ҚАЗАҚСТАН РЕСПУБЛИКАСЫ БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
«Ғ.ДАУКЕЕВ АТЫНДАҒЫ АЛМАТЫ ЭНЕРГЕТИКА ЖӘНЕ БАЙЛАНЫС УНИВЕРСИТЕТІ»
коммерциялық емес акционерлік қоғамы
Телекоммуникациялық желілер және жүйелер кафедрасы

«ҚОРҒАУҒА ЖІБЕРІЛДІ»
Кафедра меңгерушісі Темырканова Э.К, PhD докторы, доцент
(ғылыми дәрежесі, атағы, Т.А.Ж.)
_____ «____» _____2020ж.
(қолы)

ДИПЛОМДЫҚ ЖОБА

Тақырыбы: Телекоммуникация операторының мәліметтерін өңдеу кезінде машиналық оқыту әдістерін қолдану
Мамандығы 5В071900 Радиотехника, электроника және телекоммуникациялар
Орындаған Аменова Шахназ Шухратжановна Тобы РЭТ(ИКТ)-16-1
(Т.А.Ж.)
Ғылыми жетекшісі Жунусов Канат Хафизович, ф.м.ғ.к., аға оқытушы
(ғылыми дәрежесі, атағы, Т.А.Ж.)
Кеңесшілер:
экономикалық бөлім бойынша:
Тузелбаев Бакберген Ибадиллаевич, доцент
(ғылыми дәрежесі, атағы, Т.А.Ж.)
_____ «__2__» _____06_____2020ж.
(қолы)
өміртіршілігі қауіпсіздігі бойынша:
Жандаулетова Фарида Рустембековна, проф.
(ғылыми дәрежесі, атағы, Т.А.Ж.)
_____ «_6__» _____05_____2020ж.
(қолы)
негізгі бөлім бойынша:
Панченко Сергей Владимирович, доцент
(ғылыми дәрежесі, атағы, Т.А.Ж.)
_____ «__18__» _____02_____2020ж.
(қолы)

есептеу техникасын қолдану бойынша:
Панченко Сергей Владимирович, доцент
(ғылыми дәрежесі, атағы, Т.А.Ж.)
_____ «_29___» _____04_____2020ж.
(қолы)
Нормобақылаушы:Мухамеджанова Альмира Далелханкызы, доцент
(ғылыми дәрежесі, атағы, Т.А.Ж.)
_____ «_07___» _____06_____2020ж.
(қолы)
Пікір беруші: _____
(ғылыми дәрежесі, атағы, Т.А.Ж.)
_____ «____» _____2020ж.
(қолы)

Алматы 2020

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

Ғарыштық инженерия және телекоммуникация институты
Телекоммуникациялық желілер және жүйелер кафедрасы

Мамандығы 5В071900 – Радиотехника, электроника және телекоммуникациялар

Дипломдық жобаны орындауға берілген

ТАПСЫРМА

Студент Аменова Шахназ Шухратжановна
(Т.А.Ж.)

Жобаның тақырыбы Телекоммуникация операторының мәліметтерін өңдеу кезінде машиналық оқыту әдістерін қолдану

2019 ж. «11» ___11___ №147___ университет бұйрығымен бекітілді.

Аяқталған жобаны тапсыру мерзімі «25» ____05_____2020ж.

Жобаға алғашқы деректер (талап етілетін зерттеу (жоба) нәтижелерінің параметрлері және зерттеу нысанының алғашқы деректері):
Пайдаланушылардың IP мекен-жайлары = 25000000
Домендік атаулар = 25000000
Сұраныс уақыттары=25000000
Сұраныс көздері=25000000

Диплом жобасындағы әзірленуі тиіс мәселелер тізімі немесе диплом жобасының қысқаша мазмұны:
Кіріспе
1 Телекоммуникация желілеріндегі трафикті талдау
2 Ықтималды тақырыптық модельдерді құрудың қолданыстағы тәсілдерін талдау
3 Мәтіндік құжаттар ағынын ықтималды тақырыптық модельдеу
4 Өміртіршілік қауіпсіздігі бөлімі
5 Тақырыптық модельді пайдаланудағы техникалық – экономикалық негіздеме
Қорытынды
Әдебиеттер тізімі
А қосымшасы Тақырыптық модельдеу

Графикалық материалдардың (міндетті түрде дайындалатын сызбаларды көрсету) тізімі:

Телекоммуникация желілеріндегі трафикті талдау;

Машиналық оқыту;

Телекоммуникациядағы машиналық оқыту технологиялары;

Телекоммуникация желілеріндегі трафикті өңдеу қажеттілігі ;

Ықтималды тақырыптық модельдерді құрудың қолданыстағы тәсілдерін талдау;

Тақырыптық модельдеу;

Тақырыптық модельдердің аддитивті регуляризациясы (ARTM)

Тақырыптық модельдердің ішкі және сыртқы сапасын бағалау тәсілдері

Мәтіндік құжаттар ағынын ықтималды тақырыптық модельдеу

Тақырыптарды сирету және интерпретациясын жақсарту үшін регуляризаторлар комбинациясын қолдану

Негізгі ұсынылатын әдебиеттер:

1 Воронцов К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. No 3 (455). С. 268–271.

2 Воронцов, К.В. Потапенко А.А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. 2012. No 4 (4). С. 693–706.

3 Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах, Москва: Litres, 2017.

4 Potapenko A. A., Vorontsov K. V. Robust PLSA performs better than LDA // 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. — Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2013. — Pp. 784–787.

5 Жандаулетова, Ф. Р. Охрана труда: учебник для вузов / Ф.Р. Жандаулетова, Т.Е. Хакимжанов, Т.С. Санатова; МОН РК, НАО АУЭС. - Алматы : АУЭС, 2019. - 399 с.

Жоба бойынша жобаның бөлімдеріне қатысты белгіленген кеңесшілер

Бөлімдері	Кеңесшілері	Мерзімі	Қолы
Экономика	Тузелбаев Б.И.	02.06.2020	
Ө.Т.Қ.Н.	Жандаулетова Ф.Р.	6.05.2020	
Негізгі бөлім	Панченко С.В.	18.02.2020	
Есептеу техникасы	Панченко С.В.	29.04.2020	
Нормабақылау	Мухамеджанова А.Д.	07.06.2020	

Диплом жобасын дайындау

КЕСТЕСІ

№ р/с	Бөлімдердің атауы, әзірленетін мәселелердің тізімі	Ғылыми жетекшіге ұсыну мерзімдері	Ескерту
1	Кіріспе	1.02.20 - 6.02.20	Орындалды
2	Телекоммуникация желілеріндегі трафикті талдау	6.02.20 - 8.02.20	Орындалды
3	Машиналық оқыту	8.02.20 - 9.02.20	Орындалды
4	Телекоммуникациядағы машиналық оқыту технологиялары	8.02.20 - 9.02.20	Орындалды
5	Телекоммуникация желілеріндегі трафикті өңдеу қажеттілігі	9.02.20 - 10.02.20	Орындалды
6	Ықтималды тақырыптық модельдерді құрудың қолданыстағы тәсілдерін талдау	10.02.20 - 13.02.20	Орындалды
7	Тақырыптық модельдеу	11.02.20 - 13.02.20	Орындалды
8	Тақырыптық модельдердің аддитивті регуляризациясы (ARTM)	13.02.20 - 15.02.20	Орындалды
9	Тақырыптық модельдердің ішкі және сыртқы сапасын бағалау тәсілдері	15.02.20 - 18.02.20	Орындалды
10	Есептеу бөлімі	18.02.20 - 29.04.20	Орындалды
11	Өміртіршілік қауіпсіздігі бөлімі	20.04.20 - 6.05.20	Орындалды
12	Тақырыптық модельді пайдаланудағы техникалық – экономикалық негіздеме	20.04.20 - 11.05.20	Орындалды

Тапсырманың берілген уақыты «17» ақпан 2020ж.

Кафедра меңгерушісі _____ (Темырканова Эльвира Кадылбековна)
(қолы) (Т.А.Ж.)

Жобаның
ғылыми жетекшісі _____(Жунусов Канат Хафизович)
(қолы) (Т.А.Ж.)

Орындалатын тапсырманы
қабылдаған студент _____ (Аменова Шахназ Шухратжановна)
(қолы) (Т.А.Ж.)

Аңдатпа

Бұл дипломдық жобада желідегі пайдаланушылар трафигін талдау мен өңдеудің ықтималды тақырыптық модельдеу әдістері қарастырылады. Мәтіндік құжаттар жинақтарын ықтималды тақырыптық модельдеу телекоммуникацияда, негізінен, Байес әдісі аясында дамуда. Бұл жұмыста ықтималдылық логарифмінің салмақтық қосындысына және қосымша критерийлерге - регуляризаторларға негізделген желідегі пайдаланушылар трафигін талдау үшін балама тәсіл - Тақырыптық модельдердің аддитивті регуляризациясы (ARTM) ұсынылады.

Сондай-ақ, техникалық - экономикалық негіздеме жасалып, пайдалану шығындары мен күрделі салымдар есептелінді. Еңбекті қорғауға қатысты мәселелер қаралды.

Аннотация

В данном дипломном проекте рассматриваются методы вероятностного тематического моделирования для анализа и обработки трафика пользователей в сети. В настоящее время вероятностное тематическое моделирование набора текстовых документов в телекоммуникациях развивается, в основном в рамках байесовского метода. В данной работе предлагается альтернативный подход - Аддитивная регуляризация тематических моделей (ARTM) для анализа пользовательского трафика в сети, на основе максимизации взвешенной суммы логарифма правдоподобия и дополнительных критериев - регуляризаторов.

Также подготовлено технико - экономическое обоснование, рассчитаны эксплуатационные расходы и капитальные вложения. Были рассмотрены вопросы, связанные с охраной труда.

Annotation

In this diploma project are considered methods of probabilistic topic modeling for analysis and processing of user traffic on the network. Probabilistic topic modeling of text collections has been recently developed mostly within the framework of Bayesian approach. This article presents an alternative semi-probabilistic approach – Additive Regularization of Topic Models (ARTM) for analyzing user traffic on the network, based on maximizing the weighted sum of the likelihood logarithm and additional criteria - regularizers.

Also, calculated operating costs and capital investments for feasibility study, considered issues related to labor protection.

Мазмұны

Кіріспе.....8
1. Телекоммуникация желілеріндегі трафикті талдау мәселесі 10
1.1 Машиналық оқыту 10
1.2 Телекоммуникациядағы машиналық оқыту технологиялары 11
1.3 Телекоммуникация желілеріндегі трафикті өңдеу қажеттілігі.....13
2 Ықтималды тақырыптық модельдерді құрудың қолданыстағы тәсілдерін талдау.....16
2.1 Тақырыптық модельдеу.....16
2.2 Тақырыптық модельдерді Байесиялық оқыту 17
2.3 Ықтималды латентті семантикалық талдау.....19
2.4 Тақырыптарды түсіну мәселесі 21
2.5 Тақырыптық модельдердің аддитивті регуляризациясы (ARTM) 24
2.6 ARTM-нің Байес модельдерінен артықшылығы.....27
2.7 Мультимодальді тақырыптық модельдердің аддитивті регуляризациясы (ARTM) 29
2.8 Тақырыптық модельдерді регуляризациялау әдістері 30
2.9 Тақырыптық модельдердің классификациясы 31
2.10 Тақырыптық модельдердің ішкі және сыртқы сапасын бағалау тәсілдері 32
3 Мәтіндік құжаттар ағынын ықтималды тақырыптық модельдеу.....34
3.1 Python бағдарламалау тілі 34
3.2 Python кітапханалары.....35
3.3 Мәтіндік құжаттар мен мәтіндер ағындарын талдау 36
3.4 Тақырыптық модельге арналған деректер.....37
3.5 Деректерді алдын-ала өңдеу.....38
3.6 Тақырыптарды классификациялау.....41
3.7 Мультимодальды тақырыптық модельдерді регуляризациялау **Ошибка! Закладка не определена.**
3.8 Тақырыптарды сирету және интерпретациясын жақсарту үшін регуляризаторлар комбинациясын қолдану**Ошибка! Закладка не определена.**
3.9 Тақырыптарды интерпретациялау . **Ошибка! Закладка не определена.**
4 Өміртіршілік қауіпсіздігі бөлімі 51
4.1 Еңбек жағдайларын талдау.....51
4.2 Есептеу бөлімі.....55
Өміршілік қауіпсіздігі бөліміне қорытынды 59
5 Тақырыптық модельді пайдаланудағы техникалық – экономикалық негіздеме.....60
5.1 Диплом жобасын әзірлеудің орындылығын негіздеу 60
5.2 Капиталды шығындарын есептеу 60
5.3 Пайдалану шығындарын есептеу.....61

Экономика бөліміне қорытынды.....67
Қорытынды.....68
Әдебиеттер тізімі.....69
А қосымшасы Тақырыптық модельдеу.....73

Кіріспе

Желілік трафикті талдау, оның ішінде, телекоммуникация желілеріндегі трафикті талдау - байланыс операторлары үшін маңызды міндеттердің бірі болып табылады.

Қазіргі таңда цифрлық технологиялардың дамуына байланысты телекоммуникацияның әртүрлі салаларында: автоматтандырылған желілер, жаңа телекоммуникациялық қызметтер, бизнес-процестерді құру бойынша пайдаланушылық тәжірибелер мен бүкіл инфрақұрылымға қызмет көрсету үлкен деректерді талдау мен өңдеуді қажет етеді, яғни сөздерді автоматты түрде өңдеу алгоритмдерінің қажеттілігі де артып келеді. Ықтимал тақырыптық модельдеу алгоритмдері - мәтіндік құжаттар жиынтығы мен ағынын табиғи тілде талдауының перспективалық бағыттарының бірі. Желідегі кез келген трафикті талдауды автоматтандыру айтарлықтай үнемдеуге ықпал етеді, өйткені күн сайын телекоммуникация секторы миллиондаған сұраныстарға қызмет етеді Бұл ұсынылатын байланыс қызметін кеңейтуге және тұтынушылардың маңызды бағыттарын анықтауға мүмкіндік береді.

Тақырыптық модельдеу - 90-шы жылдардың аяғынан бастап белсенді дамып келе жатқан мәтінді талдау үшін машиналық оқытудың қазіргі заманғы қосымшаларының бірі [1]. Мәтіндік құжаттар жиынтығының тақырыптық моделі әр құжаттың қай тақырыпқа жататынын және қай сөздер (терминдер) тақырып құрайтындығын анықтайды.

Қазіргі уақытта мәтіндік құжаттар жинақтарын ықтималды тақырыптық модельдеу, негізінен, Байес әдісі мен графикалық модельдер аясында дамуда.

LDA Дирихленің латентті үлестрімі [2] ықтималды тақырыптық модельдеудегі басым тәсіл болып табылады. LDA мамандандырылған жүздеген модельдер жасады. Екі деңгейлі LDA моделі тақырыптар мен құжаттардағы терминдер таралуы Дирихленің үлестірімінен туындаған векторлар болжамына негізделген. Дирихленің таралуы дискретті үлестіруге біріктіріліп, Байес туындысын едәуір жеңілдетеді. Сонымен қатар, тақырыптық модельдеуге Байес тәсілін қолданудың кемшіліктері бар. Дирихленің үлестрімі сенімді лингвистикалық негіздемеге ие емес және сирету үшін шешім қабылдауға мүмкіндік бермейді. Байес тұжырымы бір уақытта көптеген қосымша талаптарды қанағаттандыратын көп мақсатты тақырыптық модельдерді құруды қиындатады.

Дипломдық жұмыс барысында, желідегі пайдаланушылар трафигін талдау үшін Байес тәсіліне балама - тақырыптық модельдердің аддитивті регуляризациясы, ARTM ұсынылады. Бұл тақырыптық модельдеуге қате қойылған мәселелерді [3] жүйелеудің классикалық теориясын қолдану. Тақырыптық модельдің құрылысы стохастикалық матрицаның жіктеу мәселесі болып табылады. Жалпы жағдайда, оның шексіз көптеген шешімдері бар, яғни дұрыс қойылмаған. Оны жүйелеу үшін айыппұл шарттары модельге

қосымша талаптарды, регуляризаторларды ресімдей отырып, ықтималдылық логарифміне қосылады.

Бұл жұмыстың мақсаты телекоммуникация желілерін бақылауды автоматтандыруға қабілетті тақырыптық модельді құру болып табылады, ол трафикті автоматты түрде анықтайды және желілік мәліметтерге негізделген тақырыптық модельдер желі пайдаланушыларының трафиктерін өңдеуде әдеттегі модель ретінде қолданыла алатындығын көрсетеді. Сонымен, тақырыптық модельдеу әдістерін жасау өзекті және маңызды міндеттер болып табылады.

Желіде пайдаланушылардың мінез-құлқы туралы тақырыптар: веб-трафик, шифрланған веб-трафик, желідегі құрылғылар, лездік хабарламалар, қосымшалар, бағыттар және басқа да трафик түрлерінен тұрады. Тақырыптық модельдеуді желідегі оқиғаларды телекоммуникация саласындағы сарапшылардың түсіндірулеріне сәйкес келетін етіп классификациялау үшін қолдануға мүмкіндік береді [4].

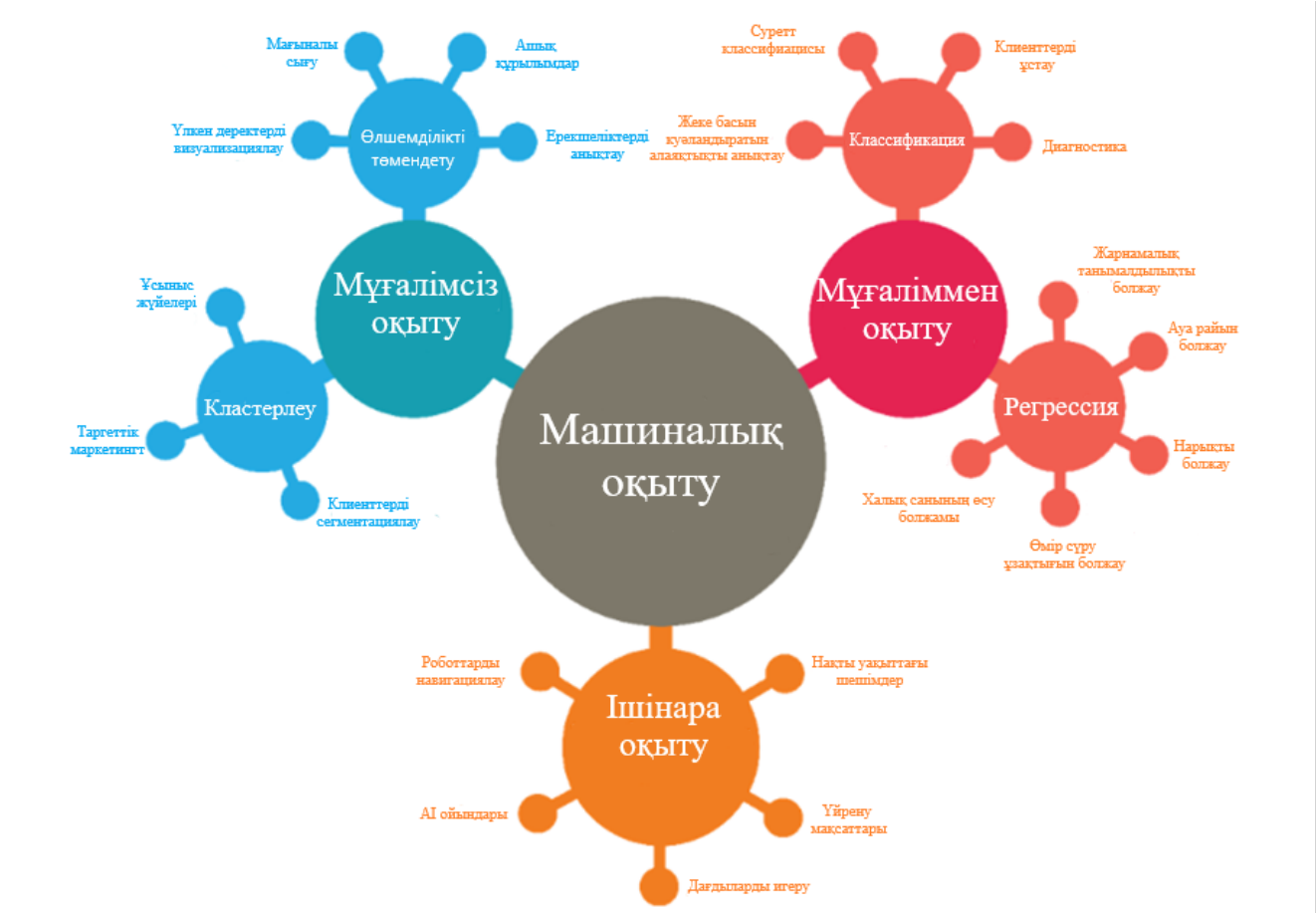
1 Телекоммуникация желілеріндегі трафикті талдау мәселесі

1.1 Машиналық оқыту

Машиналық оқыту - деректер туралы ғылымды жалпы жұртшылыққа көрсетудің негізгі әдісі. Машиналық оқытуда деректер ғылымының есептеу және алгоритмдік мүмкіндіктері статистикалық ойлауды біріктіреді. Нәтижесінде деректерді зерттеудің бірқатар тәсілдері пайда болады, олар негізінен теорияның емес, есептеудің тиімділігіне байланысты [5].

Машинамен оқыту компьютерлік жүйелерге терінің қатерлі ісігін анықтаудан бастап, жүгері бұталарын сұрыптауға дейін және жабдыққа ерте техникалық қызмет көрсетуді болжайтын жаңа мүмкіндіктер берді. Алгоритмдер дегеніміз - компьютерлерден алдын-ала болжам жасауға және қорытынды жасауға мүмкіндік беру үшін мәліметтерден шаблондар алу үшін қолданылатын әдістер.

Осы жылдар ішінде күнделікті жұмыс үстелі және веб-бағдарламалар маңызды рөл атқарды, ал бағдарламалардың тиімді жұмыс істеуін қамтамасыз ететін көптеген алгоритмдер мен әдістемелер жасалды. Алайда, біздің уақытта машиналық оқыту бүкіл әлемді жаулап алды.



1.1 сурет – Машиналық оқыту алгоритмдері [6]

Мұғалімсіз оқу кезінде олар осы белгілердің модельдеуін ұсынады. Мысалы, оларға кластерлік тапсырмалар кіреді. Бұл әдістің алгоритмдері жеке топтарға бөлінеді, өлшемді азайту алгоритмдерінің көмегімен оларды сығылған мәліметтерге қолдануға болады. Сонымен қатар, ішінара жаттығу әдістері бар. Олар басқа екі оқыту әдісі арасында орналасқан. Егер ішінара деректер болса, олар көмекші болып табылады.

Машиналарды оқыту алгоритмдері маңызды математикалық және статистикалық негізден тұрады, бірақ пәндік саланы білу маңызды емес.

Модель, мысалы, карточкалық операцияның жалған екенін неғұрлым дәл анықтай алатын болса тиімді болып есептеледі. Үлгі барлық өткен зерттеулерді ескереді, машиналық оқытумен айналысатын адам алдымен деректерді шулы компоненттерден тазартып, қажетті форматқа түрлендіруі керек. Содан кейін осы тапсырма үшін ең қолайлы алгоритмді таңдалуы тиіс. Ол алгоритм қандай әдісті, алгоритмді неғұрлым жақсы сипаттайтынын және түзететінін, оны қалай қолдануды білуі керек. Бұл алгоритмдердің математикалық негіздерін білу ең бастысы болмауы мүмкін, бірақ бұл жақсы модель құруға көмектеседі.

Бұл модельдер қол жетімді бақылау деректері бойынша оқытылғаннан кейін оларды жаңа бақылау деректерінің түрлі аспектілерін болжау және түсіну үшін пайдалануға болады.

Базалық деңгейде машиналық оқытуды екі негізгі түрге бөлуге болады: басқарылатын және бақыланбайтын оқыту.

Сонымен, оқытушымен бірге машиналық оқыту (бақыланатын оқыту) - мәліметтер атрибуттарын және мәліметтерге сәйкес белгілерді модельдеу кіреді. Үлгіні таңдағаннан кейін оны жаңа, бұрын белгісіз деректерді белгілеу үшін пайдалануға болады. Сонымен қатар, жіктеу мәселесінде белгілер дискретті, ал регрессиялық проблемаларда олар үздіксіз шамалар болып табылады.

Мұғаліммен машиналық оқытудың екі негізгі міндеті бар: классификация және регрессия. Классификацияның мақсаты - мүмкін болатын опциялардың алдын ала анықталған тізімінен таңдау болатын сынып белгісін болжау [7].

Мұғалімсіз оқыту (мұғалімсіз оқу) - мәліметтер жиынтығының сипаттамаларын ешқандай белгілерсіз модельдеуді қамтиды. Бұл модельдерге кластерлеу (өлшемдерін азайту) сияқты тапсырмалар кіреді. Кластерлік алгоритмдер мәліметтердің жекелеген топтарын бөлу үшін қолданылады, ал өлшемдерін азайту алгоритмдері деректердің неғұрлым қысылған көріністерін іздеуге арналған.

Ішінара оқыту әдістері (жетекшілік етумен) оқытушымен машиналық оқыту мен мұғалімсіз машиналық оқытудың ортасында орналасады. Жартылай оқыту әдістері тек жартылай тегтер болған жағдайда пайдалы.

1.2 Телекоммуникациядағы машиналық оқыту технологиялары

Қазіргі уақытта ұялы телефон - әлемге 90% дейін ену деңгейі бар дамушы елдердегі қарқынды дамып келе жатқан технологиялардың бірі. Бұл туралы Халықаралық электр байланыс одағының (ХЭБО) жаңа есебінде айтылған.

Қысқа хабарламалар қызметі, қоңырау, интернет, қолданушы әрекеттері, барлық сеанстар CDR-де жазылады. Телекоммуникацияда қолданылатын CDR-дің бірнеше түрлері бар. Онда көптеген ақпарат бар, мысалы, оқиғалар туралы деректер, осы оқиғалардың уақыты, әрбір оқиғаның пайда болған желідегі адресі.

Машинамен оқыту әдістері барлық деректердің белгілі бір модельге қалай жазылғанын көрсету үшін пайдалы болуы мүмкін. Машиналарды оқыту технологиялары шығындар мен оқу қисықтарын қазірдің өзінде бағалайды, сондықтан оларды көптеген салаларда қолдануға болады.

