С. АМАНЖОЛОВ АТЫНДАҒЫ ШЫҒЫС ҚАЗАҚСТАН МЕМЛЕКЕТТІК УНИВЕРСИТЕТІ

Алия Нугуманова, Мадина Мансурова

ТАБИҒИ ТІЛ МӘТІНДЕРІНДЕГІ ТЕРМИНДЕРДІ АВТОМАТТЫ ТҮРДЕ ТАНУ

Монография

Өскемен
2019

Рецензия беруші:
Макашев Е.П., ф.-м.ғ.к, әл-Фараби атындағы ҚазҰУ информатика кафедрасының доценті
Жантасова Ж.З., т.ғ.кандидаты, С.Аманжолов атындағы Шығыс Қазақстан мемлекеттік университетінің компьютерлік модельдеу және ақпараттық технологиялар кафедрасының меңгерушісі

Нугуманова Алия, Мансурова Мадина

Табиғи тіл мәтіндеріндегі терминдерді автоматты түрде тану: монография / Нугуманова Алия, Мансурова Мадина. – Өскемен, ШҚМУ 2019. – 96 б.

Пәндік сала мәтіндеріндегі терминдерді автоматты түрде шығару көптеген қосымшалары бар тапсырма болып табылады. Автоматты түрде шығарылатын терминдер құжаттарды рубрикациялауда жіктеуші белгілер ретінде, тезаурустар мен онтологияларды генерациялауда семантикалық концепт ретінде, БАҚ-ның контент-талдауында тірек түсініктер ретінде пайдаланылуы мүмкін. Іс жүзінде мәтінді автоматты түрде өңдеудің барлық тапсырмаларында, аннотациялауда, индекстеуде, жіктеуде, машиналық аудармада, білімдерді алуда және т.б. терминологияны шығару қажет болады.

Аталған тапсырманы шешу үшін пәндік сала мәтіндеріндегі терминдерді шығаруды автоматтандыруға мүмкіндік беретін көптеген тиімді әдістер құрастырылған. Бұл монография табиғи тілді өңдеу технологияларының келешегі мен даму мәселелерін анықтау үшін аталған әдістерді сипаттауға және жүйелендіруге шақырады. Анықталатын мәселелердің маңыздылығы қазіргі қоғамда орын алған ақпараттық жарылыс феноменіне негізделеді. Табиғи тілді өңдеу саласындағы ғылыми қызметкерлерге, мамандарға, ЖОО оқытушыларына, докторанттарға, магистранттарға және студенттерге арналған.

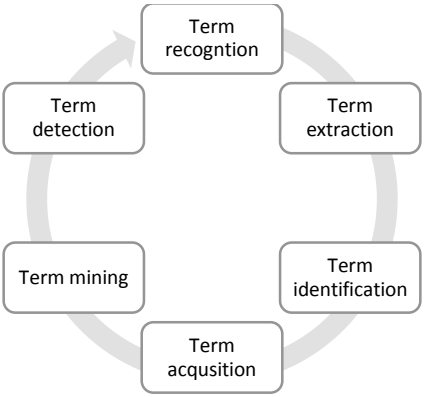
Кіріспе

Терминдерді табиғи тіл мәтіндері негізінде автоматты түрде тану - бұл ақпараттық іздеу мен білім инженериясында көптеген қосымшалары бар күрделі әрі белгілі ғылыми тапсырма. Аталған тапсырманың шешімі белгілі бір пәндік саланың терминологиялық сөздігін автоматты түрде құруға бағытталады. Тапсырманың мұндай қойылымы терминология маманы алдымен кандидат-терминдер тізімін құрып, содан кейін ақырғы сөздікті бекіту үшін пәндік сала сарапшысымен кеңесетін терминдерді дәстүрлі түрде қолмен алудың баламасы ретінде пайда болды.

Терминдерді автоматты түрде тану тапсырмасының өзектілігі терминологияны қолмен құрастыру мен сипаттаудың үнемі жаңа технологиялық салалар пайда болып, жаңа түсініктер мен терминдер туындайтын, ал техникалық лексика көлемі экспоненталды түрде артатын жылдам өзгеруші әлем жағдайында еңбек көп жұмсалатын іс екендігімен анықталады. Сол себепті терминдерді бағдарламалық құралдар көмегімен анықтау әдістері үлкен теориялық және тәжірибелік құндылыққа ие.

