауымен байланыстырады.

Жоғары қарқындылықтағы электромагниттік өрістің ағзаның имун жүйесіне әсері имунитеттің торлы Т-жүйесінің жойылу эффектісінен байқалады.

Эндокринді жүйеге әсері Эндокринді жүйе де электромагниттік сәулеленуге ұшырайды. Зерттеулер электромагниті өрістің әсер етуі кезінде гипофизарлы-адреналинді жүйенің стимуляциясы болатынын және ол қандағы адреналин көлемінің артуымен қатар жүретінін көрсетті.

Жүрек қан тамырлар жүйесіне әсері ЭМӨ әсер ету нәтижесі ретінде жүрек қан-тамырлар жүйесі қызметінің бұзылуын қарастыруға болады. Ол артерия қысымының және тамыр соғысының тұрақсыздығынан байқалады. Периферлік қан құрамының фазалық өзгеруі белгіленеді.[3]

Жыныс жүйесіне Туа бітті кемтарлық пен кемістік жағдайларының көбеюі, қыз жынысты балалардың дүниеге келу ықтималдығының артуы, спермакинездің жойылуы байқалады.

5.2 Электрмагниттік өрісінің адамға әсері және қорғану шаралары

Электрмагниттік өрістерінің кез келген өзгерістері қоршаған кеңістікті сүзіп өтетін күш сызықтарының өзгеруін туындатуы керек, яғни ортада таралатын импульстар (немесе толқындар) болу керек. Осы толқындардың таралу жылдамдығы ортаның магниттік және диэлектриктік өтімділігіне тәуелді болып, электромагниттік бірліктің электростатикалық бірлікке қатынасымен анықталады. Компьютерлік техниканың пайда болуымен қарыштап дамуы қоршаған ортада электромагниттік ахуалдың өзгеруіне алып келді. Компьютерлік техника электромагниттік өрістің сәулелену көзі болып табылады, ал ол өз кезегінде адам денсаулығына қауіп төндіретіні белгілі. Компьютерлік жұмыс орындарының дұрыс ұйымдастырылмауы электромагниттік өрістің адам денсаулығына кері әсер етуіне әкеліп соғады. Компьютердің ең қауіпті бөлігі-монитор.Ол адам ағзасын сәулелендіреді. Медицина өкілдерінің зерттеулері электромагниттік өрістің әсері метаболизм мен электромагниттік бұзылуынаықпал ететіні дәлелденген. Ол сонымен қатар жүйке жүйесі, жүрек қан тамырлары, эндокринді жүйе жұмысында да пайда болуына әсер етеді.Қалта телефонын пайдаланған кезде одан бөлінетін электромагниттік өріс тұтынушының миына әсер ететіні анықталған. Электромагниттік сәуледен адамға әсер ету деңгейі сәуленің интенсифтілігіне, жиілігіне және әсер ету уақытына байланысты. Үлкен интенсифтік өрістің адамға ұзақ уақыт әсер беруі адамның күйзелу күйіне, шаршаңқы болуына,ұйқыға әуестігіне, ұйқының бұзылуына, бастың ауыруына, гипертонияға, жүрек тұсындағы ауырлыққа әкеліп соқтырады. Өте жоғары жиілікті өрістің әсері адамның қан құрамының өзгерісіне,көздің ауруына әсер етеді.

Көп ғалымдар ұялы телефонның электромагниттік толқындарының шынында да адам ағзасына қаншалықты зияны бар екенін зерттеп қараған. Көп елдерде осыған байланысты арнайы программалар құрылып, үш жылдың көлемінде зерттеулер жүргізілген. Оған 12 елден келген 15 мың дәрігер мамандар қатысқан. Дәрігерлер зерттеу нәтижесінде құлақтың маңайында орналасқан (ми, сілекей безі, құлақ түйсіктері, көз)-дерге ұялы телефоннан шығатын электромагниттік энергияға көп көңіл бөлген. Осы сұрақ бойынша мен біздің Шет ауданындағы емханаға барып, лор дәрігерімен сауалнама өткіздім. дәрігер шынында да электромагниттік толқындардың адам ағзасына , соның ішінде құлақтың маңайында орналасқан мүшелерге көп әсерін тигізетіні туралы айтты.