Жаңа және дұрыс жолдарды іздей отырып, қызметтерді ұсынуды дамытатын және олардың осы қызмет жеткізушісі рөлін арттыратын технологиялар телекоммуникация үшін маңызды. Телекоммуникация мұның бәрін мүмкіндігінше дұрыс және ақылға қонымды қолданғысы келеді. Егер машиналық оқытудың спектрі туралы айтатын болсақ, онда ол жинақталған тәжірибеден білім алуға арналған көптеген бағдарламаларды ұсынады [8]. Бұл функция үлкен көлемді мәліметтерден алынған заңдылықтар мен ережелерден туындайды - пайдалы білім желілік деректер мен ақпарат құралдарынан алынады немесе бұрыннан қалыптасқан ережелерде айтылған, мысалы, ұжырымдарды талдауға және процесті автоматтандыруға арналған қисынды немесе дәлелді білдіреді.

Машиналарды оқыту әдістерінің негізінде нақты мақсат, функция немесе модель үшін дайындалған алгоритмдердің жиынтығы жатыр. Әлеуметтік желілерден мәліметтерді алу, әртүрлі пайдаланушылардың сегменттерін анықтау, алаяқтарды анықтау, медиа файлдарды өңдеу және т.б. көптеген мамандандырылған алгоритмдер бар. Әр жолы алгоритмдер жаңа тәжірибеден үйренеді. Белгілі бір мақсатқа жету үшін оларды дайын мәліметтер жиынтығын қолдана отырып оқыту керек. Содан кейін нәтижелер кері байланыс бере алатын адамдармен немесе қосымшалармен бағаланады, бұл ескі модельді одан әрі жетілдіруге көмектеседі. Машиналарды оқыту және әдістер мен алгоритмдер саласындағы, инновация саласындағы зерттеулер телекоммуникациялық компаниялар мен жеткізушілерге ұсыныстарын кеңейтуге, модельдерді табуға және жасауға мүмкіндік береді [9].

Машиналарды оқыту технологиясындағы үш негізгі бизнес сценарий, ұсыныстар, даралау және медианы тану. Оларды айтарлықтай артықшылықтар беру үшін пайдалануға болады. Жиналған желілік деректерді талдау негізінде ұсынылатын қызметтер ақпараттың белгілі бір түрлері үшін пайдаланушының қалауы бойынша кеңестер береді. Қызметтер тұрғысынан, бұл пайдаланушыны толығымен немесе белгілі бір жағдайда қызықтыратын

қызметтерді сәйкестендіруді талап етеді. Бүгінгі таңда ұсыныстар жиі қолданылады. Музыка, фильмдер, кітаптар сияқты ақпарат құралдары ұсынылады.

Жекешелендіру процесі байланыс операторының қызметтеріне және табылған клиенттік базаның сұранысы мен басымдығына сәйкес ұсынылады. Қазіргі уақытта клиенттік базаның ықтимал ішкі жиынтықтарын сұрау арқылы мұның бәрі қолданушы профильдері үшін, мысалы, жазылым түрлерін пайдаланатын осы операторды талдау үшін қолмен жасалады [10].

Бейнені тану дегеніміз - кескіндердегі, дыбыстық, бейнелік және жолдағы шашыраңқылықтардың кез-келген түріне жатады. Мақсаты - машиналық оқытуында топтастарыруды үйрену: белгілі үлгіні автоматты тануды қолдану, содан кейін бұл ақпаратты бұрын көрінбейтін үлгілерге жататын категорияны анықтау.

Операторлар қауымдастық пайдаланушыларына сурет метадеректерін бөлісуге рұқсат ете алады. Өз кезегінде, пайдаланушылар кескіндерді жіктеудің тиімді жүйесіне қол жеткізеді, бұл басқа абоненттердің аннотациялары арқасында әртүрлі суреттердің үлкен жиынтығын тануды үйренді. Бірлескен индекстеу көптеген пайдаланушыларға кілт сөздер сияқты метадеректерді қосуға мүмкіндік беретін процесті анықтайды. Осының арқасында қолданушы мазмұнға байланысты ойында ерекше жеңіске жетеді. Платформалық байланысты қамтамасыз ететін басқарылатын желіде пайдаланушылар жеке цифрлық медиа серверіне фотосуреттерді жүктей алады. Консольде жұмыс істейтін фотосуреттерді қарау қосымшасы аннотация дүкенін қолдана алады, мысалы, пайдаланушылар фотосуреттерді үйдегі теледидардан санаттары бойынша көре алады.

1.3 Телекоммуникация желілеріндегі трафикті өңдеу қажеттілігі

Желілік трафик дегеніміз - кез келген уақытта желі арқылы берілетін мәліметтердің мөлшері. Желілік трафикті трафик немесе жай трафик деп те атауға болады [11].

Іздеу жүйесін оңтайландыруда (SEO) желіге трафикті тікелей, органикалық немесе ақылы деп сипаттауға болады. Тікелей трафик браузерде біреудің сайттың бірыңғай ресурстарын (URL) енгізгенде пайда болады. Органикалық трафик дегеніміз - іздеу жүйесін мазмұнды іздеуде қолданатын нәтижесі, ал ақылы трафик біреудің жарнаманы басқанын білдіреді [11].

Деректер орталығын басқару кезінде желілік трафикті солтүстік-оңтүстік немесе шығыс-батыс деп сипаттауға болады. Солтүстік-оңтүстік деректер орталығы мен желіден тыс орналасқан орын арасында жылжитын клиент-сервер трафигін сипаттайды. Солтүстік-оңтүстік трафик әдетте деректер орталығына кіретін және шығатын трафикті бейнелеу үшін тігінен бейнеленген. Интернеттің алғашқы күндерінде желілік трафиктің көп бөлігі солтүстіктен оңтүстікке дейін болды.

Керісінше, шығыс-батыс желі трафигі деректер орталығындағы серверден серверге өтетін деректер пакеттерін сипаттайды. Шығыс-батыс

термині көлденең желілік трафикті (LAN) бейнелейтін желілік диаграммалармен шабыттандырды. Аппараттық құралдарды виртуализациялау және заттардың интернеті (IoT) - бұл шығыс-батыс желілік трафиктің өсуіне ықпал ететін екі ұғым.

Оператор деректерді басқару немесе басымдық беру арқылы және трафиктің мөлшері мен түрін өлшеу арқылы желілік трафикті бақылай алады. Желілік трафик пакеттерге енеді, олар желі жүктемесін қамтамасыз ететін мәліметтердің бірлігі болып табылады. Мысалы, веб-параққа кіру арқылы пайдаланушы бірқатар пакеттерді алады. Әдеттегі пакетте 1000-5000 байт болады.

Желінің өнімділігін басқару кез-келген бұзушылықтар үшін желілік трафикті талдау және басқару үшін желілік трафикті бақылауға көмектеседі. Желілік трафик анализаторы - бұл желінің жұмысына, қол жетімділігіне және қауіпсіздігіне әсер ететін процесс. Желілік трафикті бақылау компьютердің желілік трафигін тексеру үшін әртүрлі құралдар мен әдістерді қолданады.

Желілік трафиктің жалпы проблемаларына сервер, маршрутизатор немесе желіаралық қалқан сияқты компоненттердің істен шығуы немесе қиындықтар мен жоғары кідіріс сияқты трафиктің бұзылуы кіреді [12]. Қазіргі трафик көлемін өткізу үшін өткізу қабілеттілігі жеткіліксіз болған кезде проблемалар туындауы мүмкін. Жүйеге кірудің кешеуілдеуі немесе оның нәтижесіне дейін кідіруі, деректер орталығындағы құрамдас бөліктердің желінің трафигін арттыра отырып, бір-біріне ақпаратты жіберетіндігімен байланысты болуы мүмкін. Шығыстан батысқа қарай жылжу кезінде жоғары кідіріс жиі орын алуы мүмкін.

Желіні бақылаудың әртүрлі бағдарламалық жасақтамалары ұсынылатынына байланысты әр түрлі болуы мүмкін, бірақ көптеген желілік бақылау жүйелеріне сәйкес келетін бірнеше жалпы тапсырмалар бар. Бірінші қадам - желіге қандай құрылғылар қосылғанын анықтау, оның ішінде маршрутизаторлар, принтерлер, брандмауэр және т.б. Желідегі әрбір құрылғы жүйеде проблемалар тудыруы мүмкін, сондықтан бұл мәселелерді шешпес бұрын, олардың бар екенін білу керек.

Одан әрі, жүйе өзі шығаратын мәліметтерді көрсетудің қандай да бір әдісін ұсынады. Карточкасыз ақпарат жиі түсінілмегендіктен, ақпарат өте бейімделген болады. Деректерді көрсету мүмкіндігі жүйелерде әр түрлі болады, сондықтан желілік әкімшілер мәліметтерді жақсы түсіну үшін оларға берілген шектеулі бағдарламалық жасақтаманы пайдалануға дайын болуы керек.

Барлығы дайын болған кезде, бақылау процесін бастау керек. Бұл желінің сандық аспектілерін, сонымен қатар физикалық компоненттерін бақылауды қамтуы мүмкін. Сандық жағынан қолданушыларға бес маңызды аспектіні қарастыру ұсынылады: интерфейсті, жадты, дискіні, Ping қол жетімділігі мен процессорды пайдалану. Оларды бақылау, әдетте, желі операторларын сәтсіздікке әкелуі мүмкін жерлерді бақылауға мүмкіндік

береді. Көптеген желілік бақылау жүйелері пайдаланушыларға құрылғының өзі сияқты температураның физикалық жақтарын бақылауға мүмкіндік береді.

Жоғарыда айтылғандай, желілік проблемалар сөзсіз, сондықтан желіні бақылаудың келесі қадамы - хабарландыру процесі. Желілік бақылау жүйелері бірдеңе дұрыс болмай қалса, операторға ескерту жасай алады, ал көбіне бір нәрсе болып қалмас бұрын ескертуге бағдарламаланады. Мұны бақыланатын компоненттердің шекті мәндерін орнату арқылы жасауға болады. Бұл дегеніміз, егер жүйе проблема тудыратын белгілі бір шекті мәнге жетсе, әкімшілер апаттың алдын алуға мүмкіндік алады.

Сонымен, желілік бақылау жүйелері операторларға кесте деректерін жақсартуға көмектесетін есептер шығарады. Бұл тек операторларға проблемаларды анықтауға және шешуге көмектесіп қана қоймай, сонымен қатар болашақта желінің қалай жұмыс істейтіні туралы түсінік береді.

Желінің сапасын қамтамасыз ету үшін операторлар трафикті талдап, бақылап, қорғауы керек. Желіні бақылау жүйенің тұрақты жұмысын қамтамасыз ету үшін сәтсіздіктер мен кемшіліктерді компьютер желісін бақылауға мүмкіндік береді. Желіні бақылауға арналған құралдар, әдетте, пайдаланушыларға желі жұмысындағы кез келген маңызды немесе жағымсыз өзгерістер туралы хабарлайды. Желіні бақылау операторлар мен IT мамандарына желідегі кез-келген проблемаларға тез әрекет етуге мүмкіндік береді.

2 Ықтималды тақырыптық модельдерді құрудың қолданыстағы тәсілдерін талдау

2.1 Тақырыптық модельдеу

Он жылдан астам уақыт бұрын негізгі кадрлардың сценарийлеріндегі трафикті өңдеу саласында жұмыс басталды. Одан кейін интернет және ұялы желілер сияқты телекоммуникацияда деректерді беретін күрделі желілер пайда болды. Сонымен қатар, желіде жаңа трафик бұзушылықтары да пайда болды, бұл трафикті талдау жүйесінде үнемі ескеріліп отыруы керек. Тақырыпты ашу үшін жасырын модельдеу әдісі өте танымал болды.

Тақырыптық модельдеу - бұл 90-жылдардың аяғынан бастап қарқынды дамып келе жатқан, деректер туралы ғылымның бір саласы, машиналық оқытудың заманауи шешімі. Мәтіндік құжаттардың тақырыптық моделі әр құжаттың қай тақырыпқа қатысты екенін және әр тақырыпты қалыптастыруға қандай сөздер (терминдер) қатыса алатындығын анықтайды [13].

Тақырыптық модель - бұл иерархиялық модель. Ол екі деңгейден тұрады: (1) кластер шаблондарының жоғарғы деңгейі; (2) әр топта деректер үлгісін құру үшін жалпы кластерлік үлгілер қолданылатын топтың деңгейі.

Ең танымал PLSA (ықтимал жасырын семантикалық талдау) және LDA. Екеуін де үлкен құжаттар жинағында пайдалануға болады.

Ықтималдық модель құжаттық жұптың пайда болуы (d, w) үш балама жолмен жазылады:

$$p(d,w) = \sum_{t\in T} p(t)p(w|t)p(d|t) = \sum_{t\in T} p(d)p(w|t)p(t|d) = \sum_{t\in T} p(w)p(t|w)p(d|t), \quad (2.1)$$

мұндағы $p\,(t)$ - тақырыптарды жинақ бойынша тарату. Бірінші көрініс симметриялы деп аталады, екінші және үшінші - асимметриялық. Олар тақырыптық үлгіні үйренуге арналған әртүрлі итеративті процестерге әкеледі.

Модель $g \in G$ әр түрлі санау мәліметтері бар деп санайды. $\{n_1^8, \ldots, n_m^8\}$, мұндағы, m - әр түрлі санаушылардың саны, ал k әрбір токенді сипаттайды, $\{1, \ldots, m\}$ және есеп нөмірлері. Тақырып моделін мына құрылым ұсынуы мүмкін:

- $p(k|t) \sim Dirichlet(\propto)$ таңбалары бойынша ықтималдылықты бөлу), мұнда, $t \in \{1, \ldots, l\}$;
- тақырып пропорцияларын сызу, $\theta_g \sim Dirichlet(\beta), g \in G$;
- әр топ үшін$\in G$:
  - тақырыпты сызу $t \sim Multinomal(\theta_g)$;
  - токен сызу $k \sim p(k|t)$;
  - g мөлшері үшін (a) -тен қайталау.

Ықтималдық тақырыптық модель (PTM) әр тақырыпты көптеген терминдердің дискретті үлестріріммен, ал әр құжатты көптеген тақырыптардың дискретті үлестірімімен сипаттайды. Құжаттар жинағы кездейсоқ және осындай бөлу қоспасынан тәуелсіз таңдалған терминдер

тізбегі болып табылады және есеп қоспаның құрамдас бөліктерін берілген үлгінің көмегімен қалпына келтіру болып табылады. Құжат немесе термин бір уақытта әртүрлі тақырыптардың үлестрімдерімен байланысты болуы мүмкін болғандықтан, РТМ тақырыптық кластерлерінде құжаттарды, сонымен қатар термин сөздерін ұқыпты кластерлеуді жүзеге асырады. Көрсетілген тақырыптар талданатын коллекцияның негізгі тақырыптарына сәйкес келеді, тақырып модельдеу алгоритмі осы тақырыптарға сәйкес құжаттарды ұйымдастырады.

Көптеген тақырыптық модельдер деректер негізінде олардың болжамды тиімділігі бойынша бағаланады. Бәрі тақырыптық модельдердің ықтималдығын арттыруға жарамды екендігіне, яғни құжаттар апостериори жиынтығына негізделді. Ең жақсы модель жоғары ықтималдылықты анықтайды деп болжауға болады [14].

Тақырыптық модельдер ғылыми жарияланымдардағы немесе жаңалықтар арналарындағы тенденцияларды анықтау, құжат суреттері мен бейне ағындарын жіктеу және санаттарға бөлу, ақпаратты, соның ішінде көптілділікті іздеу, веб-беттерді белгілеу, мәтіндік спамды анықтау, ұсыныс жүйелері және басқа қосымшалар үшін қолданылады. Тақырыпты модельдеу алгоритмін қолдана отырып, тақырып моделінің шығыс мәндері құрылды. Олар осы айнымалылар бойынша ықтималды бөлу тобы ретінде сипатталған.

Тақырыпты модельдеуде айнымалылар бастапқыда мәтіннің негізгі бөлігінде ұсынылады. Онда бірыңғай сөздер талданады. Сондықтан, біз тақырыптық алгоритмнен алған әр тақырыпта сөздердің ықтималды үлестірімі болуы мүмкін. Бұл бөлім модельдеу тақырыптарында нақты мақсаттарға жетудің негізгі әдістерін қамтиды. Біріншіден, пікірталас тақырыптық модельдеудің негізгі идеяларын қалыптастыратын жасырын тақырыптардың қарапайым моделінен басталады.

2.2 Тақырыптық модельдерді Байесиялық оқыту

Берілген мақалада [15] трафикті талдау Байес шешімінің ережесіне негізделген, сондықтан оны статистикалық тәсіл ретінде жіктеуге болады.

Шешімдер қабылдау үшін Байес ережесі статистикалық тануда кеңінен қолданылады [15]. Модельді тану проблемасын былайша сипаттауға болады: нысандар жиынтығын бірнеше кластарға бөлуге болады. Осы объектілердің әрқайсысы үшін байқалған сипаттамалардың жұбы өлшеніп, векторға біріктірілді. Бұл бақылау векторы әр нысан үшін әр түрлі болды; вектор X байқау векторы бар объект ретінде түсіндірілді. Жаңа объектіні жіктеу үшін әр класқа X-тің ықтималдығын бөлу үйретілді, содан кейін байқау векторымен x ықтималдығы c немесе $P(c|x)$ класына жатады деп есептелді. Жаңа объект келесі түрде жіктелді:

- объект үшін бақылау векторы өлшенеді;
- әр класс үшін $P(c|x)$ ықтималдығы есептелінеді;
- объект сыныбы ретінде ықтималдығы жоғары сынып таңдалады.

Бұл шешім ережесі қателіктің минималды деңгейіне қатысты Байес шешімінің ережесі деп аталады. Шешімнің кез-келген ережесінде Байес ережесінен гөрі қателіктердің көбірек болатындығы көрсетілген. Шешімнің кез-келген ережесінде Байес ережесінен гөрі қателіктердің көбірек болатындығы көрсетілген. Постериори деп аталатын ықтималдық $P(c|x)$ келесідей өрнектелуі мүмкін:

$$P(c|x) = \frac{P(c)p(x|c)}{p(x)}, \quad (2.2)$$

$P(c|x)$ - векторын көрудің кластық шартты тығыздығы x $P(c)$ – c класының a-априори ықтималдығы, ал $P(x)$ - векторын көру ықтималдығының тығыздығы. Егер x векторы үшін $p(x)$ әр к класы үшін тұрақты болса, $p(c|x)$ класты шартты ықтималдылық тығыздығы үйрену керек және a (a) априори ықтималдығын $P(c)$ есептеуге болады в классының векторын бақылаудың салыстырмалы жиілігі ретінде. Мысалы егер біз n векторларын байқасақ және олардың n_1 векторлары c_1 класына жататын болса, $\hat{P}(c_1)$ эмпирикалық ықтималдығын келесідей есептеуге болады:

$$\hat{P}(c_1) = \frac{n_1}{n}, \quad (2.3)$$

$p(x|c)$ оқыту қиын. Қарапайым әдіс - бұл гистограммаларды қолдану: вектор кеңістігі интервалдарға бөлінді, біз әр интервалға түсетін векторлардың санын есептейміз, содан кейін осы аралықтағы векторлардың ықтималдығын осы бөліктегі векторлар санына пропорция ретінде есептедік.

Бұл әдістің кемшілігі интервалдар саны векторлармен салыстырғанда аз болатындығында (мысалы, векторлық кеңістіктің төменгі өлшемі).

Байес шешімінің ережесі GSM ұялы желілерінде пайдаланушы профильдерін жасауға қатысты қолданылды. GSM желісі [16] тарату жиілігін қайта пайдалануға мүмкіндік беру үшін таратылады. Бірнеше мобильді коммутация орталықтары (MSC) және жергілікті дерекқорлар (Visitor Location Register, VLR) жүйеге сәйкес жергілікті аудандарға (MSC-аудандар) қызмет көрсетеді. MSC-аймағы өз кезегінде бірнеше орналасу аймағына (Location Areas, LA) бөлінеді, және LA бірнеше ұяшыққа бөлінеді, олар ұялы желінің ең кішкентай бөлігі болып табылады.

Абоненттерге қызмет көрсету аймағының кез-келген жеріне бару тегін болғандықтан, олар ұяшықтарға кіріп, кететіні анықталды. Сондықтан орналасқан жер туралы ақпаратты басқару керек болды. Кейбір өнімділікті ескере отырып, желіде орналасқан жер туралы ақпарат Home Location Register (HLR) тұрғысынан қарастырылды.

Мұндай деректерді басқару үйді тіркеу регистрінің орталық дерекқорының көмегімен ұйымдастырылған.Орталықтандырылған HLR деректерді абоненттер мен мобильді станцияларда сақтайды. Абонент жаңа

MSC қызмет ететін жаңа орналасу аймағына кірген кезде VLR-ге тек тиісті деректер жүктеледі.

Ұялы желілер үшін Байес алгоритмі қолданылды. Тәуліктің уақыты бақылау векторы ретінде түсіндірілді, ал ұяшықтар сынып ретінде, проблема Байес ережелерін қолдана отырып, күннің белгілі бір уақытында жылжымалы станция үшін ең ықтимал ұяшықты табу болды: уақыт осін интервалдарға бөлу (мысалы, бір секунд) ұзындығы - δt. Қозғалыс сызбаларын қолдана отырып, t_ $t_{1,c_1}, t_{2,c_2}, \ldots, t_{n,c_n}$ векторларының тізбегі $t_{i+1,c_k} - t_{i,c_1=}\delta t$ қарастырылды. Осы векторларды қолдана отырып, $P(c)$ және $p(x|c)$ үлестірімді $\hat{P}(c)$ және $\hat{p}(x|c)$ сәйкес эмпирикалық үлестіру және осы бөлулердің көмегімен мүмкін болатын қозғалыс профилін есептелінді. Байес шешімінің ережесін қолданатын орналасу алгоритмі Байес алгоритмі деп аталады.

Болжау алгоритмдерінің жұмысын бағалау үшін метрика ретінде орташа болжам деңгейі пайдаланылды.

Орташа болжау деңгейі (Mean Prediction Level, MPL) - жылжымалы станцияның болжалды ұяшықта болуының эмпирикалық ықтималдығы. Мұны пайдаланушының нақты ұяшығын күтілетін ұяшықтармен салыстыратын трафикті анықтау жүйесінің тексеру функциясы анықтайды. Сәтті тексерулер саны hit болып, сәтсіз тексерулердің саны $miss$ жіберу:

$$MPL = \frac{hit}{hit+miss}, \tag{2.4}$$

Бұл мақалада мобильді пайдаланушыларды Байес шешімінің алгоритміне негізделген профильдеу үшін алгоритм ұсынды. Бұл тәсілді мобильді желілерді пайдаланушыларға қауіпсіздіктің жетілдірілген мүмкіндіктерін ұсыну үшін қолдануға болатындығы көрсетілді, бірақ модельдер априорлы таратуды қолдануға мәжбүр. Математикалық тұрғыдан ыңғайлы, оның көпжақты таралуы бар мультиномиальділікке байланысты. Алайда, ол ешқандай табиғи тіл құбылыстарын модельдемейді және сенімді лингвистикалық негіздемелері жоқ.

2.3 Ықтималды латентті семантикалық талдау

Уақыт өте келе тақырыптар эволюциясы мен тақырыптардың иерархиялық құрылымын көрсететін неғұрлым жетілдірілген модельдер төменде сипатталады. Дирихленің жасырын таралуы [3] әрбір деректер нүктесі бірнеше кластерге тиесілі болуы мүмкін ретінде қарастырады. Әдеттегі қосымша - бұл құжаттар жиынтығындағы тақырыптық кластерлерді сәйкестендіру. LDA тақырыптық модельдеудің нақты фактісі болды. Ол көбінесе жаңа тәсілдермен салыстыру кезінде негізгі әдіс ретінде қолданылады.

Енді біз стандартты тақырыптық үлгіні құру үшін LDA әдісінің қысқаша сипаттамасын береміз. [3] LDA - корпустық деректердің генеративті

ықтималды моделі. Негізгі идея - кіру құжаттары жасырын тақырыптарға кездейсоқ араласу түрінде ұсынылады және осы тақырыптардың әрқайсысы сөздердің үлестрімімен сипатталады. Бір құжатта бірнеше сөздерден тұрады:

$$v = (thecat, sat, on, the, mat), \tag{2.5}$$

Егер D сөздерінің қол жетімді сөздігіндегі әрбір сөз ерекше күйге берілсе (dog = 1, tree = 2, cat = 3, …), онда n^{th} құжатын сөз индекстерінің векторы ретінде ұсына аламыз:

$$v^n = (v^n, \ldots, v^n_{w_n}), \qquad v^n_i \in \{1, \ldots, D\}, \tag{2.6}$$

мұндағы, W_n–n^{th} құжаттындағы сөздер саны.

Әр құжаттағы сөздердің саны әр түрлі болуы мүмкін, бірақ олардың жалпы сөздігі бекітілген. Кез-келген құжатта бірнеше тақырып болуы мүмкін, мақсаты - құжаттағы жалпы тақырыптарды анықтау. Бірінші кезекте сөздерді құрудың әлеуетті заңдылықтарын, оның ішінде ықтимал тақырыптарды ашқан пайдалы.

Кез келген n құжат үшін бізде тақырып таралуы бар π^n, $\sum_{k=1}^{K} \pi^n_k = 1$ біз оны тақырып мүшелігіне негізделген ықтимал түрде бейнелей аламыз. Мысалы, n құжатта «жануарлар» және «қоршаған орта» тақырыптары бойынша көптеген тақырыптық хабарламалар болуы мүмкін.

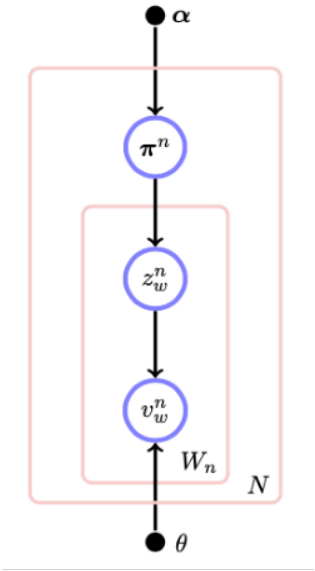
Тақырып шынымен мүмкін - «жануар» атауы, ол тақырыпты шығаратын $\theta_{i|k}$ сөздерінің түрінен шыққан. Күрделілікті басқару үшін әр нақты құжаттағы белсенді тақырыптар санын шектеместен бұрын Дирихле үлестірімін қолдана аламыз:

$$p(\pi^n|\alpha) = Dirichlet((\pi^n|\alpha), \tag{2.7}$$

мұндағы, α - тақырыптар ұзындығының векторы.

Әр құжат үшін n, алдымен модельдегі π^n тақырыптардың таралуын мойындаймыз. Осыдан кейін құжаттағы осы сөздердің $w = (1, \ldots, W_n)$ позициясы үшін z^n_w тақырыбы таратудан алынады. Тақырыпты ескере отырып, содан кейін тақырыпты сөздердің бөлінуінен аламыз. Модель параметрлері дегеніміз - сөздерді бөлу және әр тақырып бойынша тақырыптарды бөлу параметрлері.