Бағдарламалық құралдар көмегімен алынған терминдер тізімі машина оқи алатын қарапайым білім құрылымдары болып табылады, алайда қарапайымдылығына қарамастан олардың ақпараттық іздеу мен басқа да тәжірибелік қосымшалардағы алатын орнын бағалау қиынға соғады. Терминдердің машина оқи алатын тізімдері арқылы сарапшылардың көмегінсіз құжаттарды аннотациялауға және индекстеуге, олардың тақырыптық бағытын бекітуге (тарауларға бөлуге) және машиналық оқытуды сүйемелдеуді орындауға болады. Сондай-ақ терминдердің машина оқи алатын тізімдерін таксономия және онтология сияқты күрделі білім құралдары үшін құрылыс материалдары ретінде пайдалануға да болады. [1] жұмыста машина оқи алатын тезаурустардың сандық кітапханалар үшін өте маңызды екендігі жөнінде, олар кітапхана ішінде жеңіл әрі тиімді навигация ұйымдастыруға және оның қорлары бойынша жылдам іздеу жүргізуге мүмкіндік береді.

Әдебиетте терминдерді автоматты түрде тануға арналған көптеген белгілеулер бар: терминдерді шығару (term extraction), терминдерді тану (term recognition), терминдерді сәйкестендіру (term identification), терминдерді анықтау (term detection), терминдері алу (term acquisition) және терминдерді шығарып алу (term mining). Аталған белгілеулер арасында кішігірім айырмашылықтар болса да, біз бұл жұмыста оларды синоним ретінде қарастыратын боламыз (1-суретті қараңыз).



1-сурет – Терминдерді автоматты түрде тану мәселесінің әдебиетте кездесетін белгілеулері

[2] жұмыста терминдерді пәндік сала мәтіндері негізінде автоматты түрде тану үдерісін құрайын тізбекті 5 кезең көрсетіледі (2-суретті қараңыз):

1. Корпус құрастыру (corpus collection) – терминдер алынатын пәндік сала мәтіндерінің репрезентативті корпусын компиляциялау. Егер терминдерді шығару үшін қарама-қарсы әдістер қолданылатын болса, онда жалпы сипаттағы мәтіндер корпусы да қажет. Құрастырылған корпустарға терминдерді шығарудың әрі қарай пайдаланылатын әдістеріне байланысты лемматизация (сөзді қалыпты түрге келтіру), сөз таптары бойынша белгілеу (сөздерді морфологиялық белгілеу),

фрагменттеу немесе синтаксистік талдау сияқты алдын-ала өңдеу жүргізіледі;

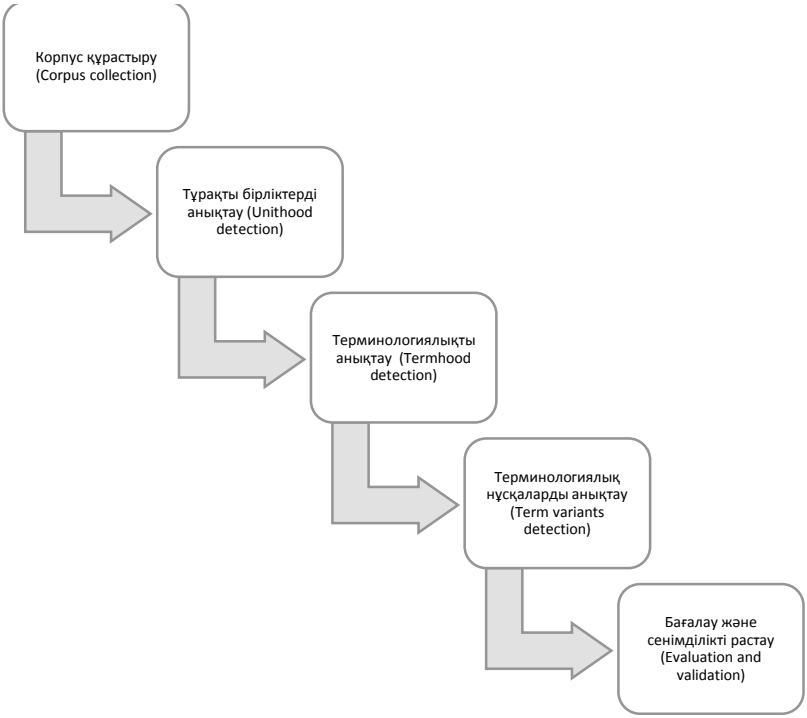
2. Тұрақты лексикалық бірліктерді анықтау (unithood detection) – бірнеше сөзден тұратын, бірақ бір ұғымдық бірлікке жататын лексикалық элементтерді сәйкестендіру;

3. Терминологиялықты анықтау (termhood detection) – шығарылған сөздің немесе тұрақты лексикалық бірліктің термин болу ықтималдығын анықтау;

4. Терминологиялық нұсқаларды анықтау (term variants detection) – пәндік саланың бірдей түсініктерінің әртүрлі лингвистикалық жүзеге асыруларын сәйкестендіру;

5. Бағалау және сенімділікті растау (evaluation and validation) – терминдерді автоматты түрде алудың пәндік сала сарапшысының қолмен жүргізетін жұмысымен салыстырғандағы сапасын бағалау тәртібі.