Шынында да ұялы телефонды тым жақын қолданғанда одан электрмагниттік энергия шығады, тура сондай энергия микротолқынды пеште тауық етін пісіргенде шығады. Бұл энергия адамның миына және басқада мүшелеріне әсер етеді. Ұялы телефонды балаларға ұстап, онымен ойнауға болмайды. Шведтық ғалымдар «2мин. артық сөйлескен кезде адамның басында шу пайда болады» -деген тұжырымдама жасады. Ал Россияның ғалымдарының зерттеу барысында NMT-450, AMPS-800 ұялы телефондарымен сөйлескен кезде адамның миының биоэлектрлік жылдамдығына әсер ететіні байқалған. Адамның миы электромагниттік толқындардың сәулелерін әртүрлі қабылдайды. Электромагниттік сәулелер адам ағзасына өте қатты зиян тіпті кейде өлімге де әкеп соқтыруы мүмкін.

Ресейлік Федерация ғылым академиясының институты микротолқынды пешке зерттеу жасаған екен. Микротолқынды пештің ішінде тағамға зерттеу жасаған, онда «С» витамині сақталып 75-98% пайыз құраған, ал күнделікті өмірде витамин 30-60% пайызды құрайды екен [4].

Тағамды жәй жылытатын болса, онда ол тағамның дәмі азайып, микрофлорасы да азаяды. Тағамды сусыз немесе аз көлемді суда пісірсе, онда тағамның ауыр металлдары мен нитраттары, нитриттері қайда кетеді?

Американдықтардың зерттеулері бойынша микротолқынды пештің арқасында «асқазан ракын» болдыртпауға болады. Себебі онда май мүлдем қолданылмайды. Ондай тағамды асқазанға пайдалы деп есептейді екен.

Ал испандық зерттеушілер керісінше микротолқынды пеште пісірілген брокколи қырыққабатының 98% пайызының витамині және микроэлементі жойылғанын байқаған.

Ең алғашқы микротолқынды пештің ішінде пісірілген тағамға швейцарлық биолог Хортеля және профессор Бернарда Блонка 1989 ж зерттеу жасаған екен, осы зерттеу барысында бір адамға пеште істелінген тағам жегізген, ал кейін микротолқынды пеште істелінген тағамды жегізген, сонда микротолқынды пеште істелінген тағамды жеген соң адамның қанының құрамында лейкоцит тым көп болып көбейіп, рак ауруына шалдыға бастағанын байқаған. Осы ауруды микротолқынды пеште істелген тағам тудырған, бірақ бұған ешкім мән бермеген екен.

Микротолқынды пеште істелген тамақтың электромагниттік толқындары микротолқынды пеште істелген етке судың молекуласының құрамымен енеді. Осыдан молекулалар қызып, бір-бірімен қақтығысады. Осыдан температура жоғарлайды, микротолқынды энергия 2,5см ден 5см ге дейін өтеді, сондықтан қалың етті ортасынан бөліп салған жөн.

Ұялы телефонның электромагниттік толқындарының адам ағзасына зияны қандай болса, микротолқынды пештің электромагниттік толқынының зияны да тура сондай болады.

Францияның, Россияның, Украинаның және Швецияның ғалымдарының зерттеуі бойынша электромагнитті ақпаратты (торс) толқын ақпаратты кеңістік болып табылады. Осы ақпаратты кеңістіктің адам денсаулығына

тигізетін зияны өте көп екен. Адамның бас ауруы, ұйқының бұзылуы, мазасыздануы осының әсерінен болады.

Биологиялық зерттеу бойынша, адамның ағзасына микротолқынды пештен бөлінетін жоғарғы жиілікті диапазонның сантиметрлік сәуле шашуы әсер етеді. Электромагниттік сәуле шашуды көріп, естуге болмайды. Бұндай сәуле шашу тез байқалмайды. Сәуле шашу микротолқынды пеште істелген тағамға да әсер етеді. Осының салдарынан электромагниттік сәуле шашудан тағамның құрамындағы молекула ионизацияланып, атом электронын жоғалтады немесе керісінше қосып алады. Бұндай жағдай тағамның құрылымын өзгертеді. Сәуле шашу тағамдағы молекуланың бұзылуын тудырады.