LDA моделін оқыту жаттығу параметрлерімен байланысты, параметрлер тақырыптар санына байланысты және әр тақырыптағы сөздердің бөлінуі θ сипатталған. Өкінішке орай, артқы жағындағы шеттерде жаттығудың қажетті формасын табу қиын. Бұл модельді тиімді жақындату - бұл зерттеу нүктесі, жақында мутация мен іріктеу әдістері жасалды [16].



2.1 сурет - Дирихленің латентті үлестрімі [1]

LDA және PLSA-да ортақ көп нәрсе бар [17], олардың екеуі де мүмкін тақырыптарды бөлуге негізделген құжаттарды сипаттайды. LDA - гиперпараметрлерді орнату сияқты мәселелерді шешу үшін максималды ықтималдылықты қолдана алатын ықтималдық модель. Екінші жағынан, PLSA матрицалық факторизация әдісі болып табылады (мысалы, PCA). Егер біз осы тексерулерді қолданатын болсақ, онда модель үшін ең жақсы PLSA гиперпараметрлерін тауып және баптай аламыз. PLSA тек бастапқы оқыту деректерін сипаттайды, ал LDA - негізінен жаңа құжаттарды синтездеу үшін пайдалануға болатын деректер үлгісі болып табылады.

2.4 Тақырыптарды түсіну мәселесі

Телекоммуникациялық желілерде тақырыптық модельдеу бойынша алғашқы жұмыс [18] LDA-ны қолданумен байланысты болды. Желі мониторингін автоматтандыру жобасы әзірленді, бұл желілердегі трафиктерді классификациялауды автоматты түрде анықтауға мүмкіндік берді. Бұл жобадағы тақырыптық модельдеу сымсыз желілерде деректерді зерттеу құралы ретінде пайдаланылды. Оқиғалар мен олардың топтамалары үшін тақырыптық модельдеу қолданылды және телекоммуникация саласындағы мамандармен өзара әрекеттесу маңызды болды. Тақырыптық модельдер трафикті анықтауға арналған дәстүрлі модель ретінде қолдануға болатын оқиғаларсыз желілік деректер негізінде ұсынылды.

Сондай-ақ, пайдаланушылардың трафиктерін талдаудың тақырыптық модельдер көрсетілді. Бұл желілік модельдер желілік мәліметтерге негізделді. Сынақ деректерінде біз әрқашан қалыпты және қалыпты емес модельдер бір-бірінен айтарлықтай ерекшеленді деп санаймыз. Сонымен қатар, осы айырмашылықтар трафиктердің ықтимал түпкі себептері туралы ақпарат береді, бірақ түпнұсқаны локализациялау процесін автоматтандырады, алайда

байланыс операторлары тақырып модельдерін қолмен анықтау үшін көп уақыт керегі анық.

Базалық станцияларда 4G тарату желілерінде мамандар тәжірибе арқылы ақпарат алады. Оларда телекоммуникацияны іске асыру уақытының ауыспалы мәні бар. Тақырыпта модельдер, айнымалылар бөлек сөздер құрайды. Барлық оқиғалар белгілі бір уақыт аралығында жазылады және ауыспалы жиіліктерде қарастырылады. Сонымен қатар, дәстүрлі тақырыптық модельдеуде мәтін құжаттағы сөздердің жиі қайталануын анықтайды. Деректер бір базалық станциядан уақыт аралығында қабылданады. Жиналған айнымалылар жиынтығы бір құжатқа жатады. Кез-келген базалық станция уақыт аралығын анықтау үшін құжат жібереді. Барлық құжаттарды әртүрлі аралықта жинау арқылы жинақ жасалады.

Нақты құжат белгілі бір процесті анықтайды. Мысалы, екі станцияның арасындағы байланыс пайда болғанда, қай ұялы телефон байланыстан тыс жерде екенін анықтап біледі [19]. Қазіргі уақытта қай базалық станция пайдаланылып жатқанын анықтаңыз. Барлық тапсырмалардың өзіндік себептері бар. Базалық станциялар уақыт өте келе әртүрлі процестерге қатысады және бірнеше тақырыпты қамтуы мүмкін. Мәтіндік құжаттар өлшенген уақыт аралығында сақталады. Егер біз телефондар мен базалық станция арасында байланыс орнатсақ, оны тақырыпты анықтайтын мысал ретінде анықтауға болады. Бүкіл процесті сигналдар тұрғысынан сипаттауға болады. Байланыс орнатуға сұраныс ұялы телефоннан келіп түседі, егер байланыс орнатылса, бір бірімен байланысады. Бұл процесс бірнеше кезеңнен тұрады. Барлық осы оқиғалар осы базалық станцияға тиесілі. Ерекше оқиғалар орындалу уақытының өзгермелі санауыштарында сақталып жазылады.

Соңғы кезеңде орнатылған байланысты немесе жойылғанын көреміз. Базалық станция оқиғаларды белгіленген қосылымда немесе істен шығу есептегішінде тіркейді. Әр түрлу базалық станциялардың арасындағы байланысты басқа мысалға тағайындауға болады. Содан кейін орнатылған абонент байланыстан шығады, базалық станция басқа станцияға жақын болады.

Сымсыз негізгі желіде тапсыру сұранысын анықтаймыз. Тағы бір базалық станцияда инфрақұрылым салу керегі анық. Ұялы телефонды негізгі станция қабылдайды. Бұл телефон екінші базалық станцияға қосылуы керек, содан кейін біз осы инфрақұрылымды аяқтауға болады. Көптеген базалық станциялар әдетте көптеген процестерді өңдейді, бірнеше телефон бір-біріне қосыла алады.

Көптеген жағдайларда ұқсас процестер бірдей оқиғаларды жинайды, ол сымсыз байланысқа қосылған болуы мүмкін. Сонымен қатар, ауыспалы есептегіштер уақыт өте келе кеңейеді. Санауыштарда көптеген оқиғалар, көптеген ақпарат жинақталады, оларды бәрін анықтау өте қиын.

Деректер жиынтығы статистикалық сипаттамаларға ие. Бұл төлсипаттар тақырып үлгісімен алынады. Сіз бұл оқиғалардың бір процесті, хабар таратуды білдіретінін байқауыңыз мүмкін. Олардың барлығы белгілі

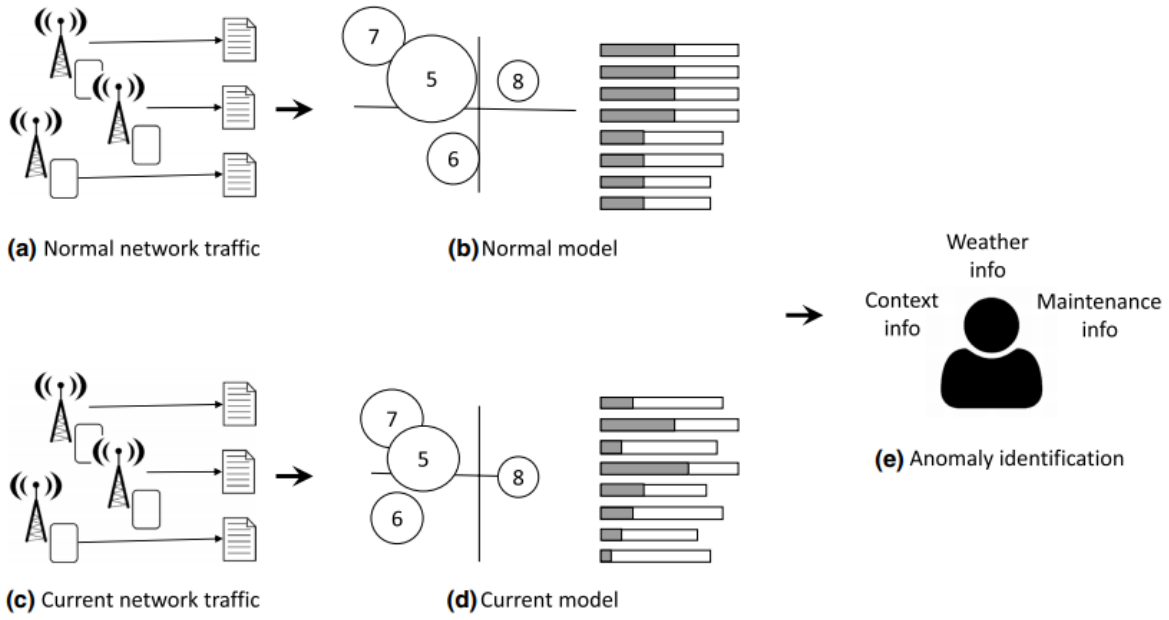
тақырыптарды құрайды. Бұл тапсырма үшін LDA Gibbs және LDAvis Sievert және Shirley пайдаланылды. LDA модель ретінде, екіншісі жинақталым түрінде болды.

Дайын модельді алғаннан кейін біз осы деректерді тексеру процесін бастаймыз. Телекоммуникация саласындағы сарапшы осы модельдерге баға береді. 2.2 Суретте тақырыптық модельдеуді қолдану арқылы пайдаланушылар тақырыбын анықтау процесіне сипаттама берілген. Нақты уақыттағы желідегі қалыптан тыс мінез-құлықты қосып, құрамын талдайды. Тақырып моделі берілген мәліметтер жиынтығынан оқытылды. Деректер бір сағаттық трафиктан алынды. Екінші модель қазірдің өзінде қалыпты емес мәліметтерге толы.

Телекоммуникация мамандары бұған дейін барлық оқиғаларды анықтаған. Олар осы басқару объектісінің (MME) қол жетімсіздігінен пайда болды. Мобильділікті басқару субъектісі - бұл ұялы байланыс тіркеулеріне, қауіпсіздікті қамтамасыз етуге арналған процестер мен процестерге қатысатын төртінші буындағы ұялы ядро желісінің басқару жазықтығының түйіндік нүктесі. Мобильділікті басқару субъектісі ұялы құрылғылардың орналасқан жері туралы деректерді сақтайды және ең жақын шлюзді таңдауды қанағаттандырады.

Визуализацияны қолдану үшін, оны жақсы білу қажет. Бір сарапшыға визуализация туралы жақсы түсінік пен танысу үшін жеткілікті уақыт керек болады. Сарапшы алдымен алғашқы эксперименттерге қосылып және LDA тақырып моделіне сәйкес тақырыптарды көрсететінін білуі дұрыс. Кейіннен сарапшымен одан бір сағаттық желілік режимнің тақырыптық моделін және бір сағаттық желінің тақырыптық моделін салыстыра отырып, пайда болған мінез-құлықтың (а) екінші тақырыпқа қатысты екендігін анықтауы сұралады. (б) тақырып моделінің сипатын анықтау мүмкін бе, деген сұрақ туындайды.

LDA тақырыбын модельдеуде телекоммуникациялар, сонымен бірге пайда болатын трафик бұзушылықтар туралы тақырып бар-жоғын тексеру үшін сарапшы трафиктің қалыпты деректері сияқты негізгі ақиқатты білуі керек. Сондықтан сарапшының өзі осы нақты кездейсоқ жағдайларды тиісті сынақ жағдайлары ретінде таңдады. Бұл сонымен бірге тарауда сипатталған қате пікірді негіздеу қаупін де білдіреді. Алайда, сарапшы, егер оның сенімділігі дәлелденбеген болса, модельдеу тақырыбын немесе трафик бұзушылықтарды анықтаудың кез-келген әдісін алға тартпайтын компанияның телекоммуникациядағы рөлі қауіпті екенін біледі. Сонымен қатар, тәжірибе кезінде сарапшы оған қандай ақпарат пайдалы екенін және неге қажет екенін анықтайды. Осы ақпаратқа телекоммуникация операторларының ақпараттық қажеттіліктеріне бағытталған әр түрлі және визуализациялар негізделеді.



2.2 сурет - Тақырыптық үлгіні қолдана отырып, қалыптан тыс тақырыптарды анықтау тәртібі. (a) Уақыт туралы деректер қалыпты және әдеттегі желілік трафиктен тұрады және жұмыс уақыты кезінде жиналады. (b) Тақырыптық модель анықталған. (c) Егер желі қосылса, деректер толтырылады. (d) Желіде күдікті оқиғалар туындаған жағдайда, тақырыптық модельді қолдануға тура келеді. (e) Сарапшы модельді алдыңғы модельмен салыстырып, соның негізінде ол желідегі мінез-құлқымен түсіндірілетін басқа құбылыстар мен оның мүмкін себептерін табады [18]

Телекоммуникациялардағы трафикті талдау кезінде тақырыптық модельдеуді қолдануға болатындығын білу үшін, ең алдымен, екі модельде де (қалыпты және кездейсоқ) тақырыптар тест жағдайлары үшін процедураларды сипаттайтынын, яғни себеп-салдарсыз болуын талдау қажет. Сарапшылар мен телекоммуникация операторларының бірдей оқиғаларды топтастыруына сәйкес келетін оқиғалар бар. Олар LDA-ны операторға интерактивті түрде трафик бұзушылықтарды анықтауға көмектесу үшін қолданылады, LDA операторлар арасындағы жақсы сәйкестік тақырыпты құрайтын түсінігіңізді қажет етеді.

Алдыңғы сараптамаларда LDA құрған тақырыптар сарапшылардың оқиғаларды топтастыруына сәйкес келеді. Осы зерттеу жағдайларын зерттей отырып, қалыпты модельдер де, оқыс оқиғалар модельдері де бұл қажеттілікті қанағаттандырды. LDA ұсынатын тақырыптар саны (k) болғандықтан, тақырыптардың қолайлы санын қалай анықтауға болады деген сұрақ туындайды. Бұл сонымен қатар тақырыптар серпінді және қалыпты модель мен оқиға моделі үшін өзгеше болуы мүмкін бе деген сұрақ туындайды. Бұл сұрақтарды телекоммуникация маманы мұқият зерттеді. Әр түрлі тақырыптағы модельдерін жасады. Мәліметтер жиынтығы үшін және сарапшы тақырыптардың әр саны үшін мазмұн тақырыптарын талдайды.

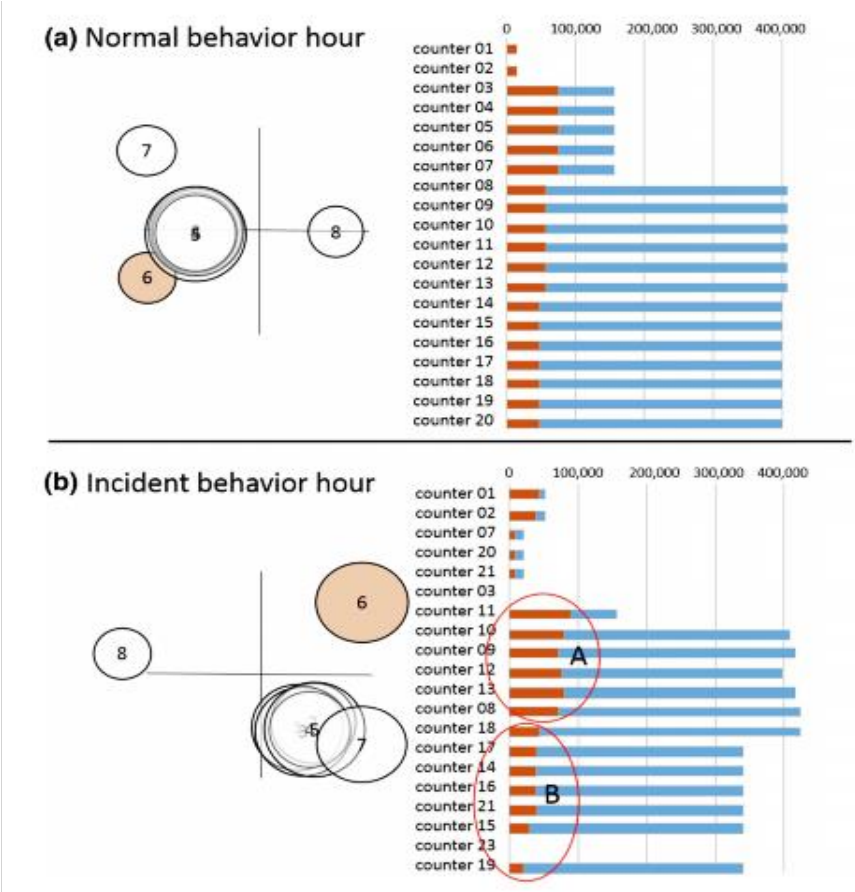
Нәтижесінде сарапшы әрбір мәліметтер жиынтығына сәйкес келетінін анықтай алды.

Жұмыстың мақсаты трафикті талдау мен өңдеу болғандықтан, қалыпты үлгіні құру үшін тақырыптардың тиісті саны бір рет анықталады. Тақырыптар шеңберіндегі мазмұн ретінде қызығушылық тудыратын айырмашылықтар инциденттер болған кезде өзгереді.

Бұл тәжірибелер сарапшы жеке тақырыптар ретінде бөлетін тақырыптардың санына байланысты LDA бір-біріне ұқсас бірнеше тақырыпты құратынын, сарапшы оларды сол тақырыптың даналары ретінде оңай анықтай алатындығын көрсетеді. LDAvis визуализациясында бұл тақырыптар диаграммада бірдей күйде орналасқан, сондықтан мұнда да бір-біріне тиесілі деп оңай анықталады.

Гистограмма қалыпты және оқиғалы модель үшін әртүрлі айнымалылар ең маңызды болып саналатынын көрсетеді. Модельдер үшін де тізімнің басында тұрған 5 және 6 есептегіштерге қосымша тізім айтарлықтай өзгерді. Айта кету керек, 5 және 6 есептегіштер инциденттер моделінде үлкен мәнге ие және 6-тақырыптан басқа тақырыптарда болмаған (бар диаграммасындағы сәйкес жолақтың көк пропорциясы арқылы байқалады). Қарапайым есептегіш модельде анықталған мәндер көптеген айнымалылар үшін шамамен тең, бұл әдеттегі телекоммуникациялық процедуралар елеулі проблемаларсыз жүретін желінің қалыпты жұмысын көрсетеді. Алайда, инциденттер моделінде құрылым құрамы мүлдем бөлек болып табылады, өйткені есептегіштердің мәндері айтарлықтай ерекшеленеді, бұл кейбір оқиғалардың басқаларына қарағанда жиі болғандығын білдіреді.

2.3 суретте көрсетілген шешуші факторлар тақырыптар мен олардың бір-біріне сәйкес орналасуы, олардың үйлестіру жүйесіндегі дүниежүзілік орналасуы емес 6, 7 тақырыптар мен 5 пен 6 арасындағы кеңістік ұлғайтылды, ал 5, 7 обьектілері стандартты модельге қарағанда көбірек сәйкес келеді. Оның трафик бұзушылықтары бар екендігі анықталғаннан кейін, басқа тақырыптар бойынша берілген мәліметтер негізгі себепті анықтауға көмектеседі. Стандарттар мен ерекше модельдер арасындағы тақырыптар бір-бірлеп талданды. Телекоммуникацияның тәжірибелі маманы жүргізген сараптамадан кейін, 5, 6 және 8-тақырыптар ерекше трафикті түсіндіруге жеткілікті қанағаттандырарлық ақпарат берді деген қорытындыға келеді.



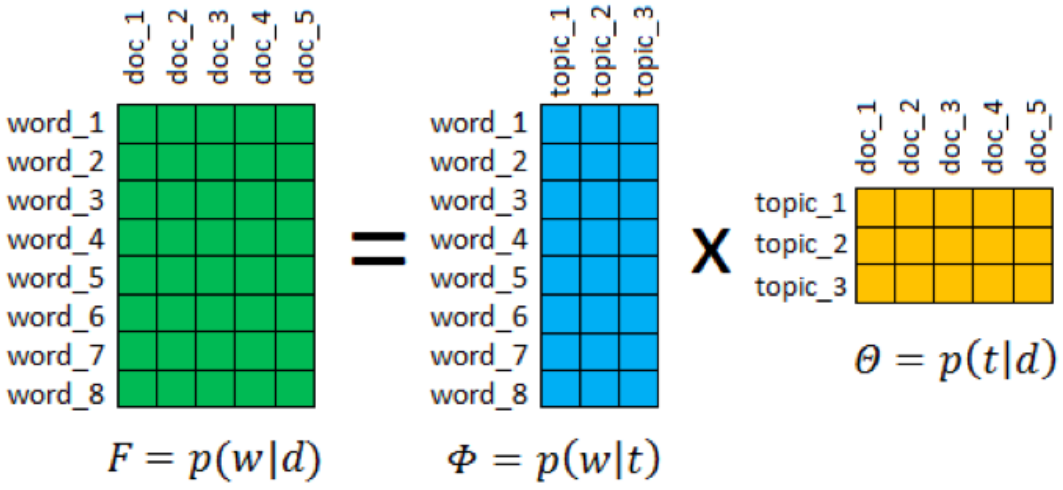
2.3 сурет - LDA моделі қалыпты және қалыпты емес уақытта жағдайында жасалған тақырыптар. 6 тақырып қызыл түспен таңдалып, есептегіштердің нәтижелеріне гистограмма көрсетілген. Көк жолақтар санауыштардың жиілігін, ал қызыл жолақтар тақырыптардың шамамен жиілігін көрсетеді. Телефон байланысының процестеріне байланысты оқиғалар А тобының есептегіштеріне тағайындалады, дабыл процестерін білдіретін оқиғалар В тобына тағайындалады [18]

Әдеттегі модельге қарағанда көбірек орын алатын және жиі пайда болатын оқиғалар саны туралы ақпарат байланыс операторы үшін маңызды мәліметтер болып табылады. Желінің басқа бөлігі табиғи, нақты қатынас стандарт ретінде қарастырылуы мүмкін. Демек, бұл трафикті өңдеу үшін әдеттегі модель немесе осы нақты стандартты жағдай туралы айтарлықтай тәжірибесі бар адам қажет.

2.5 Тақырыптық модельдердің аддитивті регуляризациясы (ARTM)

Тақырыптық модельдердің аддитивті регуляризациясы (ARTM) ықтималдық логарифмінің өлшенген қосындысын және қосымша критерий - регуляризаторларды көбейтуге негізделген. Бұл тақырыптық үлгілерді біріктіруді және кездейсоқ күрделі көп мақсатты модельдерді құруды жеңілдетеді. Көптеген танымал модельдер ARTM тұрғысынан регуляризатор

болып саналады. Деректерді матрица түрінде, ал матрицаның ыдырауы тапсырма түрінде береміз:



2.4 сурет- Тақырыптық модельдердің аддитивті регуляризациясы (ARTM) матрица түріндегі сұлбасы[3]

Тақырыптық модельдерді қолдану арқылы машиналық оқытудың ең танымал мәселелерінің бірі шешіледі - модельдің түсініктілігі. Φ матрицасы - тақырыптардағы сөздерді таратудың матрицасы, Θ матрицасы - құжаттардағы сөздерді тарату матрицасы. Ең ықтимал сөздер тізімінен осы тақырыптың не екенін түсініп, оған атау беруге болады. Содан кейін құжаттың тақырыптық профилін сипаттау оңай болады.

Нақты деректер бойынша тәжірибелерде тегістеудің, сиретудің және декорреляция регуляризаторларының комбинациясы зерттелген. Бірлесе отырып, олар тақырыптық модельдің нақтылығында іс жүзінде нашарламай, кеңістік, когеренттілік, тазалық және контраст өлшемдерін едәуір жақсарта алады.

Ықтималдылықпен қатар, r критерийлерін $Ri(\Phi, \Theta)$, максимизациялау керек деп есептейік, $i = 1, \dots r$ регуляризаторлар деп аталады [19]. Көптеген критерийлерді максимизациялау үшін біз критерийлердің сызықтық комбинациясын барынша көбейтеміз, $L\,(\Phi, \Theta)$ және $Ri\,(\Phi, \Theta)$ теріс емес регуляризация коэффициенттері τi:

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta), \tag{2.8}$$

$$L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}, \tag{2.9}$$

T тақырыбы, егер $n_{wt} + \varphi_{wt}\frac{dR}{d\varphi_{wt}} > 0$ болса, $w \in W$ кем дегенде бір мерзім үшін тұрақты деп аталады, әйтпесе t тақырыбы тым регулярлы деп айтамыз.

D құжаты, егер $n_{td} + \theta_{td}\frac{dR}{d\theta_{td}} > 0$ болса, $t \in T$ кем дегенде біреуі үшін тұрақты деп аталады, әйтпесе d құжаты қайта регулярлы деп айтамыз.

Регуляризацияланған ықтималдылық мәселесін шешу үшін модификацияланған М-қадам формулалары бар ЕМ алгоритмі қолданылады [20]:

Кіріс: D құжаттар жинағы, тақырыптар саны $|T|$;

Шығу: Φ, Θ;

1 баған векторларынφ_t , θ_dкездейсоқ түрде таңдап алу;

2 қайталау;

3 барлық $w \in W$, $t \in T$, $d \in D$ үшін n_{wt}, n_{td} нөлдеу;

4 барлық $w \in d$, $d \in D$ үшін:

5 $p(w|d) := \sum_{t \in T} \varphi_{wt}\theta_{td}$;

6 барлық $t \in T$ үшін:

7 $p(t|d,w) := \varphi_{wt}\theta_{td}/p(w|\,d)$;

8 n_{wt}, n_{td} $n_{dw}p(t|d,w)$ ұлғайту;

9 барлық $w \in W$, $t \in T$ үшін $\varphi_{wt} \infty n_{wt} + \varphi_{wt}\frac{dR}{d\varphi_{wt+}}$;

10 барлық $d \in D$, $t \in T$ үшін $\varphi_{wt} \infty n_{td} + \theta_{td}\frac{dR}{d\theta_{td+}}$;

ARTM-де регуляризаторларды біріктіру үшін жүйелеу стратегиясын ойластырған жөн:

- осы тапсырмада қандай регуляризаторлар қажет;
- қандай регуляризаторлар бір уақытта жұмыс істеуі керек, қайсысы кезекпен немесе кезекпен қажетті дайындық жұмыстарын жүргізеді
- итерация кезіндегі әр регуляризаторлар коэффициентін қалай өзгерту керек: қандай регуляризаторлар қосу, күшейту, әлсірету және өшіру қажет.

APTM тәсілінің шектеулеріне реттеу коэффициенттерін қолмен таңдау жатады. ARTM-де жүйелеу стратегиясын автоматты түрде түзету әлі күнге дейін ашық мәселе болып келеді.