Берілген монографияда терминдерді автоматты түрде танудың қолдағы бар әдістеріне шолу жасалады және танудың жоғарыда келтірілген 5 кезеңінің әрқайсысының толық сипаттамасын қамтитын тәжірибелік мысал келтіріледі. Жалпы алғанда монографияның құрылымы келесідей: 1-тарауда терминдерді автоматты түрде тану мәселесіне қатысты жалпы мәліметтер келтіріледі, кіріспе анықтамалар беріліп, терминологиялық сияқты күрделі түсінікті операционалдау әдістері талқыланады. 2-тарауда терминдерді автоматты түрде тану әдістері қарастырылады және олардың жіктемесі келтіріледі. 3-тарауда “Introduction to Information Retrieval” оқулығынан терминдерді автоматты түрде шығарудың тәжірибелік мысалы қарастырылады.



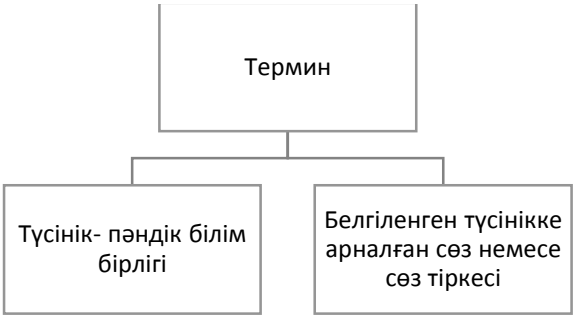
2-сурет – Терминдерді автоматты түрде тану кезеңдері

1. **ТЕРМИНДЕРДІ АВТОМАТТЫ** ТҮРДЕ ТАНУ МӘСЕЛЕСІНІҢ ҚОЙЫЛЫМЫ

1.1. Термин түсінігі және оны сөзбен қарсы қойып салыстыру. Терминологиялық белгілері

Кез келген пәндік сала сол саланың негізі болатын қатынастармен байланысатын түсініктерден тұрады. Мұндай түсініктер пәндік сала әдебиеттерінде терминдер түрінде көрсетіледі. Берілген пәндік сала үшін қай сөздердің термин, қайсыларының термин еместігін анықтау күрделі тапсырма болып табылады, оны шешу үшін бірінші кезекте не нәрсені термин деп есептеуге қатысты қойылатын талаптарды қалыптастыру қажет. Сонымен, [3] жұмыста термин дегеніміз – бұл нақты бір пәндік саладағы пәндік білімдерді жіктеуге арналған түсініктердің лингвистикалық бейнелеулері деп көрсетіледі. Басқаша айтатын болсақ, T терминін (c, t) реттелген жұбы ретінде анықтауға болады, мұндағы c – бұл пәндік сала түсінігі (білім бірлігі), ал t – оның терминологиялық түрі (3-суретті қараңыз).

Сонымен, пәндік салаға жаңа термин енгізу үшін аталған терминді көрсететін түсінік болуы керек. Терминді жіктеу үшін пәндік саладағы түсінікті басқа түсініктермен байланыстырып, топтастыру қажет. [3] жұмыста көрсетілгендей, терминдер жеке тұрғанда білім бірліктері бола алмайды, бірақ олар білім бірлігі болып табылатын түсініктерге жатады және олардың атауы белгілі бір пәнге тәуелді үлгілерді қолдануды қамтиды. Мысалы, компьютерлік ғылымдардағы жаңа терминдер әдетте қолданыста жүрген терминдерді біріктіру арқылы, ал техникалық салаларда қолданыстағы сөздерді басқа мағынада қолдану арқылы қалыптасады.



3-сурет – Терминнің формальды түрде көрсетілуі

1-кестеде жаңа терминдер қалыптастыру үшін қолданылатын негізгі 3 үлгі келтірілген.