Микротолқынды пеште өмірде қолданылмайтын жаңа атау еңгізілген- «радиолитикалық».Радиолитикалық қосылым молекулалық радиация туғызады.

Микротолқынды пеште істелген құрамында nitrosodienthanolamines к онцерогені болады. Сүттегі кейбір аминақышқылдар концерогенге айналған. Микротолқынды пеште қатып тұрған жеміс-жидекті жібіткен кезде, олардың глюкозидтары және галактозидының құрамында концерегенді элементі бар ұсақ бөлшекке айналады екен. Тағамның құрамы 60% пайыздан 90% пайызға дейін азаяды. В, С және Е витаминдерінің биологиялық қасиеті жоғалады.

Электромагнитік өрістің әсері - электр заряды не магниттік моменті бар бөлшектер арасындағы электромагнитік өріс арқылы берілетін белгілі . Адам өмірге келгеннен бастап, электромагнит сәулесінің әсерінде болады. Адамға әсер ететін жердің магниттік өрісі - табиғи электромагниттік өріс, планетарлық сарқылмайтын ресурс. Магниттік өрістің күші әржерде әртүрлі. Радиожиіліктік өрістер адам организіміне қолайсыз әсерін тигізеді. Адамға, жануарларға, өсімдіктерге, микроорганизмдерге жер қыртысынан бөлінетін гамма сәулелер және ғарыш сәулелері сырттан,организмде болатын радиоактивті элементтер сәуелері іштен әсер етеді. Егер бұл сәулелер тірі организмге артық мөлшерде өтсе, клеткалардың, органдардың тіршілігіне қауіпті ауру жабысады. Радиожиілікті қондырғылар шығаратын электромагниттік сәулелерді мөлшерден көп қабылдаған жағдайда ол адамда мамандық ауруға әкеліп соғады. Нәтижесінде нерв жүйесі журек қан тамырлары эндокриналды жүйе және де басқа да ағзаларға әсер етуі мумкін. Электромагниттік өріс әсерінде ұзақ уакыт болған жағдайда адамдар тез шаршайды. Ұйқышылдық пайда болады,ұйқысы бұзылады, жиі- жиі басы ауырады, нерв жүйесі бұзылады т.с.с. системетикалық сәулелену болған жағдайда психикалық ауру қан қысымы өзгеру жүрек соғысының баяулауы шашының түсуі байқалады. Қорғану әдістері: сәуле шығару көзіндегі сәулеленуді азайту. Өте жоғары жиілікті және ультра жиілікті қондырғыларды дұрыс орнату. Экрандалған бөлмелердегі қондырғыны алыстан бақылау. Жұмыс істеу орнын және сәуленің шығу көзін экрандау немесе мыстан жасалтын жогары өткізгіштік қасиеті бар тор металдар шағылдырғыш жерлету экран ретінде пайдалану шаралар электормагниттік сәулеленуді

дозиметр көмегімен кемінде айына бір рет тексеру, жылына медициналық тексеруден бір рет өткізу. Қосымша демалыс қысқартылған жұмыс күнін жасау жасы он сегізге толмаған және орталық нерв жүйесі жүрегі, көзі ауыратын тұлғаларды жұмысқа қабылдамау.

5.3 Электрмагниттік өрісті есептеу

Иондаушы емес электромагниттік сәулелену мен өрістерге оптикалық және радиожиілікті диапазонындағы электромагниттік сәулеленулерді, сонымен қатар шартты-статикалық электрлік және тұрақты магниттік өрістерді жатқызу қабылданған.