2.6 ARTM-нің Байес модельдерінен артықшылығы

Ықтималдық тақырыптық модельдеу негізінен Байес және графикалық модельдер аясында дамып келеді. Байес тәсілінде мәтіндер жиынтығы ықтимал генеративті модельмен сипатталады, онда қосымша деректер ескеріліп, априори тарату арқылы ресімделетін қосымша шектеулер бар. Тақырыптық модельдеудегі Байес тәсілінің кемшіліктері:

- оптимизация өлшемдері арқылы көптеген модельдік талаптарды енгізу ыңғайлы болып табылады. Олардың априори таралуы тұрғысынан қайта құрылымдалуы Байес тұжырымын едәуір қиындатады [13]. Бірнеше априори үлестірім жиынтығын жалпы түрде тиімді ету мүмкін емес;

- Байес оқыту моделінің параметрлерін өздері емес, олардың таралуы анықтайды. Алайда, тақырыптық модельдеуде табылған бөлу тек математикалық күтімдерді бағалау үшін қолданылады. Осылайша, қажет болғаннан әлдеқайда қиын міндет шешіледі;

- көптеген Байес модельдері априори Дирихленің таралуын қолдануға мәжбүр. Математикалық тұрғыдан ыңғайлы, оның көпжақты таралуы бар мультиномиальды үлестірімге байланысты. Алайда, ол ешқандай табиғи тіл құбылыстарын модельдемейді және сенімді лингвистикалық негіздемелері жоқ. Сонымен қатар, ол Φ, Θ матрицаларында нөлдік мәндерден аулақ болу үшін сиретудің табиғи талабына қайшы келеді;

- Дирихленің үлестірімі тым әлсіз регуляризатор. Ол модельдің бірлігі мен тұрақсыздығы мәселесін шешпейді.

APTM тәсілінің артықшылықтары:

- ARTM-де регуляризаторлар априори үлестірімінен болмауы керек және ықтималды түсіндірме болуы керек;

- Дирихле регуляризаторы өзінің ерекше рөлін жоғалтады, оны барлық модельдерде кез-келген модельде қолдану қажет емес;

- математикалық аппарат өте қарапайым: регуляризатор қосу үшін оның туындыларын М-қадам формулаларына қосу жеткілікті;

- Байес көптеген тақырыптық үлгілерін (немесе оларға енгізілген идеяларды) регуляризаторлар арқылы қайта құруға болады;

- әртүрлі модельдерден алынған регуляризаторларды қорытындылай келе, көп мақсатты құрама модельдерді құруға болады;

- ARTM-дегі тақырыптық модельдерді түсіну оңай, шығарылуы оңай және біріктіру оңай.

Ілеспе салалардың зерттеушілері үшін тақырыптық модельдеу саласына кіру шегі азайтылды.

2.7 Мультимодальді тақырыптық модельдердің аддитивті регуляризациясы (ARTM)

Мультимодальды тақырыптық модельдер құжаттың метадеректерін ескереді - негізгі мәтінге қосымша ақпарат. Метадеректер құжаттың тақырыбын анықтауға, керісінше, мәтіннің мәтінінен құжаттың тақырыбын анықтай отырып, метадеректерді автоматты түрде құруға, жоқ метадеректердің орнын толтыруға және пайдаланушыларға ұсыныстар жасай алады.

Тақырыптық модельдер әр түрлі метадеректерді ескере алады: авторлар, құжатты немесе оның фрагменттерін құру уақыт белгілері, категориялар, кескіндер және жеке кескін элементтері, дәйексөз құжаттар, дәйексөз авторлары, құжаттарды пайдаланушылар және т.б.

BigARTM бірнеше типтегі метадеректерді бір уақытта өңдеуге мүмкіндік беретін мультимодальды модельдерді енгізеді [21]. Әрбір режим үшін мүмкін мәндердің сөздігі жасалады. Әрбір модуль элементтерінің енгізілуі мәтінге терминдер енгізілгендей қарастырылады. Шындығында,

терминдер (сөздер мен тіркестер) модальдылықтардың біреуі ғана. Мультимодальды тақырыптық модель әр тақырып үшін ықтималдылықтың дискретті үлестірімін әр модульдің барлық элементтерінің жиынтығына (сөздікке) салады.

2.8 Тақырыптық модельдерді регуляризациялау әдістері

Тегістегіш регуляризатор және LDA моделі. Φt және θd үлестірімдері $(\beta w)_{w\in W}$ және $\alpha = (\alpha t)_{t\in T}$сәйкесінше, Кульбак- Лейблер дивергенциясына жақын болуын талап етеміз:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T}\sum_{\omega\in d}\beta_w ln\varphi_{wt} + a_0\sum_{d\in D}\sum_{t\in T} a_t ln\,\theta_{td} \to max, \quad (2.10)$$

Регуляризатор KL дивергенциясы арқылы түсіндіру априори Дирихленің таралуынан гөрі табиғи болып көрінеді.

Кескіш регуляризатор. Әр құжат және әрбір термин тақырыптардың аз санына байланысты болады делік. Сонда «wt» және «ықтималдық» арасында көптеген нөлдер болуы керек. Үлкен топтамалардың тақырыптық модельдерін құру кезінде көптеген тақырыптар, матрицалардың ығысуы есте сақтау мен уақыт шығындарын азайтуға көмектеседі. Таралуы неғұрлым сирек болса, соғұрлым оның энтропиясы азаяды. Максималды энтропияның біркелкі таралуы бар.

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in T}\sum_{\omega\in d}\beta_w ln\varphi_{wt} - a_0\sum_{d\in D}\sum_{t\in T} a_t ln\,\theta_{td} \to max, \quad (2.11)$$

Энтропияны қалыпқа келтіру идеясы PLSA динамикалық тақырыптық моделінде бейне ағындарын өңдеуде тақырыптардың уақытша таралуын сұйылту үшін ұсынылды [22].

LDA кесуге бағытталған көптеген зерттеулер өте күрделі конструкцияларға әкеледі, өйткені кеңістік пен бөлу қасиеті арасында ішкі қайшылық бар [23]. Дирихле нөлдік ықтималдылыққа жол бермейді. Біздің жұқаруға деген көзқарасымыз әлдеқайда қарапайым және табиғи. Сондай-ақ, тегістеу мен кесу бірдей сипатталғанын ескертеміз. βw, α параметрлерінің белгілеріне шектеулер қойылмайды.

Тақырыптарды декорреляциялау регуляризаторы. Тақырыптардың әртүрлілігін арттыру модельдің түсініктілігін жақсартады деген пікір бар [2221]. Баған векторларының $\varphi t, \varphi s$ арасындағы ковариантты азайтуға мүмкіндік беретін регуляризатор:

$$R(\Phi) = -\gamma\Sigma_{t\in T}\Sigma_{s\in\frac{T}{t}}\Sigma_{w\in W}\,\varphi_{wt}\varphi_{ws} \to max, \quad (2.12)$$

Осы формула бойынша итерация кезінде w сөзінің маңызды тақырыптарының ықтималдығы одан сайын артады. Аз маңызды тақырыптардың ықтималдығы біртіндеп азайып, жоғалып кетуі мүмкін.

Осылайша, бұл реттегіш сирек кездеседі. Сонымен қатар, сөздерді бөлек тақырыптарға топтастырудың қосымша пайдалы қасиеті бар [22].

Құжаттарды ковариациялау регуляризаторы. Кейде ұқсас тақырыптағы құжаттар арасындағы сілтемелер туралы қосымша ақпарат болады. Атап айтқанда, олар бір тақырыпқа сілтеме жасай алады, көбінесе бір-біріне сілтеме жасайды немесе сілтеме жасайды. Біз бұл болжамды регуляризатор көмегімен ресімдейміз:

$$R(\Theta) = \tau \Sigma_{d,c} n_{dc} \Theta_{td} \Theta_{tc} \rightarrow max, \quad (2.13)$$

мұндағы, n_{dc} - құжаттар арасындағы байланыстың салмағы, мысалы, d-ден c-ге сілтемелер саны.

2.9 Тақырыптық модельдердің классификациясы

Тақырыптық модельдердің сапасын бағалау тривиалды емес мәселе болып табылады. Классификация немесе регрессиялық тапсырмалардан айырмашылығы, «қате» немесе «жоғалту» туралы нақты түсінік жоқ.

Кластерлеу және классификациялау объектілерінің ұқсастығы осы объектілердің қасиеттерімен анықталады. Мәтіндік құжаттардың негізгі қасиеті - мәтінде сөздің болуы. Мәтіндік құжаттардың қасиеттерін компьютерлік бейнелеу үшін тұжырымдама енгізіледі - векторлық кеңістіктің моделі. «Ортақ векторлық кеңістіктегі векторлар түрінде әртүрлі құжаттарды ұсыну векторлық кеңістік моделі болып табылады және көптеген іздеу тапсырмаларының, соның ішінде сұраныстағы құжаттардың, жіктелуі және кластерлері болып табылады».[2324]

Келесі жіктеу әдістері бөлінді:

- Байес әдісі;
- шешім қабылдау ағаштары;
- сызықтық және сызықтық емес әдістер;
- қодтау векторларының әдісі (Support Vector Machines, SVM);
- регуляризациялау жүйелері;
- дискреттеу және сирек кездесетін торлар.

Кластерлеу сапасының стандартты критерийлері, мысалы орта ішкі кластерлік немесе аралық кластерлік қашықтықтар немесе олардың қатынастары құжаттар мен терминдердің «жұмсақ» бірлескен кластерленуін бағалау үшін өте қолайлы [24].

Классификациялау әр құжат C жиынының элементтерінің жиынтығына сәйкес келеді, олар класстар, класс белгілері немесе жай белгілер деп аталады. Егер құжаттарда бірдей белгілер болса, онда олардың тақырыптары да ұқсас болады деп болжанады. Сондықтан, тегтер тақырыптардың түсініктілігін жақсарта алады, тіпті класстар мен тақырыптар арасында бір-біріне сәйкес келмейтін хат алмасулар болса да. Тапсырма класстар мен тақырыптар арасындағы байланыстарды анықтау, тақырыптық модельдің сапасын

жақсарту және белгілері әлі орнатылмаған жаңа құжаттарды жіктеу алгоритмін құру болып табылады.

Мәселенің күрделілігі сол, стандартты жіктеу алгоритмдері көп мөлшерде теңгерілмеген, бір-біріне тәуелді, өзара тәуелді сыныптардың үлкен мәтіндік жинақтарында қанағаттанарлықсыз нәтижелерді көрсетеді [48].

Теңгерімсіздік дегеніміз, сабақтарда өте кішкентай құжаттар да, өте көп сан да болуы мүмкін. Сыныптар бір-біріне сәйкес келетін жағдайда, құжат бір классқа немесе көптеген кластарға сілтеме жасай алады. Өзара тәуелді сыныптар ұқсас терминдердің жиынтығына ие және құжатты жіктеу кезінде бәсекелестікке түседі.

2.10 Тақырыптық модельдердің ішкі және сыртқы сапасын бағалау тәсілдері

Ең көп таралған критерий - бұл компьютерлік лингвистикада тілдік модельдерді бағалау үшін пайдаланылатын перплексия болып табылады [21].

Бұл $p(w|d)$ моделінің сәйкессіздік немесе «таңдану» өлшемі, D жиынының d құжаттарында сақталған, логарифм ықтималдылығы арқылы анықталған:

$$P(D;p) = exp\left(-\frac{1}{n}L(\phi,\theta)\right) = exp(-\frac{1}{n}\sum_{d\in D}\sum_{\omega\in d} n_{dw}\left(p(w|d)\right), \quad (2.14)$$

Бұл мән неғұрлым аз болса, p моделі D жиынтығының d құжаттарындағы терминдердің пайда болуын жақсырақ болжайды. Егер w термині электрлік сөздікте $p\,(w) = 1/V$ біркелкі таралуынан пайда болса, онда мұндай мәтіндегі p моделінің плексиясы ұзындығының ұлғаюымен V-ге өзгереді. P-дің таралуы неғұрлым күшті болса, біркелкі, соғұрлым қиын болады. P моделі генераторлық үлестіруден неғұрлым күшті болса, соғұрлым қиын болады. $p\,(w \mid d)$ терминдерінің шартты ықтималдығы қолданылады, ал интерпретация сәл өзгеше: егер әрбір құжат V бірдей ықтимал шарттардан құрылса (әр түрлі құжаттарда мүмкін әр түрлі), онда перплексия V-қа ауысады. Тағы да үлестірім біркелкі болған сайын, соғұрлым перплексия аз болады.

Тақырыптық модельдің түсініктілігі матрицаның a сиреттету құрылымына жақындығын сипаттайтын бірнеше критерийлер бойынша бағаланады. T тақырыбының Wt өзегін жоғары шартты ықтималдығы бар терминдер жиынтығы ретінде анықтаймыз $(p \mid t \mid w)$ = Осы тақырып үшін $p(t|w) = \varphi_{wt}\frac{n_t}{n_w}$:

$$pW_t = \{w \in W|p(t|w) > 0.25\}, \quad (2.15)$$

Ядроға сәйкес t тақырыбын интерпритациялаудың үш көрсеткішін анықтаймыз:

- $pur_t = \sum_{w \in W_t} p(w|t)$- тақырыптың тазалығы (соғұрлым жақсы);

- $con_t = \frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$- тақырыптың қарама-қайшылығы (соғұрлым жақсы);

- $ker_t = |W_t|$ ядро мөлшері (шамамен оңтайлы $|W|/)|T|$).

Егер белгілі бір тақырыпта жиі кездесетін терминдер жиынтық құжаттарда жиі кездесетін болса, тақырып үйлесімді деп аталады [25].

Эксперименттер барысында сарапшылардың бағалауы адамның араласуынсыз толықтай автоматты түрде есептелуге болатын үйлесімділік сияқты сапалы өлшеммен жақсы сәйкес келетіні анықталды.

Когеренттілік үшінші тараптың коллекциясы немесе модель құрастырылған сол коллекциямен бағалауға болады [26]. Біріздiліктің бірнеше бағасы ұсынылды. Алдымен нүктелік өзара ақпарат (PMI) қолданылды [27].

$$PMI(t) = \sum_{i=1}^{k-1} \cdot \sum_{j=1}^{k} log \frac{N(w_i w_j)}{N(w_i)N(w_j)},$$

(2.16)

мұндағы, w - termwt азайту кезіндегі оныншы мерзім, N (w) - w термині кемінде бір рет кездесетін құжаттардың саны, $N\ (w, w')$ - w, w терминдері қатар-қатар пайда болатын құжаттардың саны. кем дегенде бір рет, k саны әдетте 10 деп қабылданады. «Жанама кездесу» дегеніміз - берілген h ені терезесінде, параметр, әдетте h = 10.

Содан кейін, тәжірибелерде когеренттіліктің неғұрлым адекватты өлшемі логикалық шартты ықтималдық (LCP) болатындығы, ол жиірек кездесетін жағдайдағы аз сөздің ықтималдығын бағалайтындығы көрсетілген:

$$LCP(t) = \sum_{i=1}^{k-1} \cdot \sum_{j=1}^{k} log \frac{N(w_i w_j)}{N(w_i)N(w_j)}, \quad (2.17)$$

Ықтималды тақырыптық модельдеу сапасын бағалаудың басқа әдістері бар, мысалы, сараптамалық бағалау, бірақ олар аз танымал. Ықтималды тақырыптық модельдеу бойынша зерттеу жүргізу үшін және ықтималдық тақырыптық модельдеудің сапасын бағалау үшін модель құрастырылатын тілдегі мәтіндік корпустар қажет екенін атап өткен жөн.

3 Мәтіндік құжаттар ағынын ықтималды тақырыптық модельдеу

3.1 Python бағдарламалау тілі

Python - жоғары деңгейлі, еркін, интерпретацияланған, объектіге бағытталған, кеңейтілетін, кіріктірілген бағдарламалау тілі [28].

- ақысыз - барлық аудармашылардың бастапқы кодтары мен кітапханалары кез келген, соның ішінде коммерциялық мақсатта қол жетімді;

- интерпритацияланған - өйткені «кеш байланыстыру» қолданылады;

- объектіге бағытталған - классикалық моделі, оның ішінде бірнеше мұрагерлікті қамтиды;

- кеңейтілетін - C немесе C ++ тілдерінде модульдер, типтер мен класстар құруға арналған қатаң түрде анықталған API интерфейсі бар;

- ендірілетін - аудармашыны басқа бағдарламаларға енгізу үшін қатаң түрде анықталған API интерфейсі бар;

- өте жоғары деңгей - динамикалық теру, жоғары деңгейлі мәліметтер типтері, класстар, модульдер, ерекшелік механизмі.

Python - бұл әмбебап тіл, ол бүкіл әлемде әртүрлі мақсаттарда қолданылады - мәліметтер базасы және мәтінді өңдеу, ойындарға интерпретаторлар ендіру, GUI бағдарламалау және жылдам прототиптеу (RAD).

Python Интернетті және веб-қосымшаларды - серверді (CGI), клиентті (роботтарды), веб-серверлерді және қолданбалы серверлерді бағдарламалау үшін қолданылады. Бұл тілде өте кең кітапхана бар және басқа тарап мамандары әзірлеген көптеген модульдер жиынтығы бар. Python мен оған жазылған қосымшаларды ең танымал және ірі компаниялар пайдаланады. [29]

Бұл тілде жазылған:

- Mailman - GNU жобасының тарату тізімдерінің ресми менеджері атанған тарату тізімінің менеджері;

- Medusa - HTTP, FTP, NNTP, XML-RPC және SOAP сияқты жоғары сапалы TCP / IP серверлерінің архитектурасы;

- Zope - кең танымалдыққа ие болған веб-қосымшалар сервері (Web application server).

Python қолдануы оңай, бірақ бұл программалық қамтамасыздандыру тілі, ол қабыққа қарағанда үлкен бағдарламаларды құруға және қолдауға арналған көптеген құралдарды ұсынады. Python бағдарламаларын интерпретаторда орындайды.

Әдетте python командасын енгізу арқылы интерпретатор шақырылады. Сондай-ақ, интерпретаторлардың көптеген іске асырулары және әртүрлі даму орталары бар (мысалы, Jython, IronPython, IDLE, ActivePython, Wing IDE, pydev және т.с.с.), сондықтан қоңырау шалу тәртібі туралы ақпарат алу үшін құжаттамадан кеңес алу керек. [30]

Python интерпретаторының интерактивті режимі - бұл пайдалы функциялардың бірі. Интерактивті қабыққа кез-келген жарамды нұсқауларды немесе олардың тізбегін енгізіп, бірден нәтиже алуға болады. Екінші жағынан,

ол С тілінен гөрі қателіктерді жақсы өңдейді және жоғары деңгейлі тіл ретінде оған әр түрлі мәліметтер кіреді, мысалы, С-де тиімді іске асыру сізге көп уақытты қажет етеді. Жалпы мәліметтер түрлерінің арқасында Python Awk және Perl-ге қарағанда кең ауқымды міндеттерге қолданылады, ал Python-да көптеген заттар оңай.

Python бағдарламаларды кейіннен басқа бағдарламаларда қолдануға болатын модульдерге бөлуге мүмкіндік береді. Тіл кәдімгі модульдердің үлкен кітапханасымен бірге келеді, оны бағдарламаларға негіз немесе тіл үйренудің мысалдары ретінде пайдалануға болады. Стандартты модульдер файлдармен, жүйелік қоңыраулармен, желілік қосылыстармен және тіпті әртүрлі графикалық кітапханаларға интерфейстермен жұмыс істеуге арналған құралдарды ұсынады. Python өте ықшам және оқылатын бағдарламаларды жазуға мүмкіндік береді. Python-да жазылған бағдарламалар, әдетте, С немесе С ++ тілдеріндегі баламаға қарағанда әлдеқайда қысқа: бірнеше деңгейге байланысты мәліметтердің жоғары деңгейлері күрделі операцияларды бір нұсқаулықпен көрсетуге мүмкіндік береді, нұсқаулықтарды топтау бұйра жақшалар орнына шегіністермен орындалады, айнымалыларды жариялаудың қажеті жоқ.

3.2 Python кітапханалары

Соңғы 10 жыл ішінде Python бағдарламалау тілі ғылыми талдау мен есептеудің, сонымен қатар үлкен деректер массивтерін визуализациялаудың дамыған құралына айналды [28]. Python бағдарламалау тілі негізінен үшінші тарап әзірлеушілері жасаған ашық бастапқы пакеттердің үлкен және тез дамып келе жатқан модулдерінің арқасында қолданылады:

- NumPy кітапханалары - мәліметтермен, деректер массивтерімен жұмыс істеу үшін;
- Pandas кітапханалары - әртүрлі және нақты мәліметтермен жұмыс істеу үшін;
- SciPy - ғылыми есептеу үшін;
- Matplotlib кітапханалары - мәліметтерді графика түрінде және түрлі әдістермен бейнелеу үшін;
- IPython қабықтары - интерактивті кодтауға және басқалармен код алмасуға арналған;

NumPy кітапханасында деректерді сақтауға болады. NumPy (сандық python үшін қысқа) үлкен деректермен жақсы жүреді. NumPy кітапханалық массивтері Python-ның кірістірілген деректер типіне ұқсас, бірақ деректерді сақтау мен тиімді жұмысты қамтамасыз етеді, ал жылдамдық массивтің көлеміне қарай артады.

NumPy кітапханасы бұл үшін керемет, бірақ оның шектеулері бар, олар бізге біршама икемділік қажет болғанда байқалады (деректерді белгілеу, жоқ мәліметтермен жұмыс істеу және т.б.). Бұл шектеулер, сонымен қатар, жарты биттік аударуға сәйкес келмейтін әрекеттерді орындау кезінде пайда болады (топтау, айналым кестелерін құру және т.б.). Мұндай операциялар қоршаған

әлемнің көптеген нысандарында аз дәрежеде құрылымы бар мәліметтерді талдаудың маңызды бөлігі болып табылады.

Pandas кітапханасы кестеде (құрылымдалған) деректермен жұмыс істеудің көптеген әдістерін ұсынады, ол SQL-ге ұқсас әртүрлі сұраулар мен кесте өзгертулерін жасауға мүмкіндік береді. Онда әрбір деректер бағанының өзіндік типі болады (бүтін сандар, өзгермелі нүкте нөмірлері, күндер және жолдар). Екінші жағынан, pandas кітапханалары кең көлемді файлдармен және деректер қорымен жұмыс істейді, мысалы SQL, Excel және scv файлдары мен архивтерімен бірдей. Pandas кітапханасы кестеде (құрылымдалған) деректермен жұмыс істеудің көптеген әдістерін ұсынады, ол SQL-ге ұқсас әртүрлі сұраулар мен кесте өзгертулерін жасауға мүмкіндік береді. Онда әрбір деректер бағанының өзіндік типі болады (бүтін сандар, өзгермелі нүкте нөмірлері, күндер және жолдар). Екінші жағынан, pandas кітапханалары кең көлемді файлдармен және деректер қорымен жұмыс істейді, мысалы SQL, Excel және scv файлдарымен, сонымен қатар архивтермен.

Matplotlib - бұл деректерді бейнелеуге арналған бай кітапхана, сол NumPy кітапханаларында салынған және SciPy кең стекімен жұмыс істейді [7]. Matplotlib кітапханасы 2002 жылы Джон Хантермен жасалып, IPython патчін енгізді, бұл MATLAB типіндегі графиктерді gyuplot үйлесімінде IPython командалық жолынан интерактивті түрде құруға мүмкіндік берді.

Matplotlib көптеген әртүрлі операциялық жүйелермен, әртүрлі қосымшалардың графикалық компоненттерімен үйлеседі. Matplotlib қосымшаның және шығыс түрлерінің бөліктерінен мәліметтерді жасайды, яғни сіз олармен жұмыс істей аласыз, амалдық жүйенің түріне немесе шығыс форматына қарамастан. Кітапханада әмбебап кросс-платформа қолданылады, ол пайдаланушылар санын көбейтті, әзірлеушілердің ағымын, matplotlib кеңейту мүмкіндіктерін және әлемдегі python ғылыми қауымдастығының таралуын қорытындылады.

IPython Shell - бұл пайдалы интерактивті python тілдік интерфейсі. Оның көптеген кең синтаксистік мүмкіндіктері бар. IPython браузерге блокнот немесе мәтіндік редактор ұсынатын Jupiter жобасымен байланысты. Бұл ресурстарды дамыту, бірлесіп жұмыс істеу және пайдалану, ғылыми нәтижелерді жариялау үшін ыңғайлы. IPython блокноттары - Julia, R сияқты программалау тілдеріне арналған блокноттарды қамтитын Jupiter блокнотының жалпы құрылымының ерекше жағдайы. IPython мәліметтерді үлкен көлемде өңдеуді қажет ететін интерактивті ғылыми есептеу үшін python тілін тиімді пайдалануға мүмкіндік береді [6].

Jupiter Notepad - бұл браузерге негізделген GUI, IPython қабығына және динамикалық визуализацияның бай жиынтығына арналған.

3.3 Мәтіндік құжаттар мен мәтіндер ағындарын талдау

«Кластеризация мақсаты – объектілерді «ұқсас » объектілер топтарына жіктеу, бұл топтарды кластерлер деп атайды. Кластеризация туралы есептерде әрбір мәліметтер нысаны бұрын анықталмаған бір немесе бірнеше кластарға

тағайындалады. Деректер пішіндерінің кластерлерге бөлінуі олардың қалыптасуы кезінде жүреді. Кластерлердің анықтамасы және мәліметтер нысандарының таралуы соңғы деректер үлгісінде көрсетілген, бұл кластерлер мәселесін шешеді»[8]. Құжаттың кластеризациясы дегеніміз - белгілі бір ұқсастық өлшеміне негізделген ұқсас белгілері жоқ құжаттарды топтау.

Деректер анализінің негізгі кезеңдері қамтыды:

- анализ тапсырмасын түсіну және тұжырымдау;
- автоматты талдауға деректерді дайындау (алдын-ала өңдеу);
- деректері құру (Data Mining) әдісін қолдану және модель құру;
- құрылған модельді тексеру және бағалау;
- адамдардың модельдерді интерпретациялауы.

Мәліметтердің визуалды анализі, мәтіндік құжаттардағы деректерді анықтау, нақты уақыт режимінде деректерді талдау әдістері, веб-құжаттар мен сайттарды талдау бар. Деректерді визуалды түрде өңдеу (Visual Mining) - бұл деректерді көрнекі түрде ұсыну, бұл адамға деректерге қанығуға, көрнекілікпен жұмыс істеуге, олардың мәнін түсінуге, қорытынды шығаруға және деректермен тікелей өзара әрекеттесуге мүмкіндік береді.

Құрылымсыз мәтіндердегі талдау әдістері бірнеше бағыттардың түйіскен жерінде орналасқан: Data Mining, тілдер табиғи өңдеу, ақпаратты іздеу, ақпарат алу және басқару. Мәтін өндіру (Text Mining) - бұл құрылымдалмаған мәтін деректерінде шынымен жаңа, ықтимал пайдалы және түсінікті заңдылықтарды ашуға арналған тривиалды емес процесс.

Machine Learning термині барлық деректерді өндіру технологияларына қолданылады [8]. Машиналық оқыту қолдану дегеніміз - алгоритмдер мен жүйелерді жүйелі түрде оқыту, нәтижесінде олардың білімі немесе жұмыс сапасы тәжірибемен артады. Машиналық оқытудың мәні дұрыс қойылған есептерді шешуге қолайлы модельдерді құру үшін қажетті мүмкіндіктерді пайдалану болып табылады.