1-кесте

Жаңа терминдерді құру үшін қолданылатын үлгілер

№	Үлгі сипаттамасы	Мысал
1	Қолданыстағы сөздің мағынасын теңеулер, метафоралар, ассоциациялар және т.с.с. көмегімен өзгертетін тілдік ресурстарды қолдану. Нәтижесінде көпмағыналы сөздер пайда болады.	1) Жады (көпшілік қолданатын мағынада) - өткен шақтағы әсерлерді, тәжірибені ойда сақтау және қайта шығару қабілеті, сондай-ақ ойда сақталатын әсерлер қоры. 2) Жады (компьютерлік ғылымдарда) - белгілі бір уақыт аралығында қолданылатын мәліметтерді сақтауға арналған физикалық құрылғы немесе орта.
2	Қолданыстағы сөздерді аффиксация, біріктіру, қысқарту және т.с.с. сияқты түрлендірулер арқылы өзгерту.	Информатика <- Информатика + Автоматтандыру. e-learning <- electronic + learning
3	Басқа тілдегі сөздергі енгізу, сөздерді тура калькалау, жаңа сөздерді ойлап табу және т.с.с. арқылы жаңа терминдер құру.	Цифрландыру <- Digitalization Қойма <- Storage Кубит <- q-bit <- quantum bit

[4] жұмыста көрсетілгендей, термин дегеніміз – бұл ғылыми түсінік немесе кәсіби қызмет пәнін жай ғана белгілеу емес, ол кәсіби ортада пайдаланылатын белгілеу болып табылады. Егер кәсіби қызметке тиістілігін алып тастайтын болсақ, онда соған сәйкес терминге қойылатын бірмәнділік, нақтылық, дәлдік және т.с.с. талаптар да алынып тасталады. Сонымен, [4] жұмыстың авторы кәсіби қызмет шеңберінен шығуды терминологиялық сөздің жалпы тіл лексикасымен бірігуінің алғашқы қадамы деген қорытындыға келеді. Сәйкесінше, терминологиялықтың алғашқы белгісін анықтап көрсетуге болады: қолданыс аясының шектеулілігі ("қайда қызмет етеді?" белгісі).

Қолданыс аясының шектеулілігі – бұл терминді қайдан іздеу керектігін көрсететін белгі [4]. Терминдерді танудың кандидат-терминді пәндік сала мәтіндерінде (кәсіби тақырыптағы) қолдану жиілігін жалпы тіл мәтіндеріндегі қолдану жиілігімен салыстыратын қарама-қарсы әдістері осы белгі бойынша құрастырылады (4-суретті қараңыз).

Шын мәнінде, мұнда сөздің қолданыс аясының қаншалықты шектеулі екендігін бағалау арқылы терминологиялықты операционалдау (тәжірибелік өлшеу) жайлы айтылуда. Бұл жерде операционалдау ретінде теориялық идеяны эмпирикалық түрге өзгерту үдерісі түсіндіріледі.

Көріп тұрғанымыздай, терминдерді жалпы тіл сөздерімен оппозициялау, яғни терминдер мен жалпы тілдегі "қарапайым" сөздерді қарама-қарсы қою беталысы тұрақты түрге еніп келеді [4]. [4] жұмыс авторының сөзінше, сөзді термин ретінде танудың маңызды шарты лексикада оның бірінші атының болуы болып табылады, ал терминнің өзі әдетте бірінші аты бар заттың екінші, ерекше аты ретінде түсіндіріледі. Сәйкесінше, терминологиялықтың екінші белгісін атап көрсетуге болады – жалпы тілдегі баламасының болуы. «Термин – жалпы тілдегі сөз» оппозициясы ғылыми түсінік (арнайы, пәндік, кәсіби білім) пен тұрмыстық түсінікті (аңқау білім), заттың өзі мен ол жайлы түсінікті ажыратуға мүмкіндік береді [4].

4-сурет – Терминологиялықты қолдану аясының шектеулілігі арқыты операционалдау.

Әрине, жалпы тілде барлық терминдердің тура баламалары бола бермейді, мысалы, жентек тас терминіне "жұмыр тектес тау жынысы" анықтамасы сәйкес келеді, ал көлеміне қарасақ, мұны балама деу қиынға соғады. Сондықтан терминологиялықты балама сөздерді іздеу арқылы операционалдауға негізделетін әдістер көп құрастырылмаған және қарама-қарсы әдістермен салыстырғанда аса танымал емес.

1.2. Терминдерді автоматты түрде танудың дәлдігі мен толықтығы

Терминдерді автоматты түрде шығаруға қатысты екі негізгі мәселе – бұл жалған анықтаулар (оларды шу – noise деп те атайды) және жалған рұқсат берулер (оларды тыныштық – silence деп те атайды). Ережеге сай, шу ретінде термин болмайтын, бірақ пәндік сала мәтіндерінде жиі қолданылатын жоғары жиілікті сөздер мен сөз тіркестері түсіндіріледі, ал тыныштық дегеніміз – термин болатын, бірақ пәндік сала мәтіндерінде сирек пайдаланылатын төменгі жиілікті сөздер мен сөз тіркестері.