Электромагниттік сәулеленулер (ЭМС) толқын ұзындығымен - λ(м), тербеліс жиілігімен - **f**(Гц) және таралу жылдамдығымен -V (м/с) сипатталатын элктромагниттік толқындар түрінде таралады. Бос кеңістікте ЭМС таралу жылдамдығы жарық жылдамдығына тең - $C=3\cdot10^8$. м/с

Келтірілген параметрлер бір-бірімен келесі қатынаспен байланысады:

$$\lambda = \frac{c}{f} (1)$$

Ағзаға әсер ететін факторлардың бұл тобына:

· табиғи иондаушы емес электромагниттік сәулеленулер мен өрістер;

· статикалық электрлік өрістер;

· тұрақты магниттік өрістер;

· электромагниттік сәулелену мен өнеркәсіптік жиіліктегі және радиожиілікті диапазонындағы өрістер;

· лазерлік сәулелену жатады.

Өндіріс жағдайында адамға аталған өрістер мен сәулеленулердің соңғы төрт түрі әсер етеді.

Табиғи иондаушы емес сәулеленулер мен өрістер салыстырмалы түрде жақын арада зерттеле бастады және соңғы он жылдықтарда жерде тіршілік пайда болуында, одан әрі дамуы мен реттелуінде олардың маңызды ролі дәлелденді. Табиғи электромагниттік өрістердің спектрін шартты түрде бірнеше құрам бөліктеріне бөлуге болады. Олар жердің тұрақты магниттік өрісі, немесе геомагнитті өріс (ГМӨ), және 10^{-3} -нен 10^{12} Гц дейінгі жиілік диапазонындағы электростатикалық өріс пен айнымалы электромагниттік өрістер.

Табиғи электромагниттік өрістер, оның ішінде геомагнитті өріс ағзаға әр түрлі әсер етуі мүмкін. Бір жағынан, геомагниттік ауытқулар экологиялық қауіп-қатер факторы ретінде қарастырылады – биологиялық ырғақтардың, мидың функционалдық жағдайы модуляциясының синхрондылығын бұзады, клиникалық тұрғыдан ауыр медициналық патологиялар (миокард инфарктысын, инсульттерді, жол-көлік жағдайлары мен апаттарын, соның ішінде әуе апаттарының) санының өсуіне себеп болады. Басқа жағынан, ГМӨ-нің кезеңдік емес түрінің циркадтық, инфрадтық және циркосептадтық

биологиялық ырғақтармен және олардың ара қатынастарымен байланысы анықталған.

Ағзаға тек магниттік толқу (буря) ғана қолайсыз әсер етіп қоймайды, сонымен қатар, адамның әлсіз электромагниттік өріс жағдайында ұзақ уақыт болуы факторы да, оның ішінде, жұмысы экрандалған бөлмелер мен құрылыстарда істелетін бірқатар өндірістерде де қолайсыз әсер етуі мүмкін. Мұндай жағдайда жұмыс істеушілер жиі көңіл-күйінің және денсаулығының нашарлағанына шағымданады, бұл-геомагниттік өрістің әсерін зерттейтін гигиенаның жаңа бағытының пайда болуына негіз болды. Геомагниттік өрістің төмен деңгейі тек экрандалған құрылыстарда ғана емес, сонымен қатар метрополитеннің жерасты құрылыстарында (2-5 есе), темір-бетонды конструкциялардан салынған ғимараттарда (1,3-2,3 есе), тез жүретін лифт кабинасында (15-19 есе), бұрғылау қондырғылары мен эксковаторлардың кабиналарында, жеңіл автокөліктердің салонында (1,5-3 есе) және басқаларда да байқалады.

Гипогеомагниттік өрістің орталық жүйке жүйесіне (негізгі жүйкелік үрдістердің дисбалансы, ми тамырларының дистониясы, жауап реакциясы уақыттының ұзаруы), вегетативтік жүйке жүйесіне (тамыр соғуымен, артериалды қысымның лабильділігі, гипертензивті типті нейроциркуляторлық дистония, миокардтың реполяризация үрдісінің бұзылуы), иммундық жүйеге (Т-лимфоциттердің жалпы санының, IgG және IgA концентрацияларының азаюы, IgE концентрациясы жоғарылауы) әсер ететіні анықталған.

Әлсіз геомагниттік өрістің әсерін регламенттейтін гигиеналық ұсыныстар жоқ. Берілген жергілікті жерге тән геомагниттік өріс деңгейі, адам үшін қолайлы болып есептеледі.