Мәтіндік құжаттарды талдауға арналған жүйелер мен бағдарламалық қамтамасыздандыру:

BigARTM - ықтималды тақырыптық модельдеуге арналған кітапхана ARTM әдісін енгізеді. Мәтіндік құжаттардың үлкен көлемін жұмсақ кластерлеуге мүмкіндік береді.

Мәтіндік құжаттар ағынын талдаудың негізгі белгілері - жаңа құжаттардың өңделу жылдамдығы және уақыт өте келе тұжырымдамалардың өзгеруі, ағым эволюциясы. Жаңа құжаттарды қабылдау Талдау жылдамдығы жылдамдықтан жоғары болуы керек.

Модульдік ARTM технологиясын енгізу үлкен деректерді тиімді өңдеуге мүмкіндік береді. BigARTM бұл үшін:

- орталық процессордың өзектерінде параллелдеу;
- бір реттік қажет етпейтін деректерді пакеттік өңдеу;
- жедел жадқа үлкен деректерді жүктеу;
- тиімді сызықтық есептеу алгоритмі;
- жинақ көлеміндегі және тақырыптар санындағы күрделілік;

- ең жиі жаңартылатын деректерді сақтау – тарату;
- тақырыптардағы сөздер - барлығы жедел жадыда;
- сәйкес C ++ тіліндегі негізгі кітапхананы енгізу;
- индустриалды бағдарламалаудың заманауи стандарттары;
- тәжірибелер BigARTM бірнеше есе алда екенін көрсетті;
- есептеу жылдамдығы тұрғысынан танымал алгоритмдер;
- еркін қол жетімді кітаптар Gensim және Vowpal Wabbit.

3.4 Тақырыптық модельге арналған деректер

Тақырыптық модель үшін деректерді ұлттық байланыс операторы АҚ «Қазақтелеком» ұсынды. Барлық мәліметтер 40 ГБ деректердің мұрағаты болды.

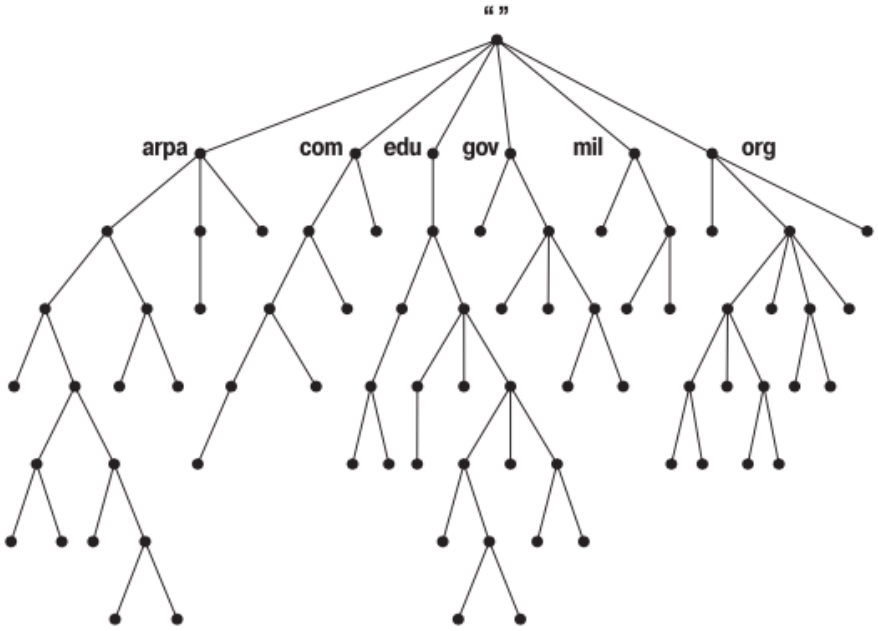
Интернеттегі компьютерлер арасында байланысатын барлық қосымшалар байланыс түйіндерін анықтау үшін IP мекенжайларын пайдаланады. Интернет пен әртүрлі компьютерлер арасындағы байланысты қамтамасыз ету үшін сәйкестендіру үшін IP мекенжайларын қолданамыз. IP қарапайым пайдаланушылар үшін есте сақтау қиын, сондықтан біз оның орнына интерфейс атауын қолданамыз. IP мекенжайлары бар барлық құрылғылардың жеке домендік атауы болады. Біз домен атауын IP мекенжайлары қажет жерде қолданамыз, жарамды атау серверін қоспағанда. Әрбір IP мекенжай бірнеше домендерді қамтуы мүмкін.

Компьютердің атауы мен IP мекен-жайы арасындағы байланыс домендік атаулар жүйесінің деректер базасында анықталған. Деректер базасы бүкіл әлемде бар. Ол ресурстардың жазбаларын қамтуы мүмкін. DNS дерекқорының жеке жазбалары бар, олар белгілі бір серверлерде орналасқан. DNS - бұл жаһандық таратылған мәліметтер базасы [31].

Атаулар серверлерінің жадында Resource Records (RR) домендік атаулар, IP туралы мәліметтер болады. Деректер атау сервері немесе DNS сервері арқылы кэшке әр түрлі жолмен жүктеледі. Сенімді деректер файлдардан оқылады. Олар дискіде болуы немесе желідегі беделді серверлерден оқи алады. Бізде басқа серверлерден рұқсат етілмеген мәліметтер бар. Дискіден жүктелетін арнайы деректер бар. DNS клиентіне DNS-тен ақпарат қажет болуы үшін, клиент DNS-тен PP сұрай алады. Мысалы, R типіндегі домен, сервер сервері немесе шешуші. Рекурсивті сұранысты өңдеу үшін DNS сервері бірнеше сұраныстарды өз бетінше жасауы керек. Алайда, серверлік процесс көптеген домендік атаулар туралы мәліметтерді алады. Сілтемеде DNS серверлерінің тізімін алған сайын ол белгілі бір аймаққа рұқсат етілген серверлерді қарап, сол серверлердің мекен-жайларын табады. Шешім қабылдау процесі аяқталғаннан кейін және бастапқы ақпарат жіберілген клиентке қажетті ақпарат қайтарылғаннан кейін, жаңа білімді кейін пайдалану үшін сақтауға болады. Ең танымал - бұл домендік атқа жақын ауданда жақсы беделі бар серверлер.

UNIX файлдық жүйелері сияқты, олардың да өз түйіндері бар, олардың жеке мәтіндік белгілері бар. Әрбір кеңейтілген ағаш мәтінді 63 таңбалы

метамәтінмен алмастырады және нүктелі үтір қолданылмайды. Сонымен қатар, ағайынды түйіннің шектеулері домендік атауымен бір ағаш түйінін біріктіреді. Деректер қоры инверттелген ағаш сияқты.



3.1 сурет- DNS атаулар кеңістігі [31]

DNS серверлерінің домендік атаулары бар. DNS хаттамалары әр түрлі операцияларды өңдейді, олардың ең көп таралғаны кіріс сұраныс болып табылады. Бұл сізге бір сұраудан бірнеше сұранысты алуға мүмкіндік береді. Қазіргі уақытта көптеген операция түрлері және DNS сұраулары бар. Қазіргі уақытта сервер 53 портына UDP арқылы сұраныс жіберуде. Егер сұраныс 512 байттан асатын болса, ол TCP хаттамасының көмегімен өңделеді [31].

Әрине, DNS серверлері деректерді мәңгі кэштай алмайды. Әйтпесе, беделді серверлердегі өзгертулер ешқашан Интернетте таралмайды. Қашықтағы серверлер кэштелген ақпаратты пайдалануды жалғастыруда. Сондықтан, деректері бар аймақтың әкімшісі, әдетте, сол деректердің қызмет ету мерзімін (TTL) анықтайды. Жарамдылық мерзімі - тегін DNS серверіне кэштелген деректерді пайдалануға рұқсат етілген уақыт мөлшері. Осы уақыттың соңында сервер кэштелген ақпаратты жойып, авторластырылған DNS серверлерінен жаңа ақпаратты алуы керек.

3.5 Деректерді алдын-ала өңдеу

Деректер домендік атаулардың көпшілігінде 2-ден 6-ға дейінгі деңгейден тұрады. Домендік атаулармен қатар, осы домендік атауларды сұраған бірегей пайдаланушылардың IP мекен-жайлары, доменге сұраныс уақыты, порт көзі және қосымша деректер болды.

Бастапқыда модельде пайдалы деректерді алу үшін деректер оқылды. Тақырыптық модельді құру үшін пайдаланушылардың IP-мекен-жайларын және домен атауларын таңдалды.

Жақсырақ түсіну үшін домен атаулары үшінші деңгейлі домендік атауларға, сонымен қатар екінші деңгейлі домендік атаулар да болды. Домендік атауларда жоғары деңгейдегі жалған домендер .local, .localdomain домендері болды. Олар жергілікті желіде zeroconf технологиясының DNS (mDNS) мультикасттық хаттамаларындағы хосттарды анықтау үшін қолданылады. mDNS Bonjour (MacOS X) және Avahi (Linux және BSD) қолданылады. Егер домен атауын анықтау мүмкін болмаса, компьютер өзін hostname.local ретінде анықтайды. Сонымен қатар, .arpa домендік атаулары алынып тасталды, олар интернет-инфрақұрылымы үшін арнайы қолданылатын жоғары деңгейдегі қарапайым домен болып табылады. Қазіргі уақытта бұл домен үшін келесі екінші деңгейлі домендер анықталған:

- e164.arpa - DNS (ENUM) телефон нөмірлерін көрсету;
- in-addr.arpa - кері DNS сұраулары үшін (IPv4 мекен-жайы);
- ip6.arpa - кері DNS сұраулары үшін (IPv6 мекен-жайы);
- uri.arpa - адрестік URI схемаларын динамикалық анықтауға арналған;
- urn.arpa - URN мекен-жай схемаларын динамикалық анықтауға арналған.

Осыдан кейін FQDN стандартына сәйкес келмейтін сұраныстардан тазала керек болды (олар домендік атаулар емес). Оны жүзеге асыру үшін регулярлы өрнек қолданылды [32].

3.1 кесте - Мәтінді алдын-ала өңдеуге арналған регулярлы өрнектер

Регулярлы өрнек	Іздеу мақсаты
'^([0-9a-z]*[-\w]*[0-9a-z]\.)+[a-z0-9\-]{2,15}$'	FQDN блогы

3.2 кесте – Өңделген деректер

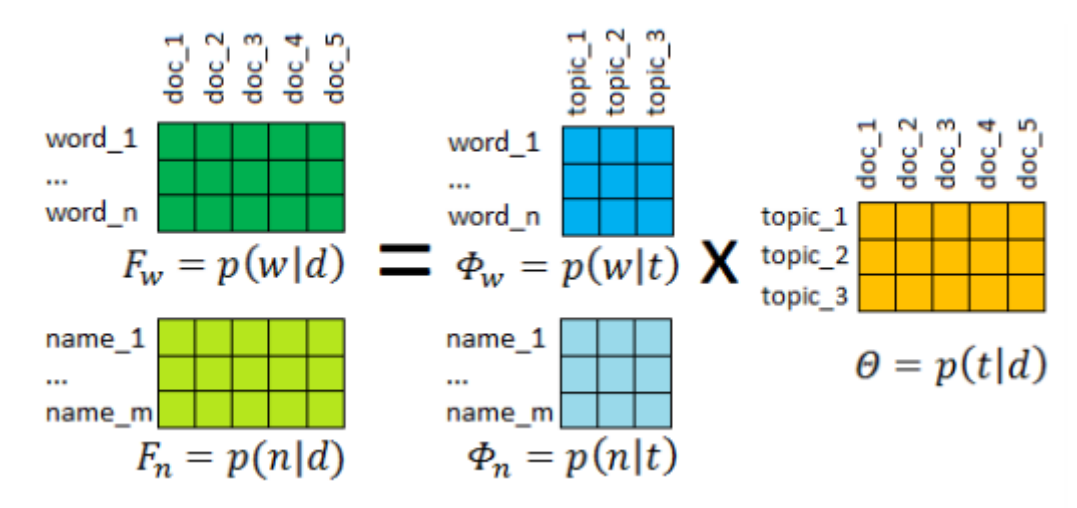
	IP	DNS
0	178.90.248.90	m31.zpmtmed.net
1	145.255.167.247	stun.l.google.com
2	2.132.10.245	www.google.com
3	95.59.77.211	a.root-servers.net
4	5.250.150.20	reserve-gb.apple.com

3.6 Тақырыптарды классификациялау

Екі өте күшті шектеулерге негізделген қарапайым тақырыптық классификация моделінен бастайық.

Тақырыптар кластармен анықталады, $C \equiv T$. Бұл сабақтардың күшті жорамалы, өйткені шартты тәуелсіздік талабы $p\,(w \mid d, c) = p\,(w \mid c)$, жасырын тақырыптар үшін жарияланған, бақыланатын сабақтар үшін қанағаттанбауы мүмкін.

Әрбір d құжат үшін $Cd \in C$ барлық кластары жиынтығы нақты белгілі. Бұл болжам тек белгілі бір тапсырмалар түрлері үшін жарамды. Уақыт пен авторларға - қолайлы; сілтемелер, санаттар, пайдаланушылар және көптеген басқа түрлер үшін - сәйкес келмейді. Мультимодальды тақырыптық модельдер процесі тек сөздерді ғана емес, сонымен қатар біздің жағдайдағыдай басқа да көптеген белгілерді қамтитын құжаттар.



3.2 сурет- Екі модальді тақырыптық модель[33]

Е-қадам және М-қадам формулалары да көп өзгертілмеген: $p\,(c \mid d, w)$, $n\hat{}dc$, θcd барлық $c \in C$ үшін есептелмейді, тек $c \in Cd$ құжат кластары үшін есептеледі.

Бұл модель үшін шартты тәуелсіздіктің екі гипотезасы орналастырылған:

- $p\,(w \mid t, c, d) = p\,(w \mid t)$ - сөздердің таралуы толығымен тақырып бойынша анықталады;

- құжат және құжаттың өзіне және оның сыныптарына тәуелді емес;

- $p\,(t \mid c, d) = p\,(t \mid c) - d$ құжатының мәні құжаттың өзіне тәуелді емес, бірақ ол қай класқа жатады;

- $p\,(c \mid t, d) = p\,(c \mid t)$ - алдыңғы жағдайға балама шарт - классификация

- D құжаты құжаттың өзіне тәуелді емес, тек оның тақырыбына байланысты.

Деректер, сонымен қатар, құжаттың метадеректерін - авторларды қамтиды. Авторлар - пайдаланушылардың IP мекен-жайы.

Деректер корпусы - белгілі ережелерге сәйкес құрылған проблемалық аймақтан алынған мәліметтердің үлгісі.Мәтіндік корпус - бұл мәтіндік мәтін немесе олардың едәуір маңызды фрагменттері, мысалы, осы проблемалық аймақтағы мәтіндердің макроқұрылымының кейбір толық фрагменттерін құрайтын мәліметтер корпусының түрі. Пайдаланушы тұрғысынан мәтіндердің мазмұнына қойылатын талаптар: анықтылық, толықтығы, пайдалылығы, материалды құрылымдау, компьютерлік қолдау.

3.7 Мультимодальды тақырыптық модельдерді регуляризациялау

Тақырыптық модель құру үшін деректерді BigARTM оқу форматына түрлендіру керек болды. Қазіргі уақытта барлық форматтар сөздердің жиынтығына (Bag-of-words representation) сәйкес келеді, яғни барлық лингвистикалық өңдеулер BigARTM-тен тыс жерде жасалуы керек. Vowpal Wabbit - бұл келесі принциптерге негізделген бір форматты файл:

- әрбір құжат бір жолдан тұрады;
- барлық токендер жол түрінде берілген (оларды бүтін санға ауыстырудың қажеті жоқ);
- токен жиілігі әдепкі бойынша 1.0, және қос нүктеден кейін міндетті түрде көрсетілуі мүмкін (:);
- атаулар кеңістігін (Batch.class_id) (|) арқылы анықтауға болады (|).

BoW (Bag of Words) моделі деректерді интерпретациялаудың ең жеңілдетілген алгоритмдерінің бірі болып табылады. «Сөздер қабы» атауы алгоритмнен шығады, ол мәтінде берілген сөздің қанша рет кездесетінін білуге тырысады. Бұл жерде сөздердің реті немесе контексті маңызды емес [33].

Деректер «сөздер қабы» ретінде:

145.255.160.118 |text www.ok.ru www.ok.ru www.ok.ru www.ok.ru i.instagram.com www.ok.ru www.ok.ru applog.uc.cn dpu.samsungelectronics.com ad.mail.ru ad.mail.ru r.mail.ru ad.mail.ru r.mail.ru rs.mail.ru rs.mail.ru rs.mail.ru an.yandex.ru an.yandex.ru an.yandex.ru an.yandex.ru www.google.com fna.fbcdn.net s1.wwhbundles.com cdn50.xcdn.me woodpecker.uc.cn rs.mail.ru r.mail.ru pimg.mycdn.me pimg.mycdn.me pimg.mycdn.me www.ok.ru www.ok.ru www.ok.ru www.ok.ru www.ok.ru www.ok.ru mqtt-mini.facebook.com ff.avast.com ff.avast.com emupdate.avcdn.net emupdate.avcdn.net ff.avast.com |subscriber 145.255.160.118

Барлық зерттеулерде тақырыптар саны | T | = 30 бекітілді, итерация саны 30-ға тең. Жинақтың жалпы ұзындығы $n \approx 5{,}4 \cdot 10^6$ сөзден тұрады. Сөздік көлемі $|W| \approx 9{,}6 \cdot 10^5$ құрады. Деректерді оқыту үшін оффлайн EM-алгоритмді таңдалды:

- коллекция бойынша қайталанатын итерация;
- құжат арқылы жалғыз өту;
- матрицаны сақтау қажеттілігі;

- φ коллекцияның әр өтуінің соңында жаңартылады;
- ол кішігірім коллекцияларды өңдеу кезінде қолданылады.

Е-қадамында:

$$\varphi_{wt} = norm_{w\epsilon W}(n_{wt} + \tau f(\varphi_{wt})d_w), \quad (3.1)$$

Берілген дистрибутивтің әртүрлі жиынтықтарын тегістеуге / бөлуге мүмкіндік береді. Сөз тарату түрін басқару үшін d_w сөздігі мен f функциясын қолдануға болады:

- d_w - бұл сөздік объектісі, жинақтау туралы ақпарат және әр сөзге қосымша өзгертілетін факторлар бар.

- f функциясы – φ_{wt} ағымдағы мәніне өзінің реттелуіне әсер етуге мүмкіндік беретін қайта құру.

М-қадамында:

$$\theta_{wt} = norm_{t\epsilon T}(n_{td} + \tau \propto_i f(\theta_{td})m_{td}), \quad (3.2)$$

- берілген дистрибутивтің әртүрлі жиынтықтарын тегістеуге / бөлуге мүмкіндік береді.

- $\propto_i$ параметрі реттегіштің берілген ішкі итерация кезінде әсер ету дәрежесін реттеуге мүмкіндік береді і

Тақырыптар мен құжаттардың таралуын бақылау үшін сіз мыналарды пайдалана аласыз:

- вектор немесе m_{td} матрица (ол туралы құжаттамада егжей-тегжейлі жазылған, қосымша фактор ретінде жұмыс істейді)

- f функциясы регуляризаторлар ағымындағы мәніне оның реттелуіне әсер етеді.

BigARTM-дегі сөздіктер үлкен рөл атқарады, олар қолданылады: тақырыптық үлгіні бастау, сапа көрсеткіштері, кейбір регуляризатор үшін. Сөздіктер туралы құжаттаманың бірнеше бөлімінен оқуға мүмкіндік береді, Python-дағы сөздікті дискіге сақтауға болады:

- artm.Dictionary.save_text (файл атауы);

- load_text () кері жүктеу.

Мәтіндік түрде Dictionary дегеніміз - жолдар жиынтығы, әр жолға (бірінші тақырыптан басқа) сәйкес келетін коллекциядағы бірегей сөз.

BigARTM сапа метрикаларын (scores) және өзіндік метрикаларды қосуға мүмкіндік береді. Деректерді оффлайн оқытқаннан кейін метрикаларды көре аламыз:

- перплексия, тілдік модельдер сапасының жалпы қабылданған өлшемі;

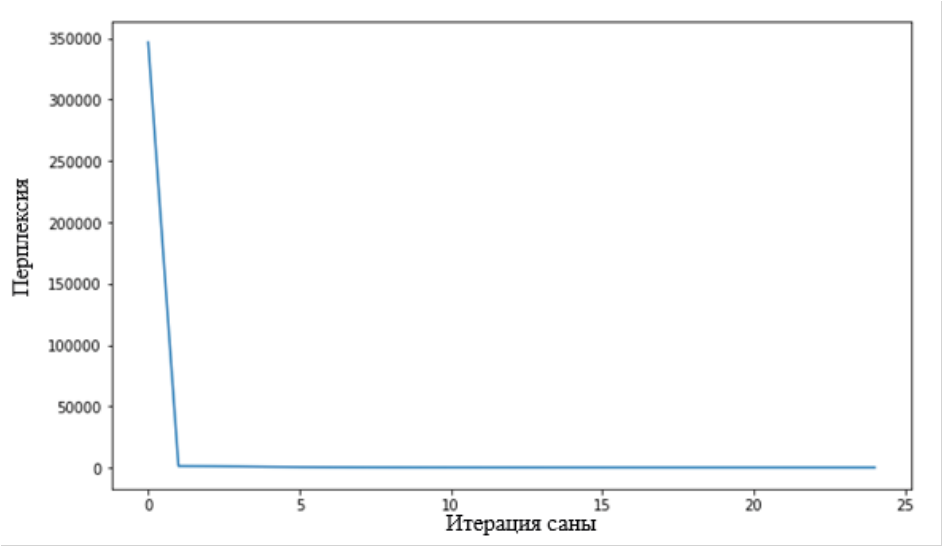
- сирек болуы, матрицада ықтималдықтардың үлес салмағы нөлге жақын;

- сәйкесінше Φ немесе Θ;

- тазалық пен контраст, тақырыптар арасындағы айырмашылықты бағалау. Бұл тақырыпты түсінудің жалпы қабылданған өлшемі.

Сапа метрикалары әр өңделген деректер пакеті үшін әр итерация кезінде қайта есептеледі.

Деректерді оқыту кезінде ең алғашқы метрикалардың бірі - бұл перплексия көрсеткіші (3.3 сурет). Мұны мәтіндегі сөздердің белгісіздігі немесе айырмашылығы деп айтуға болады. Сөздердің таралуы болса біркелкі емес, біркелкі бөлуге мүмкіндік беретін мәнмен салыстырғанда күдік азаяды. Сондай-ақ, перплексия - бұл мәтіннің тармақталу коэффициенті, яғни құжаттағы әр сөзден кейін орташа есеппен әр түрлі күтілетін сөздер саны деп айтуға болады.

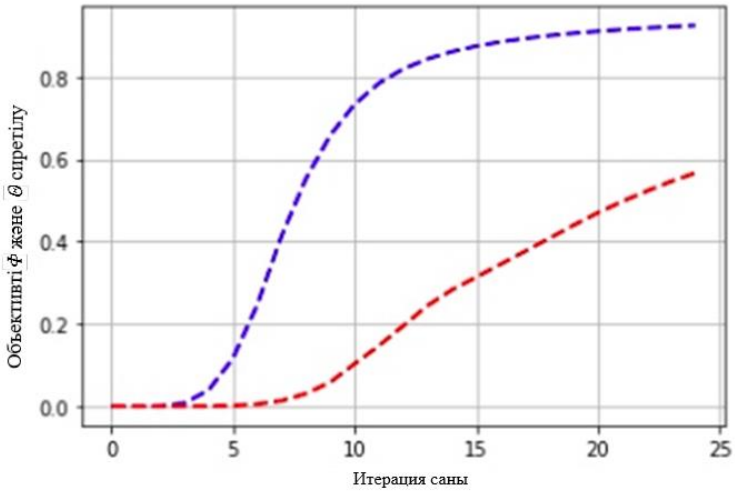


3.3 сурет- Перплексия көрсеткіші

Ірі жинақтардағы тәжірибелер оқытуда және басқада қиындықтардың арасындағы айырмашылық тым аз болатынын көрсетеді.

Үлкен топтамалардың тақырыптық модельдерін құру кезінде көптеген тақырыптар, матрицалардың ығысуы жад мен уақыт шығындарын азайтуға көмектеседі.

Φ және Θ таралуы неғұрлым сирек болса, соғұрлым оның энтропиясы азаяды. 3.4 суретте біз Φ және Θ матрицалардың итеративті процестерінде қалыпты сиретілгенін көреміз. Жалпы Φ матрицасы 92%, Θ матрицасы 56% құрады.



3.4 сурет- Φ және Θ сиретілу көрсеткіштері

Әртүрлі модельдерді салыстыру үшін сынақ модельдері әдетте маңызды емес. Сондықтан өте үлкен үлгілерде күдік тудыратын hold-out perplexity қарастырылмады, бірақ қарапайым деректер негізгі деректерге сәйкес қарастырылды.

Перплексия тақырыптардың түсініктілігі туралы ештеңе айтпайды, тек матрицаның жіктелуі қаншалықты жақсы жасалғандығы туралы айтады.

Яғни, құрастырылған модель соңғы қосымшалар үшін қаншалықты пайдалы болатыны туралы ештеңе айтпайды. Сондықтан тақырыптар қаншалықты жақсы және түсінікті болатынын өлшейтін сапа шаралары жасалды. Мұндай бағалауды мамандардың көмегімен ғана жасауға болады.

Φ және Θ көрсеткіштері толығымен сирек емес, тақырыптар нашар интерпретацияланады, регуляризаторларды қолдану арқылы модельді жақсартуға болады.

3.8 Тақырыптарды сирету және интерпретациясын жақсарту үшін регуляризаторлар комбинациясын қолдану

Регуляризаторлардың үйлесімі модельге олардың әсер ету күшін басқаратын коэффициенттерді таңдау үшін тәжірибелерді қажет етеді. Қалыптастыру стратегиясы регуляризаторлардың жиынтығымен, оларды енгізу дәйектілігімен және итерация кезінде коэффициенттерді өзгерту ережелерімен анықталады [33].