Жоғарыда айтылғандай, терминдерді танудың канондық үдерісі сенімділікті растау (тексеру) үдерісімен аяқталады. Сенімділікті растауды пәндік сала сарапшылары қолмен немесе жартылай автоматты тәртіпте жүргізеді және ол алынған сөздердің терминологиялық жағдайын растауға немесе теріске шығаруға мүмкіндік береді. Осы себепті [3] жұмыстың авторлары алынған сөздерді терминдер емес, сенімділікті растау үдерісі барысында пәндік саланың ақиқат терминдері таңдалатын кандидат-терминдер деп атауды дұрыс деп есептейді.

Егер сенімділігі расталған терминдер жиынын графикалық түрде бейнелейтін болсақ, онда ол келесі екі жиынның қиылысу облысын көрсетеді: барлық кандидат-терминдер жиыны және ақиқат терминдер жиыны (5-суретті қараңыз). Шу – бұл кандидат-терминдер мен расталған терминдер жиындарының айырмасы, ал тыныштық – ақиқат терминдер мен расталған терминдер жиындарының айырмасы.

Осылайша, терминдерді шығару толықтығы мен дәлдігін оңай есептеуге болады. Кандидат-терминдер жиынын C_t, ақиқат терминдер жиынын T_t деп белгілейік, расталған терминдер жиыны, жоғарыда айтылғандай, $C_t \cap T_t$ болады. Мұндай жағдайда толықтық ($Recall$) пен дәлдік ($Precision$) келесі түрде анықталады:

$$Recall = \frac{\|C_t \cap T_t\|}{\|T_t\|} \quad (1)$$

$$Precision = \frac{\|C_t \cap T_t\|}{\|C_t\|} \quad (2)$$

Кандидат-терминдер жиыны ақиқат терминдер жиынымен сәйкес келген кезде толықтық та, дәлдік те 100%-ға тең, бірақ мұндай жағдай өте сирек кездеседі. Ережеге сай, толықтық жоғары болған сайын дәлдік төмен келеді, және керісінше. Сондықтан терминдерді автоматты түрде тану сапасын бағалау үшін дәлдік пен толықтықтың F-өлшем деп аталатын орташаланған көрсеткіш пайдаланылады. F-өлшем дәлдік пен толықтықтың орташа үйлесімі болып табылады, яғни келесі түрде анықталады:

$$F1 = \frac{2Precision*Recall}{\text{Precision+Recall}} \quad (3)$$

Терминдерді тану дәлдігі мен толықтығын есептеуді келесі мысал арқылы қарастырып көрейік. Жүйе аналитикалық химия мәтіндерінен құралған корпустан жиіліктік қағида бойынша 11 кандидат-термин шығарсын: *белсенділік, гидролиз, қышқыл, мөлшер, жылдамдық константасы, масса, моль, молярлық масса, дәреже, жылдамдық теңдеуі, электролит*. Сарапшы олардың ішінен 9-ын растап, яғни аналитикалық химияның түсініктік аппаратына қатысты деп, *мөлшер* және *дәреже* терминдерін жалпы лексика сөздері ретінде алып тастады. Аналитикалық химия бойынша терминдердің қолдан құрастырылған толық глоссарийінде 262 термин бар. Жүйе шығарған терминдер тізімінің дәлдігі мен толықтығы қандай? Бізде бары: $\|C_t\| = 11$, $\|T_t\|$=262, $\|C_t \cap T_t\| = 9$. Мұндай жағдайда дәлдік 9/11 ретінде анықталады, пайыздық түрде 81,8%-ға тең, толықтық 9/262 ретінде анықталады, пайыздық түрде 3,4%-ға тең. Сәйкесінше, F-өлшем 6,5%-ға тең.

Расталған терминдер Кандидат-терминдер Ақиқат терминдер

5-сурет – Берілген пәндік салаға қатысты расталған терминдер

Орыс тіліндегі мәтіндер үшін терминдерді танудың заманауи әдістеріндегі дәлдік пен толықтық ережеге сай 50%-дан аспайды, бірақ [5] жұмыста дәлдік 40%, ал толықтық 68,6% болатын тәжірибе нәтижелері келтіріледі.

Терминдерді шығару дәлдігі мен толықтығын арттыру үшін мәтіндегі көп сөзді терминдердің мүмкін болатын түрленулерін ескеру қажет. [6] жұмыста көрсетілгендей, «түрлену мәселесінің зерттелгеніне көп болған жоқ және терминдер мәтінде қолданылу барысында арнайы бір түсінікті білдірсе де, түрі бойынша ерекшеленуі мүмкін (мысалы: *желі архитектурасы – желілік архитектура*).