Статикалық электрлік өрістер (СЭӨ). Бұл-қозғалыста болмайтын электр зарядының өрісі, немесе тұрақты тоқтың стационарлық элетрлік өрісі. Бұлар электрмен, газбен тазалауда, рудаларды және материалдарды электростатикалық жолмен бөліп алуда, сырларды және полимерлі материалдарды электростатикалық жолмен жағу кезінде кеңінен қолданылады. Сондай-ақ диэлектрлік материалдарды жасау өңдеу, тасымалдау кезінде өнімнің электрленуінен электростатикалық зарядтар мен өрістер пайда болатын бірқатар өндірістер мен технологиялық үрдістері (тоқыма, ағаш өңдейтін, целлюлозді-қағаз, химия өнеркәсіптері және басқалары) бар.

Статикалық электрлік өрістің негізгі физикалық параметрлеріне өрістің кернеулігі және жеке нүктелердің потенциалдары жатады. СЭӨ нүктелік зарядқа әсер ететін күштің заряд шамасына қатынасымен анықталады және вольттің метрге қатынасымен өлшенеді (В/м). Статикалық электрлік өрістің энергиялық сипаттамасы өріс нүктелерінің потенциалымен анықталады.

СЭӨ әсер ету жағдайдағы жұмыс істеушілердің денсаулығында анықталатын өзгерістер, әдетте функционалды сипатта болады және астеноневроздық синдромы мен вегетативтік-тамыр дистониясы шеңберінде

болады. Объективті көрінісінде өзіне тән көріністері жоқ, айқындылығы онша емес, функционалдық ауытқулар білінеді. Жұмыс орындарындағы статикалық электрлік өрістің кернеулілігінің шектік рұқсат етілген мәндері жұмыс күні бойына әсер ету уақытына байланысты орнатылады. Жұмыс орындарындағы электростатикалық өрістің шектік рұқсат етілген кернеулігі (E_{ngy}) 1 сағатқа дейінгі уақытта әсер еткенде 60кВ/м артық болмауы тиіс, ал одан да ұзағырақ жұмыс істеген кезде келесі өрнек (формула)бойынша анықталады:

$$E_{ngy} = \frac{60}{\sqrt{t}} \quad (2)$$

мұндағы t – 1-ден 9-ға дейінгі сағатпен берілген уақыт.

Электростатикалық өрістің шектік рұқсат етілген кернеулігін есептеу үшін жұмыс уақытын таңдау қажет.

2 – формулаға сәйкес:

t=8 сағ

$E_{ngy} = \frac{60}{\sqrt{t}} = \frac{60}{\sqrt{8}} = \frac{60}{2.8}$=21.4 в/м

Тұрақты магниттік өрістер. Жұмыс орындарындағы тұрақты магниттік өрістің көздеріне тұрақты магниттер, электрлік магниттер, тұрақты токтың күшті ток жүйелері (тұрақты токтың берілу желісі, электрлік-магниттік ванналар және т.б.) жатады.

Тұрақты магниттер мен электрлік магниттер аспап жасауда, көтергіш крандардың магнитті шайбаларында, магниттік сепараторларда, суды магнитпен өңдеуге арналған құрылғыларда, магниттік гидрадинамикалық генераторларда (МГД), ядролық магниттік резонанс (ЯМР) және электрондық парамагниттік резонанс құрылғыларында, сондай-ақ физиотерапия практикасында кең қолданылады.

Тұрақты магниттік өрісті сипаттайтын негізгі физикалық параметрлері өріс кернеулігі (Н), магнит ағыны (Ф) және магнит индукциясы (В) болып табылады. Магниттік өріс кернеулігінің СИ жүйесіндегі өлшем бірлігі ампердің метрге қатынасы (А/м), магниттік ағынның бірлігі– Вебер (Вб), магниттік ағын тығыздығының (магниттік индукция) бірлігі – тесла (Тл) болып табылады.

ТМӨ 2 Тл дейінгі деңгейлері ағзаға айтарлықтай әсер көрсетпейді. Сонымен қатар, ТМӨ көздерімен жұмыс істейтіндердің денсаулық жағдайындағы өзгерістер болатыны анықталған. Бұл көріністер жиі вегетодистония, астеновегетативті және шеткі вазовегетативті синдромдар немесе олардың қатар жүретін түрінде байқалады. Қан құрамындағы эритроциттер саны мен гемоглобин мөлшерінің төмендеу жағына беталуы, орташа дәрежеде лимфо- және лейкоцитоз болуы мүмкін.