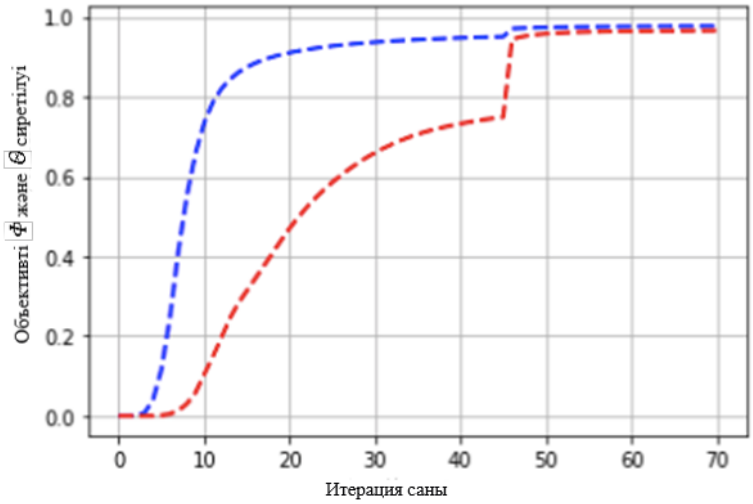
Әзірге стратегияны тәжірибе мен мысалдар мен мақалалардағы ұсыныстарға назар аудара отырып, қолмен таңдауды қажет етеді. Көбінесе регуляризаторлар тәжірибе кезінде кезекпен қосылып, әрқайсысы үшін регуляризатор коэффициенті сұрыпталды.

Φ және Θ матрицалардың құрылымын сиретуін қамтамасыз ету үшін үш регуляризатордың қосындысы ұсынылады: Φ және Θ матрицаларын сирету және тегістеу регуляризаторлары, Φ матрицасындағы тақырыптарды декорреляциялау регуляризаторы.

Тегістейтін және сирету регуляризаторлары бірдей класс әдісімен artm.SmoothSparsePhiRegularizer анықталады: егер коэффициент оң болса, онда регуляризатор тегіс болады, егер теріс болса - сиретіледі.

3.3 кесте - Модель көрсеткіштерін салыстыру

Sm	Sp	Dc	perp	Φ	Θ	con	pur
-	-	-	403	0.92	0.56	0.77	0.85
+	+	-	435	0.96	0.77	0.87	0.94
+	-	+	393	0.96	0.78	0.84	0.87
+	+	+	391	0.96	0.81	0.92	0.95



3.5 сурет - Φ және Θ сиретілу көрсеткіштері

Негізгі қорытынды: Матрицалардың сирек болуына қол жеткізгенде (Φ және Θ матрицаларын сирету 98%), тақырыптардың интерпретациялауы нашарлады, тақырыптарды түсіну жағынан ең жақсы 4-модель таңдалды. Сонымен бірге, регуляризаторлар үйлесуі перплексияның болмашы нашарлауымен барлық сапалық өлшемдерді жақсартуға мүмкіндік береді. Сирету Φ матрица элементтерінің 96% және Θ матрица элементтерінің 81% нөлдейді. Декорреляция регуляризаторы тақырыптардың тазалығы мен когеренттілігін арттырды.

Сирету регуляризаторларын тақырыптар жүйелік коэффициенттерді итеративті процесс жинақтала бастағаннан және нөлге жақын Φ және Θ матрицаларының элементтері анықталғаннан кейін қосу ұсынылды. Ертерек немесе неғұрлым тез сирету перплексияны нашарлатады.

Сирету регуляризаторының коэффициенті ретінде tau=-2e6 , тегістеу регуляризаторының коэффициенті tau=1e5 таңдалды. Декорреляция бірнеше рет тақырыптардың тазалығын және келісімділігін арттырады, бірақ матрицаны әлсіретеді және матрицаны мүлдем сиретпейді. Декорреляцияның

сұйылтумен үйлесуі тазалық пен келісімділікті төмендетпестен сирек кездесетін сирек кездесуге мүмкіндік береді.

Тақырыптардың декорреляциясы 15-ші итерациядан басталды, регуляризатор коэффициенті тұрақты және жоғары болды, бұл кезде перплексияның айтарлықтай жоғарылауы байқалмады, осы жинақ үшін γ = 750000 мәні таңдалды.

3.9 Тақырыптарды интерпретациялау

Тақырыптық модельдеудің негізгі нәтижелері екі матрицада орналасқан:

- «сөз - тақырып» матрицасында бөлу бар әр тақырып бойынша сөз ықтималдығы; олар тақырыптарды түсіндіру және оларды пайдаланушыларға көрсету үшін қажет;
- «құжат-тақырып» таратуды қамтиды, әр құжат үшін тақырыптардың ықтималдығы; олар құжаттарды іздеу, жіктеу, визуализация үшін құжаттардың векторлық көрінісі ретінде қолданылады.

Сонымен қатар, модельдеудің жанама нәтижелері бар:

- әр құжаттағы әр сөз үшін ықтималдықты бөлу; олар құжаттың тақырыптық құрылымын талдауға және құжаттар ішіндегі ақпаратты іздеуге қызмет етеді;
- әр итерация кезінде есептелген модель сапасының көрсеткіштері; олар итерациялық процесті бақылау үшін қолданылады.

Ең жақсы тақырыптардағы мысалдары TopTokensScore() көрсеткіші арқылы алынған:

- apple users: g.aaplimg.com, push.apple.com, www.apple.com, itunes.apple.com, apple.com, www.icloud.com, ls.apple.com, time-ios.apple.com, gateway.icloud.com, xp.apple.com, gs-loc.apple.com, guzzoni.apple.com, iphonesubmissions.apple.com, mesu.apple.com, imap.mail.ru, imap.gmail.com, captive.apple.com, iphone-ld.apple.com, smoot.apple.com, query.yahoo.com, news.apple-dns.net, g03.yahoodns.net;
- kaspersky users: ksn-stat-geo.kaspersky-labs.com, ksn-pp.kaspersky-labs.com, ksn-url-geo.kaspersky-labs.com, ksn-file-geo.kaspersky-labs.com,ksn-cinfo-geo.kaslabs.com, ksn-crypto-info-geo.kas-labs.com.com,saping.igamecj.com, naping.igamecj.com, ds.kaspersky.com, hkping.igamecj.com, euping.igamecj.com, ksn-crypto-url-geo.kaspersky-labs.com, ksn-verdict-geo.kaspersky-labs.com, ksn-kas-geo.kaspersky-labs.com,ksn-fr-geo.kaspersky-labs.com, meping.igamecj.com, tplay.qq.com, krping.igamecj.com, ksn-crypto-info-geo.kaspersky-labs.com;
- vkontakte users: query.yahooapis.com, sun9-25.userapi.com, sun9-27.userapi.com, sun9-23.userapi.com, sun9-38.userapi.com, sun9-5.userapi.com, sun9-29.userapi.com, sun9-4.userapi.com, sun9-30.userapi.com, sun9-26.userapi.com, sun9-32.userapi.com, sun9-39.userapi.com, sun9-28.userapi.com, sun9-41.userapi.com, sun9-21.userapi.com, sun9-2.userapi.com, sun9-8.userapi.com, sun9-9.userapi.com, sun9-12.userapi.com, sun9-10.userapi.com;
- instagram users: fna.fbcdn.net, graph.instagram.com, i.instagram.com, edge-mqtt.facebook.com, scontent.cdninstagram.com, baidu.co.th, scontent-waw1-

1.cdninstagram.com, s.cyingv.com, scontent-arn2-2.cdninstagram.com, scontent-arn2-1.cdninstagram.com,scontent-lga31.cdninstagram.com,www.testserver4.com, scontent-frt3-1.cdninstagram.com, scontent-lax3-1.cdninstagram.com, scontent-yyz1-1.cdninstagram.com, scontent-vie1-1.cdninstagram.com, supl.google.com, scontent-lga3-1.cdninstagram.com;

- facebook users: graph.facebook.com,mqttmini.facebook.com, api.vk.com g.whatsapp.net,connectivitycheck.gstatic.com,clients.google.com,bapi.facebook.com,bgraph.facebook.com,play.googleapis.com,clients3.google.com,lh3.googleusercontent.com,android.googleapis.com,api.facebook.com,appmetrica.yandex.net,fna.whatsapp.net,static.whatsapp.net,www.googleapis.com,graph.instagram.com,settings.crashlytics.com,edge-mqtt.facebook.com.

Әрі қарай матрицасын pandas.DataFrame түрінде model.get_phi () әдісі арқылы авторлардың тақырыптар бойынша үлестірімін аламыз (авторлардың модальділігіне сәйкес келетін екі матрицаның бірі). Жол бойында біздің жинақта кездесетін сөздер бар. Мәндер - бұл олардың ықтималдығы.

	topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	...	topic20	topic21	topic22	topic23	topic24	topic25
(subscriber, 92.47.154.230)	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.00000	0.000000e+00	0.0	0.0
(subscriber, 92.47.154.23)	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.00000	0.000000e+00	0.0	0.0
(subscriber, 92.47.154.229)	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.00000	0.000000e+00	0.0	0.0
(subscriber, 92.47.154.228)	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.00000	0.000000e+00	0.0	0.0
(subscriber, 92.47.154.227)	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.00000	0.000000e+00	0.0	0.0
(subscriber, 92.47.154.226)	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.00008	0.000000e+00	0.0	0.0
(subscriber, 92.47.154.221)	0.0	0.0	0.0	0.0	0.0	0.000033	0.0	0.0	0.0	0.0	...	0.0	0.0	0.00000	0.000000e+00	0.0	0.0
(subscriber, 92.47.154.218)	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.00000	6.619193e-11	0.0	0.0
(subscriber, 92.47.154.216)	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.00000	0.000000e+00	0.0	0.0
(subscriber, 92.47.154.215)	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.00000	0.000000e+00	0.0	0.0

3.6 сурет - Авторлардың тақырыптар бойынша үлестірімі

Сонымен қатар, матрицасын pandas.DataFrame түрінде model.get_theta() әдісі арқылы шығарып аламыз. Θ матрицалары - бұл құжаттарда тақырыптардың қалай үлестірілгенін бағалаудың ең оңай тәсілі. Құжаттардың тақырыптық профильдері:

	95.57.44.77	95.57.44.78	95.57.44.8	95.57.44.81	95.57.44.82	95.57.44.83	95.57.44.84	95.57.44.85	95.57.44.89
topic0	0.00000	7.692306e-02	0.0	0.000000	0.0000	0.000000	0.000000	0.000000e+00	0.00
topic1	0.00000	0.000000e+00	0.0	0.000000	0.0000	0.000000	0.116976	3.537577e-07	0.08
topic2	0.53125	0.000000e+00	0.0	0.086402	0.1875	0.000000	0.046667	3.474145e-01	0.00
topic3	0.00000	0.000000e+00	0.0	0.000000	0.0000	0.000000	0.000000	0.000000e+00	0.00
topic4	0.00000	0.000000e+00	0.0	0.000000	0.0000	0.000000	0.000000	0.000000e+00	0.00
topic5	0.00000	1.321409e-15	0.0	0.000000	0.0000	0.181818	0.000000	0.000000e+00	0.00
topic6	0.00000	7.897697e-02	0.0	0.186249	0.0000	0.000000	0.000000	0.000000e+00	0.00
topic7	0.00000	0.000000e+00	0.0	0.000000	0.0000	0.000000	0.000000	8.163264e-02	0.00
topic8	0.00000	0.000000e+00	0.0	0.000000	0.0000	0.000000	0.000000	0.000000e+00	0.00
topic9	0.00000	0.000000e+00	0.0	0.000000	0.0000	0.000000	0.040975	0.000000e+00	0.00

3.7 сурет - Тақырыптардың құжаттар бойынша үлестірімі

Авторлардың барлық жинағында әр тақырыптың үлесі $p(t)p(t)$ - анықтай аламыз. Толық ықтималдық формуласына сәйкес бұл мәндер келесідей есептелуі керек $p(t) = \sum dp(t|d)p(d)p(t) = \sum dp(t|d)p(d)$. Ықтималдық модель бойынша, $p(d)p(d)$ құжаттың d ұзындығына пропорциянал:

```
print('Topic popularity:\n', subcribers_phi.astype(bool).sum(axis=0).sort_values(ascending=False)
      /subcribers_phi.astype(bool).sum(axis=0).sum())

Topic popularity:
 topic10    0.082399
topic14    0.065782
topic20    0.061596
topic16    0.060465
topic17    0.059170
topic19    0.054124
topic2     0.050530
topic5     0.046392
topic23    0.043272
topic12    0.042091
topic26    0.035311
topic7     0.034701
topic0     0.033158
topic1     0.032731
topic21    0.030160
topic27    0.029958
topic24    0.027742
topic6     0.026100
topic15    0.025152
topic9     0.022482
topic18    0.022384
topic4     0.018857
topic11    0.017966
topic28    0.017041
topic22    0.016603
topic13    0.015114
topic3     0.013384
topic29    0.006049
topic25    0.004693
topic8     0.004590
dtype: float64
```

3.8 сурет - Авторлардың тақырыптар бойынша үлесі

Авторлардың көпшілігі бірнеше тақырып бойынша маңызды, бұл қисынды. Кез-келген автор үшін әр тақырыптың ықтималдығы:

```
print(theta.loc[:, '178.89.61.113'].sort_values(ascending=False))
```

```
topic23    0.614577
topic19    0.274312
topic8     0.111111
topic29    0.000000
topic13    0.000000
topic1     0.000000
topic2     0.000000
topic3     0.000000
topic4     0.000000
topic5     0.000000
topic6     0.000000
topic7     0.000000
topic9     0.000000
topic10    0.000000
topic11    0.000000
topic12    0.000000
topic14    0.000000
topic28    0.000000
topic15    0.000000
topic16    0.000000
topic17    0.000000
topic18    0.000000
topic20    0.000000
topic21    0.000000
topic22    0.000000
topic24    0.000000
topic25    0.000000
topic26    0.000000
topic27    0.000000
topic0     0.000000
Name: 178.89.61.113, dtype: float32
```

3.9 сурет - Тақырыптар үлесі

Сонымен, модельді үйреткен кезде біз екі матрицаны аламыз - Φ және Θ. Φ матрицасы - бұл тақырыптардағы сөздерді тарату матрицасы, ал Θ матрицасы - құжаттардағы тақырыптарды тарату матрицасы. Біріншісі біздің тақырыптарымыз туралы, оған қандай сөздер кіретінін, ал екіншісі - қандай құжаттар туралы, біздің құжаттарымызға қандай тақырыптар кіретінін түсінуге мүмкіндік береді.

$p\ (t)$ - әр тақырыптың авторлардың модальділігіндегі үлесін бағалай аламыз. Ең белсенді пайдаланушыларды аламыз:

```
#most active iphone user:
top_iphone_users = subcribers_phi.loc[subcribers_phi.topic19 > 0, 'topic19'].sort_values(ascending=False).head()
print('top_iphone_users: ', top_iphone_users)
```

```
top_iphone_users:  (subscriber, 95.58.171.227)    0.00003
(subscriber, 2.135.252.196)    0.00003
(subscriber, 2.133.13.40)      0.00003
(subscriber, 213.211.73.31)    0.00003
(subscriber, 37.151.1.74)      0.00003
Name: topic19, dtype: float32
```

3.9 сурет - Белсенді авторлардың тақырыптар бойынша үлесі

4 Өміртіршілік қауіпсіздігі

4.1 Еңбек жағдайларын талдау

Тіршілік қауіпсіздігі бөлімінің негізгі міндеті - өндірістік жарақаттанудан қорғау және алдын-алу, кәсіптік аурулар және әлеуметтік салдарды азайту. Басқаша айтқанда, еңбекті қорғаудың негізгі міндеті - әрбір жұмыс орнында әлеуметтік қолайлы жағдай жасау [34].

Телекоммуникациялардағы инженерлердің негізгі жұмысы ұйым үшін үлкен мүмкіндіктер ашатын, дұрыс мәліметтер жиынтығы мен айнымалыларды анықтайтын, деректерді талдаумен тікелей байланысты. Әртүрлі көздерден құрылымды және құрылымды емес мәліметтердің үлкен жиынтығын жинау мен дәлдігін, толықтығын және біркелкілігін қамтамасыз ету үшін деректерді тазалау және тексеру, ірі ақпарат өндірудің модельдері мен алгоритмдерін модельдеу және қолдану. Бөлмеде мынандай жабдықтар бар: дербес компьютерлер, дерекқор серверлері, модемдер және Wi-Fi маршрутизаторлары, кабельдер мен сымдар, кіріс-шығыс жабдықтары (сканер, принтер). Дербес компьютерде жұмыс жасау кезінде қауіпті және зиянды өндірістік факторлар: электр өрісі кернеуінің жоғарылауы, электр тогының соғуы, электр жабдықтарының істен шығуы, табиғи жарықтың болмауы немесе жетіспеуі, жұмыс аймағының ауа температурасының жоғарылауы немесе төмендеуі.

Электр тогының соғуынан қорғау шаралары [35], ең алдымен, электр қондырғыларының қауіпсіздігі келесі қорғаныс шараларымен қамтамасыз етіледі:

- сенімді оқшаулау;
- тірі бөлшектердің болмауы;
- нөлдеу;
- жерге қосу;
- потенциалдарды теңестіру;
- автоматты өшіру;
- ескерту дабылдары, жазулар мен плакаттар.

Жабдық келесі жағдайларда оңтайлы жұмыс істейді:

- бөлмедегі температура көрсеткіші 0-ден 40° - қа дейін;
- бөлмедегі ылғалдылық 95%-ке дейінгі мөлшерде, конденсацияланбаған;
- қуат: ауыспалы ток - кернеуі 100-ден 220 В-қа дейін, жиілігі 50 / 60Гц, ток 2 - 5 А; тұрақты ток - 48 - 60 В, жүктеме тогы 2 - 4 А.

Барлық жабдықтар сертификатталғандықтан, кәсіби тәуекел дәрежесі ең төменгі ретінде анықталған. Қауіпсіздік шаралары бойынша құрылғылар жұмыс кернеуі 1 кВ дейінгі құрылғыларға жатады.

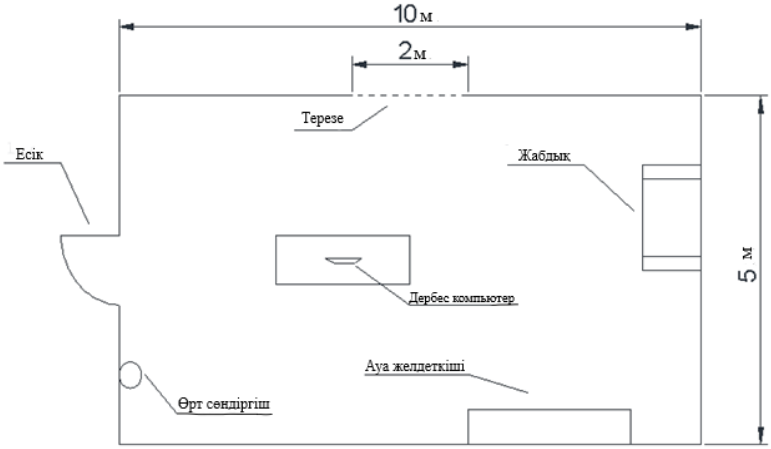
Электр тогының соғу қаупі дәрежесіне сәйкес бөлме қауіпті дәрежеге ие емес, себебі ол келесі талаптарға жауап береді:

- құрғақ;

- қалыпты температурамен;
- едендермен оқшауланған;
- шаңсыз.

Қоршаған ортаның табиғатына сәйкес бөлме «қалыпты құрғақ» класына жатады, салыстырмалы ылғалдылық 60%-дан аспайды. Қол жетімділік тұрғысынан ол электр санатына жатады, яғни жабдыққа қол жеткізуді тек электр мамандары жүзеге асырады. Электр қондырғыларына техникалық қызмет көрсетуді кезекші персонал жүзеге асырады, олардың білікті тобы кемінде III топ болады.

Біздің кеңседе ОП-4 ұнтақты өрт сөндіргішін қолданамыз. Габариттік өлшемдері: биіктігі - 205 мм, диаметрі - 155 мм.



4.1 сурет – Жұмыс бөлмесін жоспарлау

4.1.1 Өрт қауіпсіздігі. Есептеу жұмысы берілген әдістемелік нұсқаулық бойынша жасалды. Өрт үлкен материалдық шығын әкеледі және кейбір жағдайларда адам өмірін жоғалтуға әкеледі. Өрттен қорғау - қоғамның әр мүшесінің маңызды міндеті және ұлттық деңгейде жүзеге асырылады.

Өрттен қорғау өрттің алдын-алудың мақсаты экономикалық тұрғыдан тиімді және техникалық жолдары мен құралдарын табуға және оларды ең аз шығынмен жоюға күштер мен техникалық құралдарды ұтымды пайдалана отырып жоюға бағытталған.

Өрт қауіпсіздігі - өрт қаупі жоқ объектінің жай-күйі, және ол туындаған кезде қауіпті өрт факторларының адамдарға, құрылыстар мен материалдық құндылықтарға жағымсыз әсерін жою үшін қажетті шаралар [36].

Өрт қауіпсіздігі өрттің алдын алу және өрттен белсенді қорғау шараларымен қамтамасыз етіледі. Өрттің алдын алу өрттің алдын алуға немесе оның салдарын азайтуға бағытталған іс-шаралар кешенін қамтиды. Өрттің белсенді қорғанысы - өртке немесе жарылыс жағдайларына сәтті күресуді қамтамасыз ететін шаралар.

Технологиялық процестің ерекшелігі, қолданылатын заттар мен материалдардың қасиеттері, сондай-ақ электронды жабдықтардың,

қондырғылардың, серверлердің, дербес компьютерлер өрт қауіптіліктің Д категориясына жатқызылады.

Көп мөлшерде жылу, ыстық газдар мен булардың шығуы қоршаған ортаға жоғары қысым жасайды және ғимаратқа, жабдықтарға және адамдарға қауіп төндіреді. Осы жылу нәтижесінде пайда болатын жылу қоспаның бөлшектерін тұтатады, жану процесі өте жылдам болады.

Өрттің себептері электрлік және электрлік емес әсерлер болып табылады. Электрлік әсерлердің себептері: электр аппараттарында, машиналарда, электростатикалық разрядтар мен найзағай соққыларында ұшқын, қысқа тұйықталу токтары және электр құрылғыларының сымдары мен орамаларының шамадан тыс жүктелуі, олардың жоғары температураға дейін қызуы, сымдардың қосылу метаметіндегі жағымсыз түйісулер, жылудың көп мөлшері шығарылатын өтпелі қарсылықтың жоғарылауына әкеледі [37].

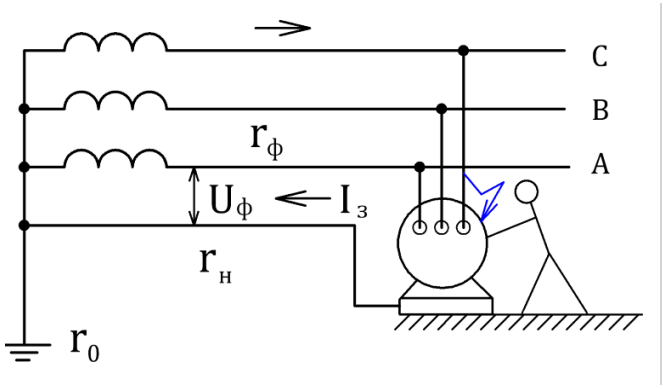
Электрлік емес өрттің себебі газбен дәнекерлеу жабдықтары мен соққылардың дұрыс пайдаланылмауы, сондай-ақ кабельдік массалар мен сіңдіргіш қосылыстардың дұрыс қызбауы, жылыту құрылғыларының дұрыс жұмыс істемеуі және олардың жұмыс режимін бұзу, өндірістік жабдықтың дұрыс жұмыс істемеуі және процестің бұзылуы, жанғыш газдардың, булардың немесе шаңның қоршаған ортаға шығуы, өрт қауіпті жерлерде темекі шегу, кейбір материалдарды өздігінен тұтату.

4.1.2 Электр қауіпсіздігі. Бүгінгі таңда біздің өмірімізді электр құрылғыларының барлық түрлерін күнделікті пайдаланбай елестету қиын. Алайда, токты қорғаныс жүйесінсіз іс жүзінде пайдалану қауіпсіз емес. Қорғаныс құрылғылары (штепсельдер, ажыратқыштар және т.б.) жұмыс істемеуі мүмкін, нәтижесінде ішкі оқшаулау бұзылып, жабдықтың металл корпусында жоғары кернеу пайда болады.

Электр және тұрмыстық құрылғыларды пайдалану кезінде адамды электр тогының соғуынан қорғау үшін әр түрлі қорғаныс шаралары, соның ішінде нөлге айналдыру шаралары әзірленді [38].

Нөлдеу электр қауіпсіздігі жүйелерін PEN, PE немесе N өткізгіштерімен қамтамасыз ету үшін қолданылады. Олардың құрамына төмен жерге тұйықталған бейтарап желілер жатады: TN-C, TN-S және TN-C-S. Бұл жүйелер үшін нөлденуді ұйымдастырудағы басты айырмашылық нөлдік қорғаныс және жұмыс өткізгіштерін қосу схемасы болып табылады.

Электрлік нөлдеу - бұл электр қауіпсіздігін қамтамасыз етуге арналған трансформатордың немесе генератордың жерге тұйықталған бейтарапқа өткізгіш бөлшектерін әдейі қосу. Нөлдеу - электр желілеріндегі электр тізбегін қорғаудың негізгі шарасы, электр желілерінде кернеуі 1000 В дейін, төмен жерге тұйықталған бейтарап әдіс.



4.2 сурет - Электрлік нөлдеу мысалы

4.1.3 Микроклимат параметрлерін нормалау. Кеңсе бөлмесінде 4 адам жұмыс істейді, көлемі 10х5х3 метр, көлемі 150 м^3. Бөлме ішіне келесі ауа көлемі енгізіледі: бөлменің текше метрі үшін бір жұмысшыға 30 м^3 дейін, бір адамға 20 м^3 / сағ. Кеңсе кеңістігіне кіретін ауа шаң мен микроорганизмдерден тұратын ластаушы заттардан тазартылады.

Өндірістегі микроклимат жұмыс аймағында, яғни жұмысшылардың тұрақты немесе уақытша болу орындары орналасқан еденнен немесе платформадан 2 м биіктіктегі кеңістікте бағаланады. Ылғалдылық ондағы су буының құрамымен анықталады.

Сандық электр құрылғы жабдықтарының қалыпты жұмысына сәйкес микроклиматтық жағдайларды сақтау үшін оператор бөлмесінде кондиционер орнатылған. Микроклиматтың стандартты көрсеткіштері 4.1 кестеде келтірілген.