1.3. Терминдерді танудың лингвистикалық әдістері

Терминдерді шығарудың лингвистикалық әдістері интуитивті түрде түсініктірек болса да, оларды автоматтандыру оңайға соқпайды. Атауы айтып тұрғандай, ол терминдердің лингвистикалық, яғни тілдік белгілеріне негізделеді және тұрақты терминологиялық тіркестер деп аталатын күрделі көп сөзді терминдерді анықтау үшін пайдаланылады. [6] жұмыста көрсетілгендей, кез келген пәндік саланың терминдер жүйесіндегі көп сөзді терминдер үлесі қарапайым (бір сөзді) терминдермен салыстырғанда жоғары болады, сондықтан мұндай күрделі терминдерді автоматты түрде анықтау мәселесі өзекті болып табылады.

Лингвистикалық әдіс нақты тілдегі терминологиялық тіркестер белгілі бір морфологиялық және лексикалық-синтаксистік үлгілерге сәйкес құрылады деген болжамды пайдаланады [7,8]. Морфологиялық және лексикалық-синтаксистік үлгі ретінде тіл құрылымының морфологиялық,

лексикалық және синтаксистік қасиеттерін көрсететін үлгілері түсіндіріледі. Мысалы, егер берілген сөз өзінің есім тобымен бірге көмектес септігінде тұрса және *аталады* сөзінің алдында келсе, онда аталған сөздің термин болуы әбден мүмкін (*сараптамалық жүйе ...* аталады). Осылайша, мұндағы үлгі (Есім тобы көмектес септікте) + "*аталады*" + (?) құрылымы болып табылады.

[8] жұмыста терминологиялық тіркестерді шығаруға мүмкіндік беретін лексикалық-синтаксистік үлгілердің 6 тобы қарастырылады, мұндай үлгілердің бірнеше мысалы 2-кестеде келтірілген. [9,10] жұмыстарда көп сөзді терминдерді іздеуде қолданылатын морфологиялық үлгінің 9 түрі қарастырылады. Аталған үлгілердің сипаттамасы 3-кестеде келтірілген.

2-кесте

Терминдерді шығаруға арналған лексикалық-синтаксистік үлгілер мысалы ([8] жұмыстағы кесте негізінде)

№	Үлгілер тобы	Үлгілер мысалы	Терминдер мен оларды пайдалану мысалдары
1	Терминдердің морфо-синтаксистік үлгілері	(зат есім.)	• жентек тас
		(сын есім.) + (зат есім)	• кесек жыныс
		(зат есім) + (ілік септіктегі зат есім)	• жыныстың жарылуы
2	Авторлық терминдерді анықтау мәнмәтіні	(?) + "деп" "атаймыз" + (?)	• Мұндай жыныстарды *кесек жыныс деп атаймыз* -> кесек жыныс
		"ретінде" + (?) + "түсіндіріледі" (?)	• *Қайраң ретінде ,...* жер *түсіндіріледі*-> қайраң
		(?) + "бұл" "бөлік" +(?) (?) + "бұл" "тарау" +(?) (?) + "бұл" "қасиет" +(?)	• *Ішкі жүйе –бұл қандай да бір белгісі бойынша ерекшеленген жүйе бөлігі* -> ішкі жүйе • *Магниттілік– бұл минерал қасиеті ...* -> магниттілік
3	Терминдердің бірігуі	(?) + "," + (?) + {"және" | "немесе"} + (?)	• *мекен-жай, мәліметтер және басқару шинасы* -> мекен-жай шинасы, мәліметтер шинасы, басқару шинасы; • *күміс, алтын немесе платина* -> күміс, алтын,

			платина
		"те" + (?) + "," "те" "және" + (?)	• *жіңішке клиент те, жуан клиент те* -> жіңішке клиент, жуан клиент
		(?) + "," + (?) +{"," + (?) } "және" "басқалары" + (?)	• Мұндай жағдайда *гематит, сомтума күкірт, бор қышқылы, реальгар, аурипигмент, киноварь және басқа минералдар түзіледі* -> гематит, сомтума күкірт …

3-кесте

Орыс тіліндегі көп сөзді терминдерді шығаруға арналған морфологиялық үлгілер

№	Морфологиялық үлгі	Мысал
1	[зат есім+сын есім(І.с.)+зат есім (І.с.)]	• қағазсыз құжат айналымы технологиясы; • тірек векторлар әдісі.
2	[сын есім+сын есім+зат есім]	• серверлік операциялық жүйе; • интеллектуалды транспорттық жүйе.
3	[сын есім +зат есім+ зат есім (І.с.)]	• терминдерді автоматты түрде тану; • виртуалды қолжетімділік нүктесі.
4	[зат есім + зат есім (І.с.)+ зат есім (І.с.)]	• мәліметтер қоры сервері; • иерархияларды талдау әдісі.
5	[сын есім + зат есім]	• есептеу жүйесі; • ақпараттық іздеу.
6	[үстеу+ зат есім]	• жылжымалы нүкте; • алыстан қолжетімділік.
7	[зат есім + зат есім (І.с.)]	• жақындық өлшемі; • сығу коэффициенті.
8	[зат есім + зат есім (К.с.)]	• көпіршікті сұрыптау; • енгізу арқылы сұрыптау.
9	[зат есім "-"+ зат есім]	• веб-браузер; • веб-сервис.