Жұмыс орындарындағы ТМӨ кернеулігі 8 кА/м (10мТл) аспауы тиіс. (1991 ж.) иондаушы емес сәулеленулер жөніндегі Халықаралық комитет ұсынған тұрақты магниттік өрістің рұқсат етілген деңгейлері жұмыс істейтін адамдар, денеге әсер ететін жері және жұмыс уақыты бойынша дифференциаланған. Кәсібі тікелей тұрақты магниттік өріспен байланысты мамандар үшін – толық жұмыс күнінде (8 сағат) әсер еткенде 0,2 Тл; денеге қысқа уақыт әсер еткен кезінде - 2 Тл; қолдарға қысқа уақыттық әсер еткен кезінде - 5 Тл. Халық үшін тұрақты магниттік өрістің үздіксіз әсер ету деңгейі - 0,0011 Тл аспауы керек.

Өнеркәсіптік жиіліктегі электрмагниттік өрістер (ӨЖ ЭМӨ). Соңғы жылдарда жиілігі 50 Гц электрмагниттік өрістер өз алдында жеке диапазонға бөлінген. Олардың негізгі көздеріне айнымалы токтың өндірістік және тұрмыстық электржабдықтарының әр түрлі түрлері, сондай-ақ аса жоғары кернеулі (АЖК) электр берілісінің әуе желілері мен подстанциялары жатады.

Өнеркәсіптік жиіліктегі электрмагниттік өрістердің әсеріне өндіріс жағдайында ұшыраған жұмысшылардың денсаулық жағдайында өзгерістер байқалады. Олар негізінен ағзаның неврологиялық статусындағы өзгерістерді (бас ауруы, жоғары ашушаңдық , тез қажығыштық, салғырлық, ұйқышылдық), сонымен қатар жүрек-тамыр қызметінің бұзылыстарын (тахикардия және брадикардия, артериалық гипертензия немесе гипотония, тамыр тұрақсыздығы, гипергидроз) және асқазан-ішек жолдарындағы өзгерістерді білдіретін шағымдар түрінде болады. Шеткі қан құрамында өзгерістер-орташа дәрежеде тромбоцитопения, нейтрофильді лейкоцитоз, моноцитоз, ретикулопенияға бетбұрыс болуы мүмкін.

Өнеркәсіптік жиіліктегі электрлік өрістердің ШРЕД-і толық жұмыс күні үшін 5 кВ/м деңгейінде орнатылады, ал 10 минуттан аспайтын әсеріне арналған максималды ШРЕД-і 25 кВ/м құрайды, қарқындылығы 5-20 кВ/м аралығындағы рұқсат етілген болу уақыты келесі өрнек бойынша анықталады:

$$T = \frac{50}{E_{ngy}} - 2 \quad (3)$$

мұндағы Т – электрлік өрістің әсерінде болатын рұқсат етілген уақыты, сағатпен

Е – кВ/м берілген бақыланатын зонадағы электрлік өрістің әсер ететін кернеулілігі.

Электрлік өрістің әсерінде болатын рұқсат етілген уақытты табу үшін 2 формуламен есептелінген электр өрістің әсер ететін кернеулігін қолданып табу керек.

3 – формулаға сәйкес:

$$T = \frac{50}{E_{ngy}} - 2 = \frac{50}{21.4} - 2 = 2.3 - 2 = 0.3\text{сағ.}$$

Қорытынды

Бұл жұмыста ВКонтакте әлеуметтік желісіндегі орыс тілді пікірлердің тоналдылықты бағалау бойынша қолданыстағы әдістерге талдау жүргізілді, мұғаліммен Машиналық оқыту әдістерін қолдану үндестік талдауға ерекше назар аударылды. Терең оқытудың нейрондық желілері негізінде деректерді жинау мен зияткерлік талдауды тиімді жүзеге асыруға мүмкіндік беретін python бағдарламалау тіліндегі кітапханалар толық зерттелді және олардың негізінде тиым салынған контентті табу үшін Интернет желісіндегі ақпараттық объекттерді интеллектуалдық талдау жүйесі әзірленді. Мәтінде қазақ рәміздерінің болуын тексеру үшін тұрақты өрнектер негізінде алгоритм әзірленді және Python тіліндегі функция түрінде іске асырылды. Бұл бағдарлама жағымсыз контентті сүзу, құқыққа қарсы әрекеттер және т. б. туралы ақпаратты қамтитын хабарламаларды анықтау үшін кеңінен қолданылуы мүмкін.

Әзірленген жүйеге сынау және реттеу жүргізілді. Сынау үшін Вконтакте әлеуметтік желісінде АЭжБУ тобы пайдаланылды.Жүргізілген тесттер жүйенің тұрақты және тез жұмыс істейтінін көрсетті.

Қысқартулар тізімі

ЖНЖ – жасанды нейрондық желі есептердің кең көлемін шешуге арналған математикалық және алгоритмдік әдістер жиыны.

NLP – neuro-linguistic programming психотерапия мен практикалық психологиядағы академиялық қауымдастық мойындамайтын, қандай да бір салада табысқа жеткен адамдардың вербалды және вербалды емес мінез-құлқын модельдеу техникасына және сөйлеу нысандары, көз, дене қозғалысы мен жады арасындағы байланыстар жиынтығына негізделген бағыт

EM – expectation-maximization ықтималдық модельдер параметрлерінің максималды шынайылығын бағалауды табу үшін математикалық статистикада қолданылатын алгоритм

SVM – support vector machine классификация және регрессиялық талдау есептері үшін қолданылатын оқытушымен ұқсас Алгоритмдер жиынтығы.

LTSM – long term short memory рекуррентті нейрондық желілер архитектурасының түрі.

ӨУ – өмір уақыты өтулердің шекті саны немесе пакет жоғалғанға дейін болуы мүмкін уақыт кезеңі.

CBR - case Based Reasoning кең мағынада белгілі шешімдер негізінде жаңа проблемаларды шешу әдісі болып табылады.

Қолданылған әдебиеттер тізімі

1. Ю. В. Рубцова. Построение корпуса для настройки тонового классифи катора // Программные продукты и системы, – 2015, - №1(109), - 72-78 с.

2. Панг Б. Thumbs up? Sentiment Classification using Machine Learning Techniques / Б. Панг, Л. Ли. – М. : Вильямс, 2002. – 312 с.

3. М. В. Клековкина, Е. В. Котельников, Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики / URL: http://ceur-ws.org/Vol-934/paper15.pdf (қараған мерзімі: 25.05.2019)

4. Регулярные выражения / URL: https://ru.wikibooks.org/wiki/Регулярные_выражения (қараған мерзімі: 21.05.2019)

5. Обработка естественного языка / URL: https://ru.wikipedia.org/wiki/Обработка_естественного_языка

6. Mikolov T. Distributed Representations of Words and Phrases and Their Compositionality //Advances in Neural Information Processing Systems, 2013, – 3111-3119 с.

7. Юсупова Н. И., Богданова Д. Р., Бойко М. В. Алгоритмическое и программное обеспечение для анализа тональности текстовых сообщений с использованием машинного обучения //Вестник Уфимского государственного авиационного технического университета. – 2012. – Т. 16. – №. 6.

8. Пазельская А. Г., Соловьев А. Н. Метод определения эмоций в текстах на русском языке //Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). М.: Изд-во РГГУ. – 2011. – №. 10. – С. 17.

9. Саймон Хайкин. Нейронные сети: полный курс = Neural Networks: A Comprehensive Foundation. – 2-е изд. – М.: «Вильямс», 2006. – С. 1104.

10. Фомин В.И. Экономика информационного бизнеса и информационных систем. Учебное пособие. СПб: Изд-во СПбУУЭ, 2014 -248 с.

11. Mullen T., Collier N. Sentiment Analysis using Support Vector Machines with Diverse Information Sources //EMNLP. – 2004. – Т. 4. – С. 412-418.