4.1 кесте - Қалыпты жұмысты орындау кезінде өндірістік үй-жайлардың микроклиматы

Жыл мезгілі	Температура, С		Оңтайлы ылғалдылық,%		Ауа жылдамдығы, м / с	
	Оңтайлы	Рұқсат етілген	Оңтайлы	Рұқсат етілген	Оңтайлы	Рұқсат етілген
Суық мезгіл	18-20	17-23	40-60	75	0,2	0,1-ден аспайды
Жылы мезгіл	21-23	18-27	40-60	26 -65	0,3	0,2-0,4

4.2 Есептеу бөлімі

4.2.1 Ұнтақты өрт сөндіру қондырғыларын есептеу. Есептеу жұмысы берілген әдістемелік нұсқаулық [38] бойынша жасалды. Өрт шыққан кезде өрт сөндіру бөлімін тез арада шақыруға арналған қондырғылар бар. От, телефон және радио байланысы туралы хабарлау үшін сирена қолданылады. Барлық өрт сөндіру техникалары тәуліктің кез келген уақытында қол жетімді. Бөлменің биіктігі 4 м, бір хабарлағыш бақылайтын аумақ 50 м^2 құрайды. Формула бойынша Скиф-Д4 санын анықтаймыз:

$$M = Ц \cdot (S/S0), \quad (4.1)$$

мұндағы, Ц – бүтін санға дейін дөңгелектеу;

S – бөлменің ауданы;

S0 – бір ОП-4 басқаратын аудан;

$M = Ц\,(3 \cdot 2{,}5/50) = 0{,}15 = 1$ (аппараттық);

$M = Ц \cdot (18 \cdot 12/50) = 4{,}32 = 5$ (операторлық).

Біз 5 өрт хабарлағышын ғимаратқа орналастырамыз. Біздің кеңседе ОП-4 ұнтақты өрт сөндіргішін қолданамыз.

Габариттік өлшемдері: биіктігі - 205 мм, диаметрі - 155 мм.

4.2 кесте - ОП-4 өрт сөндіргішінің сипаттамасы

Атауы	Өрт сөндіргіштер мөлшерінің нормативтері
Сөндіргіш заттың массасы, кг	7
Ұнтақ ағынының ұзындығы, м; аз емес	6
Өрт сөндіргішті іске қосу уақыты, с; бұдан артық емес	5
Ұнтақтың шығу уақыты, с; аз емес	10
Сөндіретін ұнтақтың қалдығы,%; бұдан артық емес	9
Қолдануға болатын қоршаған орта температурасы, С	-30
Қолдануға болатын қоршаған орта температурасы, С	+50
Жалпы өлшемдер:	
Диаметрі, мм	155
Биіктігі, мм	205
Зарядталған өрт сөндіргіштің салмағы, кг	12,5
Өрт сөндіру ауданы В, м^2; аз емес	4
Жұмыс қысымы, МПа	1,2

ОП типіндегі бірыңғай ұнтақты өрт сөндіргіштер А класындағы (қатты заттар), В класындағы (сұйық заттар), С класындағы (газ тәрізді заттар) және 1000 В дейінгі электр қондырғыларындағы өртті сөндіруге арналған, өрт сөндіргішті қосу ережелері көрсетілген. Барлық өрт сөндірушілер тексеріліп, қайта зарядталады.

Көлемді өртті сөндіруге арналған ұнтақ құрамының есептік массасы m_d, кг, формула бойынша анықталады, құрамы ескерілмеген:

$$m_d = k \cdot g_n \cdot V, \quad (4.2)$$

мұндағы, k = l,2 - шығындардың орнын толтыру коэффициенті;

gn = 0,4 – құрамның стандартты массалық концентрациясы.

V – бөлменің көлемі:

$$V = A \cdot B \cdot H, \quad (4.3)$$

мұндағы, А = 10 м – бөлменің ұзындығы;

В = 5 м – бөлменің ені;

Н = 4 м – бөлменің биіктігі.

Содан кейін:

$$V = 10 \cdot 5 \cdot 4 = 200 \text{ м}^3.$$

Сондықтан:

$$m_d = 1{,}2 \cdot 0{,}4 \cdot 200 = 96 \text{ кг}$$

Тұрақты ашық саңылаулар болған кезде, ғимарат конверті ауданының 1% -дан 10% -на дейін, 1 м^2 ұнтақтың құрамына 5 кг қосымша тұтынады, бұл 1 м^2 саңылауға 5 кг құрайды. (96 + 5 = 101 кг).

Цилиндрлердің есептік саны ξ 20 литрлік цилиндрдегі сыйымдылығы 12,5 кг ұнтақ құрамынан есептеледі.

Магистральдық құбырдың ішкі диаметрі d_i, мм, формула бойынша анықталады:

$$d_i = 12 \cdot \sqrt{2} = 16 \text{ мм} \quad (4.4)$$

Магистральдық құбырдың ұзындығы l_2, м, формула бойынша анықталады, жергілікті шығындарды есепке алмағаны үшін өтемақыны төлеу үшін:

$$l_2 = k_2 \cdot l_1 \quad (4.5)$$

мұндағы, k_1 = 1,2 – құбырдың ұзындығының өсу коэффициенті;

l = 4 м – жобаға сәйкес құбырдың ұзындығы;

l_2 = 1,2 · 4 = 4,8 м.

A_3, мм^2 суарғыштың шығуының көлденең қимасы мына формула бойынша анықталады:

$$V = S / x_1, \quad (4.6)$$

мұндағы, S – магистральдық құбырдың қималық ауданы, мм^2;

x_1– шашыратқыштардың саны.

Содан кейін:

$$A_3 = \frac{3{,}14 \cdot 8{,}5^2}{1} = 226{,}9 \text{ мм}^2$$

Ұнтақтың Q, кг/с, құбырдың тең ұзындығына және диаметріне байланысты, 1,5 кг/с құрайды.

Ұнтақты құрамның болжамды берілу уақыты t, мин, формула бойынша анықталады:

$$t = \frac{m_d}{60Q}, \quad (4.7)$$

$$t = \frac{101}{60 \cdot 1{,}5} = 1{,}12 \text{ мин}$$

Ұнтақ құрамының негізгі құрамының массасы m, кг, формула бойынша анықталады:

$$m = 1{,}1 \cdot m_d \cdot \left(1 + \frac{k_2}{k_1}\right), \quad (4.8)$$

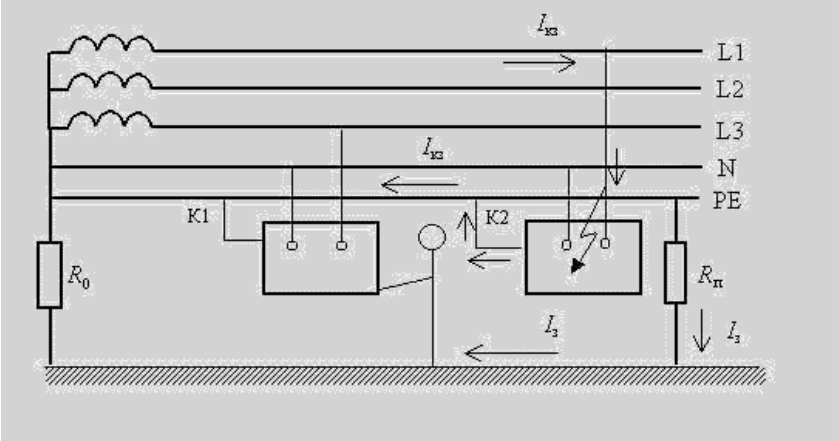
мұндағы, k_2 = 0,2 – баллондар мен құбырлардағы ұнтақ құрамының тепе-теңдігін ескеретін коэффициент.

Онда:

$$m = 1{,}1 \cdot 101 \cdot \left(1 + \frac{0{,}2}{1{,}2}\right) = 129{,}62 \text{ кг.}$$

4.2.2 Нөлдеу. Есептеу жұмысы берілген әдістемелік нұсқаулық [39] бойынша жасалды.Тәжірибелік мысал: екі нөлдендірілген электр көзі бір TN-S желісінен қуат алынады, кернеуі 380/220 V. Фазалық сым екінші тұтынушының ісіне қысқартылды, ал қазір адам бұл мәселені қозғауда электр энергиясының алғашқы тұтынушысы (4.3 сурет).

Электр энергиясын бірінші тұтынушының денесіне тиген адамның ағынын анықтаймыз, егер $R_{L1} = R_{PE(2)} = 0{,}23$ Ом, $R_0 = R_П = 4$ Ом, $l_{K1} = 0{,}3 l_{K2}$:



4.3 сурет – Тәжірибелік мысал

ПЭ өткізгіш секцияларының кедергісін бірінші және екінші ғимараттарға мына формулалар арқылы анықтауға болады:

$$R_{PE(1)} = \rho \frac{l_{k1}}{S_{PE}}, \quad (4.9)$$

$$R_{PE(2)} = \rho \frac{l_{k2}}{S_{PE}}, \quad (4.10)$$

$l_{K1} = 0{,}3 l_{K2}$, $R_{PE(1)} = 0{,}3 R_{PE(2)}$. Екінші тұтынушы жағдайында фазалық сым жабылған кезде қысқа тұйықталу тогы пайда болады, оны келесідей есептеуге болады:

$$I_{кт} = \frac{l_L}{R_{L1} + R_{PE(2)}} = \frac{220}{0{,}4} = 550 \text{ м}A;$$

Жерге тұйықтау кезінде пайда болатын токтың шамасы келесідей анықталады:

$$I_ж = \frac{I_{SF} \cdot R_{PE(2)}}{R_0 + R_C} = \frac{110}{6} = 18{,}3 \text{ м}A;$$

Жерге қатысты нөлдік нүктенің кернеуінің мәні:

$$U_0 = R_0 \cdot I_ж = 18{,}3 \cdot 3 = 55 \text{ Вт};$$

Бірінші денеге тиген адам үшін жанасудың кернеу мәні осы дененің жерге қатысты кернеу мәніне тең болады, оны өрнектен анықтауға болады:

$$U_{C1} = U_0 - I_S \cdot R_{PE(2)} = 55 - 550 \cdot 0{,}2 \cdot 0{,}3 = 22\text{Вт};$$

Нәтижесінде бірінші денеге тиген адамның денесі арқылы өтетін токтың қалаған мәні болады:

$$I_{\text{а}} = \frac{U_{C1}}{R_h} = \frac{22}{1} = 22\ \text{м}A.$$

4.2 Өміртіршілік қауіпсіздігі бөліміне қорытынды

Тіршілік қауіпсіздігі бөлімінде кәсіпорындағы еңбек жағдайына талдау жасалынды. Кеңсе қызметкерлері еңбек барысында ұшырайтын негізгі қауіпті және зиянды факторлар анықталды. Бұл жұмыста өрт қауіпсіздігі, электр қауіпсіздігін қамтамасыз ету үшін нөлдеу есептелді.

Өрт қауіпсіздігін қамтамасыз ету үшін 5 Скиф-4 өрт хабарлағыштары қолданылады, сонымен қатар ОП- 4 типіндегі өрт сөндіргіштер орнатылады.

Бөлмедегі жабдықпен жұмыс кезінде кез-келген қысқа тұйықталуды болдырмас үшін электрлік нөлдеу есептелінді. Электрлік нөлдерді есептеу кезінде өрт қауіпсіздігі ережелерін құрулып және ОП-4 ұнтақты өрт сөндіргішті кеңседе қолдануға шешім қабылданды. Жалпы өлшемдері: биіктігі - 205 мм, диаметрі - 155 мм. Электрлік құрамдас бөліктермен байланыстырылған кедергі мәні 0,1 Ом-дан аспайды.

5 Тақырыптық модельді пайдаланудағы техникалық – экономикалық негіздеме

5.1 Диплом жобасын әзірлеудің орындылығын негіздеу

Әзірленген бағдарламалық қамтамасыздандыру компанияларда әр түрлі қызмет салаларында орын тауып, сәтті қолданылады. Бұл әр түрлі көздерден алынған деректерді өңдеу мен сақтауды оңтайландыруға мүмкіндік береді. Телекоммуникацияда деректерді талдаушы өте кеңінен қолданылады.

Бұл тарауда ақпараттық-аналитикалық жүйені имплантациялаудың әсерін желідегі пайдаланушылады тақырыптарға бөлу міндеті ретінде бағалайды. Кіріс деректердің көлемі уақыт өте келе арта түседі және бұл жүйені енгізу телекоммуникация операторларына пайдаланушылар туралы түсініктерін жақсартуға, қосымшаларды кеңейтуге, жаңа пайдаланушыларды тартуға мүмкіндік береді, бұл өз кезегінде бизнесте пайда әкеледі [40].

Радио және теледидарлық станциялардан келетін мәліметтердің көбеюін ескере отырып, бұл аналитикалық жүйені енгізу пайдаланушыларды тартуға жарнама мен мазмұнды жақсартуға үлкен әсер етеді, бұл бизнес кірісіне және клиенттер үшін желімен жақсы оңтайлы әрекеттесуге әкеледі. Бұл бір жағынан желінің қол жетімділігін көрсетеді [41].

Мұндай параметр еліміздің ақпараттық қауіпсіздігінің бөлігі ретінде стратегиялық маңызды рөл атқарады. Телекоммуникация ретінде осы салалардағы қамтуды жақсарту тек байланыс операторлары үшін ғана емес, сонымен бірге Қазақстан Республикасының Ақпарат және Коммуникациялар Министрлігі мен Ұлттық Қауіпсіздік Комитеті үшін де маңызды.

Ақпаратты жедел өңдеудің және телекоммуникация, интернет арқылы халыққа берудің маңыздылығын бағалау мүмкін емес. Сондықтан телекоммуникация желілерінде орын алатын жағдайлар туралы аналитикалық ақпараттық жүйені құру маңызды нәтиже береді.

5.2 Капиталды шығындарын есептеу

Капиталды шығындарды бөлу әр кәсіпорынның қызметін қамтамасыз ететіндігіне қарамастан, оларды екі топқа бөлуге болады:

- негізгі құралдарды сатып алуға;
- негізгі құралдарға қызмет көрсету.

5.1 кесте - Негізгі құрал-жабдықтардың шығындарын есептеу

Құрылғы	Бірлік құрылғының құны, тенге	Саны	Сомасы, тенге
Shell T330 8B LFF Hot-Plug	753900	1	753900
HP ProBook 450 G6 15.6	373500	1	373500

5.2 кестенің жалғасы

Құрылғы	Бірлік құрылғының құны, тенге	Саны	Сомасы, тенге
Қуат көзі	35625	1	35625
Кабель 15 метр	5000	1	5000
Patch cord Ethernet 15 метр	8000	1	8000
Барлығы			1230025

Капиталды шығындарға осы жабдықтың құны, орнату шығындары және басқалар кіреді. Жалпы қаржыландыру формула бойынша табылады:

$$К_с = К_ж + К_{орн} + К_т + К_ш, \quad (5.1)$$

мұндағы, $К_ж$ - жабдықты сатып алу құны;

$К_{орн}$ - орнату құны;

$К_т$ - транзиттік тасымалдау құны;

$К_ш$ – жобалық шығындар.

Жабдықты сатып алу құны:

$$К_ж = 1230025 \text{ (теңге)}$$

Көлік құны жабдық құнының 5%-ын құрайды:

$$К_т = К_ж \cdot 0.05$$

$$К_т = 1230025 \cdot 0.05 = 61501 \text{ (теңге)}$$

Жобаның құрылымы мен дизайны барлық жабдық құнының 3% құрайды және шлем формулаларына негізделеді:

$$К_ш = К_ж \cdot 0.03$$

$$К_ш = 1230025 \cdot 0.03 = 36901 \text{ (теңге)}$$

Айналымдағы жалпы капитал шығындары:

$$К_с = 1230025 + 61501 + 36901 = 1328427 \text{ (теңге)}$$

5.1 сурет – Капиталды салым құрылымы

5.2 кесте - Жобаның күрделі шығындары

Шығындардың атауы	Құны, теңге
Құрылғы	1230025
Көлік (жабдық құнынан 5%)	61501
Жобаның құны (жабдық құнынан 3%)	36901
Күрделі шығындар	893669

5.3 Пайдалану шығындарын есептеу

Кәсіпорынның операциялық шығындары басқа кәсіпорындармен бәсекелестікте маңызды рөл атқарады. Бұл сізге кәсіпорынды тиімді басқаруға және басқару қызметкерлерімен өзара әрекеттесуге мүмкіндік береді. Өнімдерді, жүйелерді жарнамалау және жылжыту қосымша шығындарды талап етеді. Инновациялық жүйелер ең перспективалы, олардың құны да артып келеді.

Операциялық шығындарға тауарларды сатып алу құны кіреді. Оларды сатылған тауарлардың құны деп те атайды. Мұндай шығындар жалпы кірістен алынады. Ол сонымен бірге салықтар мен несиелер бойынша сыйақыларды алып тастайды, бұл кәсіпорынның таза кірісін есептеу үшін қажет. Операциялық шығындарға жалдау, сату, әкімшілік шығындар, маркетинг, жалақы, кеңсе шығындары кіреді. Олар жалпы құнның бір бөлігі. Операциялық шығындар ауыспалы және тіркелген шығындардың қоспасы [42].

Бекітілген шығындар әрдайым өзгермейді, ал ауыспалы құрам үнемі өзгеріп отырады. Бекітілген шығындар - шикізат, жалдау ақысы, жалақы, коммуналдық шығындар. Бұл шығындар тікелей және жанама болып бөлінеді. Тікелей шығындар белгілі бір өнім түрін өндірумен байланысты және тікелей есептеледі (еңбек шығындары, өндірістік цехтардың, материалдардың және басқалардың тозуы).

Жанама шығындарға арнайы есептеулерді қолдана отырып, бағалар құрамына кіретін әр түрлі өнімдерді өндіруден (жабдықты ұстау және пайдалану, жалпы өндіріс шығындары, жалпы шаруашылық шығындар)

шығындар жатады. Жанама (үстеме) үстеме шығыстардың мысалы көмекші материалдар, өндірістік және өндірістік персоналға сыйақы және басқа үстеме шығындар болып табылады. Өндіріске мамандандырудың жоғары деңгейі бар кәсіпорындарда өнімнің жекелеген түрлерін шығаратын жеке цехтар бөлу, механикаландырылған есепке алу, жалпы көлемдегі тікелей шығындардың үлесі артады. Бұл өнімнің жекелеген түрлеріне шығындардың дәлдігін жақсартады, басқарудың экономикалық негіздерін нығайтады.

Жылдық шығындарға жылдық өндірістік шығындар немесе нақты өндіріс шығындары қосылады:

$$E_o = C_a + \mathrm{C}_{\text{әқ}} + \mathrm{C}_{\text{мат}} + \mathrm{C}_{\text{э}} + \mathrm{C}_{\text{амор}} + \mathrm{C}_{\text{ж}} + \mathrm{C}_{\text{ққ}}, \quad (5.2)$$

мұндағы, Са - айналым капиталы;

$C_{\text{әқ}}$- әлеуметтік қамсыздандыру шығындары;

$C_{\text{мат}}$ - материалдар мен қосалқы бөлшектердің құны;

$C_{\text{э}}$ - электр энергиясының құны;

$C_{\text{амор}}$ - амортизациялық шығындар;

$C_{\text{ж}}$ - мүлікке ғимараттарды жалға беру ;

$C_{\text{ққ}}$: қосылған құн.

5.3 кесте - Жұмысшылардың жалақысы

Орындаушы	Жұмыс түрі	Айлық мөлшерлеме, теңге	Жылдық жалақы, теңге
Деректерді зерттеуші	Тақырыптарды таңдау және міндеттерді анықтау.	500 000	6 000 000
Барлығы			6 000 000

Жалақы құны мына формула бойынша анықталады:

$$W = W_{\text{н}} + W_{\text{мин}}, \quad (5.3)$$

$$W_{\text{н}} = W_{\text{к}} \cdot 12, \quad (5.4)$$

мұндағы, W – жалақы;

$W_{\text{к}}$ – қызметкерлердің жалақысы.

Жұмысшылардың жалақысы 5.2- кестеде келтірілген. Жұмысшылардың негізгі жалақысы:

$$W\text{н} = 500000 \cdot 12 = 6000000 \text{ (теңге)}$$

Ең төменгі жалақы айына табыстың 30% құрайды (мысалы, басқа жұмыс және т.б.):

$$W_{\text{мин}} = W_{\text{н}} \cdot 0.3 = 6000000 \cdot 0.3 = 1800000 \text{ (теңге)}$$

Демек,

$$W = 6000000 + 1800000 = 7800000 \text{ (теңге)}$$

Қызметкерлердің жалақысы, әлеуметтік аударымдар, таза зейнетақы шығындары:

$$\text{О}_{\text{әқ}} = N \cdot (W - PF), \quad (5.5)$$

мұндағы, $\text{О}_{\text{әқ}}$ – әлеуметтік қажеттіліктерге қаражат бөлу;

PF – зейнетақы қоры (10%);

N – бекітілген стандарттардан қаражат бөлу (жалақының 11%).

Әлеуметтік салық - бұл кәсіпкерлікпен айналысатын кәсіпорындарға қолданылатын салық салу түрі. Қызметкерлердің жалақысынан (RFP) салық заңнамасының ережелеріне сәйкес белгілі бір шегерімдер жасалады. Қызметкерлерге әлеуметтік салықты жұмыс беруші төлейді.

Заңды тұлғалар жұмыскерлердің жалақысынан HF аударады, ал жеке кәсіпкерлер (жеке кәсіпкерлер), жеке адвокаттар мен сот орындаушылары жұмыс істейтін адамдардың санына салық төлейді [43].

Заңды тұлғалар жұмыскерлердің жалақысынан HF-ті қосады, ал жеке кәсіпкерлер (жеке кәсіпкерлер), жеке адвокаттар мен сот орындаушылары жұмыс істейтін адамдардың санына салық төлейді.

СТ есептеу таңдалған салық режиміне байланысты жасалады. Қазақстанда бірнеше салық режимі бар: жалпы белгіленген тәртіп, арнайы салық режимі, жеңілдетілген декларация.

Мүмкіндігі шектеулі жандар жұмыс істейтін кәсіпорындар, сондай-ақ арнайы салық режимінде жұмыс істейтін шаруа қожалықтары HF төлеуден босатылады.

Әлеуметтік салық:

$$C_{\text{с}} = 0.11 \cdot (7800000 - 0.1 \cdot 7800000) = 772200 \text{ (теңге)}$$

Дәстүрлі түрде амортизация дегеніміз - компания активтері мен материалдық емес активтерінің құнын өндірілген өнімнің жалпы құнына біртіндеп беру.

Амортизация активтердің жағдайына қарамастан, компанияның капиталын (дәлірек айтқанда оның құны) өзгермеуін қамтамасыз ету үшін қажет. Амортизациялық қорлар қаржылық шегерімдер есебінен құрылады,

олардың көмегімен компанияның негізгі құралдарын құрайтын тозған объектілерді қалпына келтіру жүзеге асырылады.

Басқаша айтқанда, амортизация - бұл компанияның қаржылық және басқа активтерін сақтаудың сенімді әдісі. Амортизациялық аударымдар тозуға жататын негізгі ресурстар құнының аз пайызы. Аталған ресурстар, атап айтқанда, мыналарды қамтиды: жабдық, жылжымайтын мүлік, өндірістік қуаттылық [43].

Амортизациялық аударымдар:

$$А_{амор} = NA \cdot К/100, \quad (5.6)$$

мұндағы, NA – амортизация деңгейі (капиталдың 18%)

Байланыс жүйесінің тозуы:

$$А_{амор} = 1328427 \cdot 0.18 = 239117 \text{ (теңге)}$$

Өндіріс қажеттілігіне энергия шығындары жабдықтар мен жабдықтардың құнын қамтиды:

$$С_э = С_{эк} \cdot С_{кк}, \quad (5.7)$$

Жабдықтар үшін электростатикалық зарядтың құны келесі формула бойынша анықталады:

$$С_{эк} = W \cdot T \cdot S, \quad (5.8)$$

$$T = 24 \text{ с} \cdot 24 \text{ т} \cdot 12 \text{ э} = 2912 \text{ сағ}$$

мұндағы, W - В - номиналды қуат, Вт = 21 кВт;

T - зауыттың жылдық жұмыс режимі, T = 2912 сағат;

S - электр энергиясының құны, S = 21,99 tg/kW · h.

$$С_{эк} = 21 \cdot 2912 \cdot 21.99 = 1344732 \text{ (теңге)}$$

Қосалқы бөлшектер мен техникалық қызмет көрсету құны күрделі шығындардың 0,5% құрайды:

$$E_{мат} = 0{,}005 \cdot К, \quad (5.9)$$

Материалдардың құны:

$$E_{мат} = 0.005 \cdot 1328427 = 6642 \text{ (теңге)}$$

Шот-фактура сомасы:

$$С_{ш} = 0.75 \cdot PC = 0.75 \cdot 7800000 = 5850000 \text{ (теңге)}$$

Алматыдағы ғимаратты жалға беру, 32 ш.м., бір шаршы метрі 4400 теңге айына:

$$С_{ж} = 32 \cdot 4400 \cdot 12 = 1689600 \text{ (теңге)}$$

Шығын келесі түрде есептеледі:

$$С = C_a + С_{с} + С_{амор} + С_{эк} + С_{ж}, \quad (5.10)$$

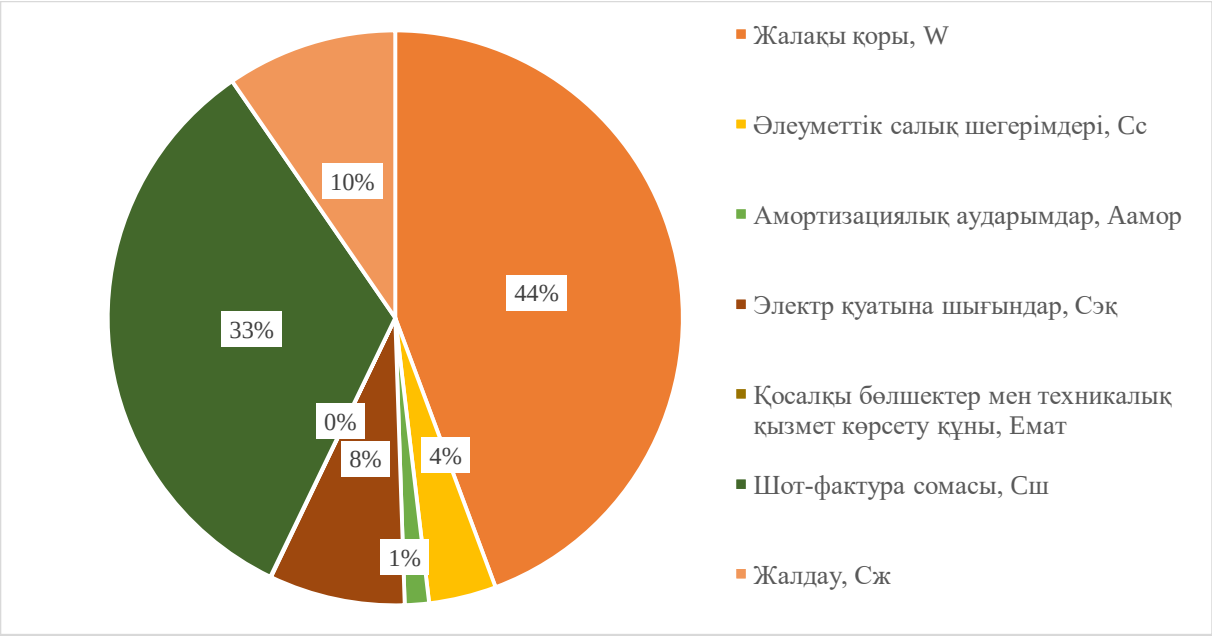
$$C = 7800000 + 772200 + 239117 + 1344732 + 1689600 = 11845649 \text{ (теңге)}$$

Сонымен, пайдалану шығындарының мөлшері:

$$E_{ш} = 7800000 + 772200 + 239117 + 1344732 + 6642 + 5850000 + 1689600 = 16357579 \text{ (теңге/жылына)}$$

5.4 кесте - Операциялық шығындардың құрылымы

Атауы	Абсолюттік мәні, теңге	%
Жалақы қоры, W	7800000	44,44988734
Әлеуметтік салық шегерімдері, $С_с$	617760	3,520431077
Амортизациялық аударымдар, $А_{амор}$	239117	1,362656886
Электр қуатына шығындар, $С_{эк}$	1344732	7,663228962
Қосалқы бөлшектер мен техникалық қызмет көрсету құны, $Е_{мат}$	6642	0,037850789
Шот-фактура сомасы, $С_{ш}$	5850000	33,3374155
Жалдау, $С_{ж}$	1689600	9,628529442
Барлығы	17547851	100



5.2 сурет – Операциялық шығындар құрылымы

Экономика бөліміне қорытынды

Дипломдық жобаның бұл бөлімінде жүргізілген есептеулер ақпараттық-аналитикалық жүйені енгізу экономикалық жағынан тиімді екенін көрсетті.