Терминологиялық тіркестерді лингвистикалық әдістер негізінде анықтау саласындағы жұмыстарға жасалған талдау үлгілерді пайдалану терминдерді автоматты түрде тану дәлдігі мен толықтығын арттыратындығын көрсетті [10].Әрине, үлгілер көмегімен алынған сөз тіркестері арасында ақиқат

терминдер де, жоғары жиілікті сөз тіркестері де кездесетін болады (*кесек жыныс* пен *мұндай жынысты, мәліметтер шинасы* мен *мәліметтер мөлшерін* салыстырып көріңіз). Бірінші және екіншіні жіктеу күрделі есеп болып табылады, бірақ оны теориялық болмаса да, тәжірибелік деңгейде шешуге болады [11].

1.4. Терминдерді тану әдістерінің жіктелуі

Терминдерді алудың лингвистикалық әдістерін мәтіндегі лексикалық бірліктердің кездесу жиілігін есептеуге негізделетін статистикалық әдістермен жиі салыстырады. Сонымен қатар, кілттік сөздер мен терминологиялық тіркестер пәндік сала мәтіндерінде көпшілік пайдаланатын сөздермен салыстырғанда жиі кездеседі деп есептеледі (кәсіби сөздерді есептемегенде). [12] жұмыста терминдерді алу әдістерінің мұндай бинарлық жіктеуі толық емес деп көрсетіледі және анықтаушы үш белгі негізіндегі күрделірек жіктеу келтіріледі (6-суретті қараңыз):

- оқыту жүйесінің болуы бойынша;
- қолданылатын лингвистикалық ресурстардың болуы бойынша;
- терминологиялық белгілерін анықтау әдісі бойынша (операционалдау).

6-сурет – Терминдерді алу әдістерінің жіктелуі

[12] жұмыстың авторлары көрсеткендей, оқыту жүйесінің болуы бойынша әдістер үш топқа жіктеледі: оқытылмайтын, оқытылатын және өздігінен оқитын. Оқытылатын әдістер терминдерді анықтауға арналған белгілі бір белгілер жиынын пайдаланатын машиналық оқыту технологияларына негізделеді. Содан кейін аталған белгілер жиыны бойынша сәйкестендірілген терминдері бар мәтіндік мәліметтерге оқыту жүргізіледі. Әдістердің бұл тобындағы ең күрделі тапсырма белгілік кеңістікті анықтау болып табылады. [13] жұмыста барлық белгілер шартты түрде екі түрге бөлінеді: (i) кіріс корпустан статистикалық, лингвистикалық және біріккен білімдерді алатын белгілер, мысалы TF-IDF өлшемі және сөздің POS-белгісі, және (ii) аталған білімдерді кіріс корпустан басқа корпустардан алатын білімдерді пайдаланатын өлшемдерден алатын белгілер.

Кіріс саладан өзге пәндік аймаққа тиісті корпустар қарама-қарсы деп, ал қандай да бір нақты пәндік салаға жатпайтын корпустар жалпы деп аталады. Атап айтатын болсақ, авторлар жалпы корпустан Freq GC белгісін – осы корпустағы

кандидат-терминнің салыстырмалы жиілігін алады және кандидат-термин жалпы корпуста аса жиі кездеспеуі керек деп есептейді.

Қолданылатын лингвистикалық ресурстардың болуына байланысты әдістер 4 топқа бөлінеді: ресурстар қолданбайтын, сөздіктер қолданатын, онтологиялар қолданатын, мәтін корпустарын қолданатын. Соңғылары өз кезегінде 2 ішкі топқа бөлінеді: белгілері бар корпустарды пайдаланатын және белгіленбеген корпустарды пайдаланатын. [14] жұмыста лингвистикалық ресурс ретінде Википедия, ал терминологиялықтың негізгі белгісі ретінде "Википедияда гиперсілтеме болуы ықтималдығы" пайдаланылады. Авторлардың сөзінше, "жалпы лексика бөлігі болып табылатын, яғни қандай да бір пәндік салаға тиісті болмайтын сөздер мен сөз тіркестері үшін аталған белгінің мәні нөлге жуық болады".