Бағдарламық қамтасыз етудің техникалық-экономикалық негіздемесі жасалып, жобаның құны талданды. Негізгі құрал-жабдықтардың шығындары, оның ішінде, жобаның күрделі шығындары, кәсіпорынның операциялық шығындары, энергия шығындары, жылдық шығындар анықталды. Есептелінген шамалардың визуалды графиктері тұрғызылды. Жобаның тасымалдау шығындары жабдық құнының 5%, ал жобаның өзіндік құны 3% құрады. Операциялық шығындардың басым бөлігіне жалақы қорының 44% мен шот-фактура сомасының 33% тиесілі болды. Сонымен бірге, жұмысшылардың жалақы құны, материалдар құны, амортизациялық аударымдар мен әлеуметтік салық есептелінді.

Ақпаратты жедел өңдеудің және телекоммуникация, интернет арқылы тұтынушыларға берудің маңыздылығы бағаланды. Бұл бизнес кірісін және тұтынушылар үшін желімен оңтайлы әрекеттесуін арттыра түсті.

Қорытынды

Бұл дипломдық жобада мәтіндік құжаттарды талдаудың негізгі әдістері және деректер ағынын талдау алгоритмдері қарастырылған, бұл мәтіндік құжаттар жиынтығы мен ағынын талдау үшін ықтималды тақырыптық модельдеуді қолдану перспективаларын растауға мүмкіндік берді. Мәтіндік құжаттар ағынын талдау жылдамдығы жаңа құжаттарды қабылдау жылдамдығынан жоғары болуы қажеттілігі анықталды. Мәтіндік құжаттарды талдау жүйесіне жалпы талаптар негізінде тақырыптық модельдеуге қойылатын нақты және негізгі талаптар тұжырымдалды.

Бұл жұмыста мәтіндер топтамасының тақырыптарын модельдеуге жартылай ықтималды көзқарас тақырыптық модельдердің аддитивті регуляризациясы (ARTM) ұсынылды. Тақырыптық модельдің құрылымы критерийлерді масштабтау арқылы бір критерий мәселесіне дейін азайтылатын көп өлшемді оңтайландыру мәселесі ретінде қарастырылады. Оңтайландыру мәселесін шешу үшін кез-келген регуляризаторларды немесе олардың сызықтық комбинацияларын алмастыруға болатын EM-дің алгоритмі ұсынылды.

Байесиялық әдіспен салыстырғанда ARTM шамадан тыс ықтималды болжамдардан бас тартуға, математикалық аппаратты жеңілдетуге және регуляризаторлардың ерікті комбинацияларын қолдануға мүмкіндік берді.

Тақырыптық модельдеуге арналған мәліметтер кезең-кезеңімен өңдеуден өтті. Алынған модельді бағалау үшін сыртқы және ішкі сапа өлшемдері пайдаланылды. Модельді оңтайландыру үшін регуляризаторлар қолданылды: Φ матрицасын сирету және Θ матрицаның декорреляциялау. Тақырыптардың регуляризация коэффициенттерін итеративті процесс жинақтала бастағаннан және нөлге жақын Φ және Θ матрицаларының элементтері анықталғаннан кейін ғана қосылды.

Тақырыптық модельдеу желідегі оқиғаларды телекоммуникация саласындағы сарапшылардың түсіндірулеріне сәйкес келетін етіп топтастыруға мүмкіндік берді.

Сонымен бірге, техникалық-экономикалық негіздеме беріліп, пайдалану шығындары мен капитал салымдары есептелінді. Өміртіршілік қауіпсіздігі жағдайлары талданды.

Әдебиеттер тізімі

1 D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of machine Learning Research, (3):993–1022, 2003.

2 Воронцов К.В., Потапенко А.А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. 2012. No 4 (4). С. 693–706.

3 Воронцов К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. No 3 (455). С. 268–271.

4 Игликов И. В., Окшин В.П., Туманбаева К. Х., Сагинтаев Ж. Вероятностное тематическое моделирование данных мониторинга, No 1 (68) http://www.ntokaxak.kz/wp-content/uploads/2020/04/Kaxak-168 2020.pdf#page=62

5 Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. – 2010. –Vol. 4, no. 2. –Pp. 280–301.

6 Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018. — 576 с.: ил. — (Серия «Бестселлеры O'Reilly»).

7 Mueller J.P., Massaron L. Machine Learning For DummiesJohn Wiley & Sons, 2016. — 435 p. — (For Dummies). — ISBN: 1119245516, 9781119245513.

8 Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах, Москва: Litres, 2017.

9 Martin Svensson, Joakim Söderberg Machine-learning technologies in telecommunications, Ericsson review, 2008: https://pdfs.semanticscholar.org/a367/f8cad03c1353e9fc36970e4cb4b8edc21fc0.pdf

10 S.M. Abdullah-Al-Mamun, JuhaValimaki Anomaly Detection and Classification in Cellular Networks Using Automatic Labeling Technique for Applying Supervised Learning: https://www.sciencedirect.com/science/article/pii/S1877050918320015

11 Xiangping Bryce Zhai, Bing Chen, Kun Zhu, Machine Learning and Intelligent Communications 4th International Conference, MLICOM 2019, Nanjing, China, August 24-25, 2019, Proceedings

12 A Ghayas Mobile communications technologies made easy: simplified view of the different generations of mobile cellular networks (Telecom networks), Independently published (July 9, 2017)

13 Воронцов К.В. Вероятностное тематическое моделирование // Москва. 2013.

14 Evaluation Methods for Topic Models: http://dirichlet.net/pdf/wallach09evaluation.pdf

15 Bayesian Reasoning and Machine Learning. David Barber. Cambridge University Press, Feb 2, 2012 - Computers - 697 pages.

16 Карпович С.Н. Тематическая модель с бесконечным словарем // Information & Control Systems/Informazionno-Upravlyaushie Sistemy. 2016. No 6 (85).

17 Y. W. Teh, D. Newman, and M. Welling. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, Advances in Neural Information Processing Systems (NIPS), number 20, pages 1481–1488, Cambridge, MA, 2008. MIT Press.

18 M. Girolami and A. Kaban. On an equivalence between PLSI and LDA. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pages 433–434, New York, NY, USA, 2003. ACM Press

19 H. Joe Steinhauer, Tove Helldin, Gunnar Mathiason, Alexander Karlsson, Topic modeling for anomaly detection in telecommunication networks, Journal of Ambient Intelligence and Humanized Computing 2019 https://link.springer.com/article/10.1007/s12652-019-01372-5

20 Tikhonov A. N., Arsenin V. Y. Solution of ill-posed problems. — W. H. Winston, Washington, DC, 1977.

21 Khalifa O., Corne D., Chantler M., Halley F. Multi-objective topic modelling //7th International Conference Evolutionary Multi-Criterion Optimization (EMO 2013). — Springer LNCS, 2013. — Pp. 51–65.

22 Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions. — 2010.

23 Tan Y., Ou Z. Topic-weak-correlated latent dirichlet allocation // 7th International Symposium Chinese Spoken Language Processing (ISCSLP). — 2010. — Pp. 224–228.

24 Ganesan A. [и др.]. LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation // preprint arXiv:150706593. 2015. 2015.

25 Newman D., Lau J. H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — HLT ’10. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. — Pp. 100–108.

26 Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — EMNLP ’11. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. — Pp. 262–272.

27 Newman D., Noh Y., Talley E., Karimi S., Baldwin T. Evaluating topic models for digital libraries // Proceedings of the 10th annual Joint Conference on

Digital libraries. — JCDL ’10. — New York, NY, USA: ACM, 2010. — Pp. 215–224.

28 Маккинли У. Python и анализ данных. / У. Маккинли, Москва: ДМК Пресс., 2015.

29 Г. Россум, Ф.Л.Дж. Дрейк, Д.С. Откидач, М. Задка, М. Левис, С. Монтаро, Э.С. Реймонд, А.М. Кучлинг, М.-А. Лембург, К.-П. Йи, Д. Ксиллаг, Х.Г. Петрилли, Б.А. Варсав, Дж.К.Ахлстром, Дж.Роскинд, Н.Шеменор, С.Мулендер. Язык программирования Python. / 2001 — 454 с.

30 Бизли Д. Python. Подробный справочник. – Пер. с англ. – СПб.: Символ-Плюс, 2010. – 864 с., ил.

31 L.Dostálek, A.Kabelová DNS in Action: A detailed and practical guide to DNS implementation, configuration, and administration, Paperback – 4 May 2006

32 Фридл Д. Регулярные выражения / Д. Фридл, под ред. А. Переводчики Матвеев, Е. Киселев, Санкт-Петербург: Символ-Плюс, 2008. 608 с.

33 Potapenko A. A., Vorontsov K. V. Robust PLSA performs better than LDA // 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. — Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2013. — Pp. 784–787.

34 Жандаулетова, Ф. Р. Охрана труда : учебник для вузов / Ф.Р. Жандаулетова, Т.Е. Хакимжанов, Т.С. Санатова; МОН РК, НАО АУЭС. - Алматы : АУЭС, 2019. - 399 с.

35 ҚР ҚН 2.02-01-2014. Ғимараттар мен имараттардың өрт қауіпсіздігі, Қазақстан Республикасы Ұлттық экономика министрлігі Құрылыс, тұрғын үй-коммуналдық шаруашылық істері және жер ресурстарын басқару комитеті «ҚазҚСҒЗИ» АҚ, «ЗЦ АТСЭ» ЖШС, Астана, 2015

36 ҚР ҚН 4.04-07-2013. Электр-техникалық құрылғылар, Қазақстан Республикасы Ұлттық экономика министрлігі Құрылыс, тұрғын үй-коммуналдық шаруашылық істері және жер ресурстарын басқару комитеті «ҚазҚСҒЗИ» АҚ, «ЗЦ АТСЭ» ЖШС, Астана, 2015

37 СН РК 4.02-01-2011 «Отопление, вентиляция и кондиционирование воздуха», Комитет по делам строительства, жилищно-коммунального хозяйства и управления земельными ресурсами Министерства национальной экономики Республики Казахстан, АО «КазНИИСА», ТОО «Сюрвейный центр», Астана, 2018

38 Абикенова А.А., Санатова Т.С. Безопасность жизнедеятельности. Методические указания к выполнению раздела «Пожарная профилактика» в выпускных работах для всех специальностей. Бакалавриат - Алматы: АИЭС, 2009. - 32 с.

39 Санатова Т.С., Мананбаева С.Е., Абдимуратов Ж.С. Тіршілік қауіпсіздігі «Нөлдеуді есептеу». Барлық мамандықтардың барлық түрінде оқитын студент-бакалаврлардың бітіру жұмысына арналған әдістемелік нұсқаулар. - Алматы: АЭжБУ, 2011. – 16 б

40 Базылов К.Б., Алибаева С.А., Нурмагамбетова С. С. Бітіруші жұмысының экономикалық бөлімі үшін әдістемелік нұсқаулар. 050719 – Радиотехника, электроника және телекоммуникация мамандығының барлық оқу түрінің студенттеріне арналған.Алматы. АУЭС.2009.

41 Экономика предприятия: Учебник / Под ред. Горфинкеля В.Я.. - М.: Юнити, 2018. - 56 с

42 Карминский, А. М. Информационно-аналитическая составляющая бизнеса / А.М. Карминский. - М.: Финансы и статистика, 2015. - 272 с.

43 Бариленко, В.И. Информационно-аналитические методы оценки и мониторинга эффективности инновационных проектов. Монография / В.И. Бариленко. - М.: Русайнс, 2015. - 747 с.

А қосымшасы

Тақырыптық модельдеу

```
1. #Тақырыптық модельдеу
2. #BigARTM-ді импорттау:
3. import pandas as pd
4. import numpy as np
5. from matplotlib import pyplot as plt
6. %matplotlib inline
7. import artm
8. #Алғашқы мәліметтерді оқу (адам үшін ыңғайлы форматқа
   түрлендіру):
9. batch_vectorizer =
   artm.BatchVectorizer(data_path='tm_data.wv',
   data_format='vowpal_wabbit',
   target_folder="top_batches")
10.    #batch_vectorizer =
   artm.BatchVectorizer(data_path="top_batches",
   data_format='batches')
11.    #Модель объектісін құру:
12.    T = 30 #тақырып саны
13.    topic_names=["topic"+str(i) for i in range(T)]
14.    model = artm.ARTM(num_topics=T,
   topic_names=topic_names, class_ids={'text':1.0,
   'subscriber':2.0},
15.    reuse_theta=True, cache_theta=True)
16.    #Сөздік құрып, оны инициализациялау:
17.    np.random.seed(1)
18.    dictionary = artm.Dictionary('dictionary')
19.    dictionary.gather(batch_vectorizer.data_path)
20.    model.initialize(dictionary=dictionary)
21.    #Көрсеткіштер
22.    model.scores.add(artm.PerplexityScore(name='perpl
   exity_score', dictionary=dictionary))
23.
24.    model.scores.add(artm.SparsityPhiScore(name='spar
   sity_phi_score', class_id="subscriber"))
25.
26.    model.scores.add(artm.SparsityThetaScore(name='sp
   arsity_theta_score'))
27.
28.    model.scores.add(artm.TopTokensScore(name='top_te
   xt', class_id='text', num_tokens=30))
29.    model.scores.add(artm.TopTokensScore(name='top_su
   bscriber', class_id='subscriber', num_tokens=30))
```

```
30.
31.   model.scores.add(artm.TopicKernelScore(name='Topi
cKernelScore', class_id='text',
32.   probability_mass_threshold=0.3))
33.
34.   #Модельді құру. Offline барлық коллекцияны
бірнеше рет өтеді:
35.   %%time
36.   model.num_document_passes = 1
37.   model.fit_offline(batch_vectorizer=batch_vectoriz
er, num_collection_passes=25)
38.   #Перплексия кестесіне сәйкес итерацияның қажетті
санын бақылауға болады.Ол өзгерді тоқтатқан кезде,
модель келісті:
39.   first_perplex_list =
model.score_tracker['perplexity_score'].value
40.   fig, ax = plt.subplots(figsize=(10,6))
41.   ax.plot(first_perplex_list)
42.   ax.set_xlabel('Iteration number')
43.   ax.set_ylabel('Perplexity')
44.   #Матрицалардың итерация саны бойынша сиретілу
графигі:
45.   plt.plot(range(model.num_phi_updates),
model.score_tracker['sparsity_phi_score'].value, 'b--
',
46.   range(model.num_phi_updates),
model.score_tracker['sparsity_theta_score'].value,
'r--', linewidth=2)
47.   plt.xlabel('Iterations count')
48.   plt.ylabel('Objective Phi sparsity, Theta
sparsity')
49.   plt.grid(True)
50.   plt.show()
51.   #матрицалардың сиретілуі:
52.   print(model.score_tracker['sparsity_phi_score'].l
ast_value)
53.   print(model.score_tracker['sparsity_theta_score']
.last_value, "\n")
54.   print(model.score_tracker['perplexity_score'].val
ue, "\n")
55.   print(model.score_tracker['perplexity_score'].las
t_value)
56.
57.   #Ядроға сәйкес t тақырыбын түсіндірудің екі
көрсеткішін анықтаймыз:
58.   – тақырыптардың тазалығы
```

```
59.    — тақырыптардың контрастілігі
60.    print(model.score_tracker['TopicKernelScore'].las
   t_average_purity) #30%
61.    print(model.score_tracker['TopicKernelScore'].las
   t_average_contrast)
62.
63.    Топ сөздерді алу:
64.    for topic_name in model.topic_names[:30]:
65.    print('\n' + topic_name + ': \n')
66.    for word in
   model.score_tracker["top_text"].last_tokens[topic_nam
   e]:
67.    print(word, end=', \n')
68.    print(' ‘)
69.
70.    smooth_phi_reg =
   artm.SmoothSparsePhiRegularizer(name='SmoothPhi',
   tau=1e5, dictionary=dictionary,
71.    class_ids={'text': 1.0}, topic_names=topic_names)
72.    model.regularizers.add(smooth_phi_reg,
   overwrite=True)
73.    %%time
74.    model.num_document_passes = 1
75.    model.fit_offline(batch_vectorizer=batch_vectoriz
   er, num_collection_passes=15)
76.    print(model.score_tracker['sparsity_phi_score'].l
   ast_value)
77.    print(model.score_tracker['sparsity_theta_score']
   .last_value)
78.    print(model.score_tracker['perplexity_score'].las
   t_value)
79.    plt.plot(model.score_tracker[‘perplexity_score'].
   value)
80.
81.    for topic_name in model.topic_names[:30]:
82.    print('\n' + topic_name + ': \n')
83.    for word in
   model.score_tracker["top_text"].last_tokens[topic_nam
   e]:
84.    print(word, end=', \n')
85.    print(' ‘)
86.
87.    #Тақырыптарды декорреляциялау регуляризаторы
88.    #Тақырыптардың әртүрлігін арттыру модельдердің
   түсініктілігін жақсартуға мүмкіндік береді
```

```
89.   decor_reg
   =  artm.DecorrelatorPhiRegularizer(name='Decorrelator
   Phi', tau=750000)
90.   model.regularizers.add(decor_reg, overwrite=True)
91.   %%time
model.num_document_passes = 1
92.   model.fit_offline(batch_vectorizer=batch_vectoriz
   er, num_collection_passes=20)
93.   print(model.score_tracker['sparsity_phi_score'].l
   ast_value)
print(model.score_tracker['sparsity_theta_score'].last_
value)

94.   print(model.score_tracker['perplexity_score'].las
   t_value)
95.   plt.plot(model.score_tracker[‘perplexity_score’].
   value)
96.
97.   print(model.score_tracker['TopicKernelScore'].las
   t_average_purity)
98.   print(model.score_tracker[‘TopicKernelScore'].las
   t_average_contrast)
99.
100.  for topic_name in model.topic_names[:30]:
101.  print('\n' + topic_name + ': \n')
102.  for word in
   model.score_tracker["top_text"].last_tokens[topic_nam
   e]:
103.  print(word, end=', \n')
104.  print(' ‘)
105.
106.  #Сирету регуляризаторы
107.  #Әр құжат және әрбір термин тақырыптардың аз
   санына байланысты болады делік.Сонда матрицалар
   арасында көптеген нөлдер болуы керек.Таралуы неғұрлым
   сирек болса, энтропиясы азаяды
108.  phi_reg =
   artm.SmoothSparsePhiRegularizer(name='SparsePhi',
   tau=-2e6, dictionary=dictionary,
109.  class_ids={'text': 1.0}, topic_names=topic_names)
110.  model.regularizers.add(phi_reg, overwrite=True)
111.  %%time
112.  model.num_document_passes = 1
113.  model.fit_offline(batch_vectorizer=batch_vectoriz
   er, num_collection_passes=15)
```

```
114.  print(model.score_tracker['sparsity_phi_score'].l
   ast_value)
115.  print(model.score_tracker['sparsity_theta_score']
   .last_value)
116.  print(model.score_tracker['perplexity_score'].las
   t_value)
117.  plt.plot(model.score_tracker['‘perplexity_score'].
   value)
118.  for topic_name in model.topic_names[:30]:
119.  print('\n' + topic_name + ': \n')
120.  for word in
   model.score_tracker["top_text"].last_tokens[topic_nam
   e]:
121.  print(word, end=', \n')
122.  print(' ‘)
123.
124.  model.theta_columns_naming = "title" #құрамына
   Theta бағандарының ішкі идентификаторы емес, сілтеме
   атаулары кіреді
125.  phi_a = model.get_phi()
126.  theta = model.get_theta()
127.
128.  model.save("tm_model")
129.  #model.load(«tm_model")
130.  #Тақырыптарды талдау
131.  #Бұдан әрі құжаттарды (Θ матрицасы) жіне
   тақырыптардағы авторлардың(Φ матрицасы , авторлар
   модальдігіне сәйкес келеді)
132.  #Құжаттардағы тақырыптардың ықтималдығы
   матрицасы:
133.  model.theta_columns_naming = "title" # #құрамына
   Theta бағандарының ішкі идентификаторы емес, сілтеме
   атаулары кіреді
134.  phi_a = model.get_phi()
135.  theta = model.get_theta()
136.  phi_a.head(10)
137.  theta.head(10)
138.
139.  #Тақырыптарды интерпретациялау
140.  #Құрылған score көмегімен тақырыптардағы топ
   сөздер мен топ авторлар ретінде көрсетеміз.
141.  import seaborn as sns
142.  subcribers_phi =
   model.get_phi(class_ids=['subscriber'])
143.  plt.figure(num=None, figsize=(20, 10), dpi=80,
   facecolor='w', edgecolor='k')
```

```
144.  sns.heatmap(subcribers_phi,yticklabels=False,
  xticklabels=topic_labels)
145.  theta.to_csv('theta.csv', header=True)
146.  subcribers_phi.to_csv('phi.csv', header=True)
147.
148.  #andoid users
149.  android_users =
  subcribers_phi.loc[(subcribers_phi.topic0 > 0),
  ['topic0','topic10']].sort_values('topic0',
  ascending=False)
150.  print("android_users:",
  len(android_users)/len(subcribers_phi))
151.  #windows users
152.  windows_users =
  subcribers_phi.loc[(subcribers_phi.topic1 > 0),
  ['topic1']].sort_values('topic1', ascending=False)
153.  print("windows users:",
  len(windows_users)/len(subcribers_phi))
154.  #instagram users
155.  instagram_users =
  subcribers_phi.loc[(subcribers_phi.topic2 > 0),
  ['topic2']].sort_values('topic2', ascending=False)
156.  print("instagram users:",
  len(instagram_users)/len(subcribers_phi))
157.  #kaspersky
158.  kasper_users =
  subcribers_phi.loc[(subcribers_phi.topic3 > 0),
  ['topic3']].sort_values('topic3', ascending=False)
159.  print("kaspersky:",
  len(kasper_users)/len(subcribers_phi))
160.  #tracker
161.  tracker =
  subcribers_phi.loc[(subcribers_phi.topic4 > 0),
  ['topic4']].sort_values('topic4', ascending=False)
162.  print("tracker:",  len(tracker)/len(subcribers_ph
  i))
163.  #samsung users
164.  samsung_users =
  subcribers_phi.loc[(subcribers_phi.topic5 > 0),
  ['topic5']].sort_values('topic5', ascending=False)
165.  print("samsung users:",
  len(samsung_users)/len(subcribers_phi))
166.  #vk
167.  vk_users =
  subcribers_phi.loc[(subcribers_phi.topic6 > 0),
```

```
['topic6','topic27']].sort_values('topic6',
ascending=False)
168.  print("vk:", len(vk_users)/len(subcribers_phi))
169.  #tiktok
170.  tiktok_users =
subcribers_phi.loc[(subcribers_phi.topic7 > 0),
['topic7']].sort_values('topic7', ascending=False)
171.  print("tiktok users",
len(tiktok_users)/len(subcribers_phi))
172.  #facebook
173.  facebook_users =
subcribers_phi.loc[(subcribers_phi.topic14 > 0),
['topic14']].sort_values('topic14', ascending=False)
174.  print("facebook users:",
len(facebook_users)/len(subcribers_phi))
175.  #netflix
176.  netflix_users =
subcribers_phi.loc[(subcribers_phi.topic17 > 0),
['topic17']].sort_values('topic17', ascending=False)
177.  print("netflix users:",
len(netflix_users)/len(subcribers_phi))
178.  #mail.ru
179.  mail_ru_users =
subcribers_phi.loc[(subcribers_phi.topic18 > 0),
['topic18']].sort_values('topic18', ascending=False)
180.  print("mail.ru:",
len(mail_ru_users)/len(subcribers_phi))
181.  #iphone users
182.  iphone_users =
subcribers_phi.loc[(subcribers_phi.topic19 > 0),
['topic19','topic23',
'topic21']].sort_values('topic19', ascending=False)
183.  print("iphone
users:",len(iphone_users)/len(subcribers_phi))
184.  #gamers
185.  gamer_users =
subcribers_phi.loc[(subcribers_phi.topic16 > 0),
['topic16']].sort_values('topic16', ascending=False)
186.  print("gamers:",
len(gamer_users)/len(subcribers_phi))
187.  #telecom
188.  telecom_users =
subcribers_phi.loc[(subcribers_phi.topic29 > 0),
['topic29']].sort_values('topic29', ascending=False)
189.  print("telecom users:",
len(telecom_users)/len(subcribers_phi))
```

```
190.
191.  print('Topic popularity:\n',
   subcribers_phi.astype(bool).sum(axis=0).sort_values(a
   scending=False)
192.  /subcribers_phi.astype(bool).sum(axis=0).sum())
193.  #most active iphone user:
194.  top_iphone_users =
   subcribers_phi.loc[subcribers_phi.topic19 > 0,
   'topic19'].sort_values(ascending=False).head()
195.  print('top_iphone_users: ', top_iphone_users)
196.  top_android_users =
   subcribers_phi.loc[subcribers_phi.topic0 > 0,
   'topic0'].sort_values(ascending=False).head()
197.  print('top_android_users: ', top_android_users)
198.  most_active_user =
   subcribers_phi.sum(axis=1).sort_values(ascending=Fals
   e).head(10)
199.  print('most_active_user: ', most_active_user)
200.  print(theta.loc[:,
   '178.89.61.113'].sort_values(ascending=False))
```