Терминологиялық белгілерін сипаттау тәсілі бойынша әдістер 3 топқа бөлінеді: терминологиялықтың статистикалық белгілеріне негізделетін, терминологиялықтың құрылымдық белгілеріне негізделетін (лингвистикалық үлгілерге немесе графтық бейнелеулерге) және біріккен әдістер.

Егер лингвистикалық және статистикалық әдістер интуитивті бейнелеулерге негізделетін болса (бірінші жағдайда – термин қасындағы лексикалық-синтаксистік маркерлердің болуы жайлы, екінші жағдайда – пәндік сала мәтіндеріндегі терминнің қолдану жиілігінің артуы жайлы), графтық әдістер графтар теориясының математикалық аппаратына негізделеді. Графтық әдістер мәтінді төбелері сөздер немесе сөз тіркестері, ал бүйірлері олардың арасындағы қатынастар болатын граф түрінде көрсетеді. Қатынастар әртүрлі әдістер арқылы анықталуы мүмкін: мысалы, бір сөйлемде немесе берілген өлшемдегі мәтін терезесінде бірігіп кездесуді, семантикалық жақындықты көрсету. Графтың барлық төбелері ішіндегі қандай да бір орталық белгісі бойынша ең сенімді төбелер есептеледі және олар кілттік сөздер ретінде таңдалады.

Терминдер мен кілттік сөздерді алудағы ең танымал графтық әдіс нұсқаларының бірі TextRank алгоритмі болып табылады [15]. Гуглдың веб-парақшаларды саралау үшін қолданатын PageRank формуласы аталған алгоритм құрастырушыларына жігер берген екен. V_i – ағымдағы веб-парақша, ал $In(V_i)$ – бұл оған сілтеме жасайтын V_j парақшалар жиыны болсын. Онда V_i веб-парақшасының рангі (маңыздылығы) келесі формула бойынша анықталады:

$$PR(V_i) = (1 - \mathrm{d}) + d * \sum_{j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|}$$

мұндағы d – тоқтау коэффициенті, ол берілген парақшаға кірген қолданушының шығыс сілтемелердің, яғни осы парақшада тұрған сілтемелердің біріне өту ықтималдығын білдіреді (әдеттегі формулада бұл коэффициент 0.85-ке тең); $|Out(V_j)|$ – бұл қарастырылып отырған әрбір V_j парақшасындағы шығыс сілтемелер саны.

PageRank формуласы кілттік сөздерді шығару үшін келесі түрде пайдаланылды:

1. Мәтіндердің бастапқы топтамасы токенделді (лексикалық бірліктерге, яғни сөздерге бөлінді) және сөз таптары бойынша белгіленді (PoS).
2. Анықталған барлық лексикалық бірліктерге синтаксистік сүзгі жасалды (яғни төбе ретінде тек зат есімдер мен етістіктер қалдырылды).
3. Екі төбе оларға сәйкес келетін лексикалық бірліктер N сөзден тұратын терезе шеңберінде кездескен жағдайда ғана бүйірмен жалғанды. Нәтижесінде бейөлшемді бағытталмаған граф шықты.
4. Сөздерді олардың маңыздылығы бойынша саралау үшін алынған графқа PageRank алгоритмі орындалды.
5. Саралау тізімінің ең жоғары бөлігіндегі, яғни рангі ең жоғары сөздер ғана таңдалды.
6. Іргелес кілттік сөздер кілттік сөз тіркесіне жылжытылды.

1.5. Кілттік сөздер мен терминдерді алудың үлгілік мысалы

Біз бұл тарауда Алан Тьюрингтің 1950-жылы жазылған "Есептеу машиналары және сана" мақаласындағы кілттік сөздер мен терминдерді автоматты түрде шығару үшін R экожүйесін (тіл және онымен қатар жүретін табиғи тілді өңдеу кітапханалары) пайдаланамыз [16]. Тьюрингтің "машиналар ойлай ала ма" деген сұраққа арналған мақаласы орыс тілінде негізгі тарауларының саны бойынша 7 мәтіндік құжаттан (файлдан) тұратын корпус түрінде жүктелді (4-кестені қараңыз).

Алан Тьюринг өз мақаласында 2000-нан астам бірегей сөз пайдаланған, соның ішінде *машина* сөзі 297 рет кездеседі. Бұл мақаладағы "ішінде" көмекші сөзінен кейінгі ең жиі кездесетін екінші сөз. Көмекші сөздер мен кейбір байланыстырушы етістіктер сияқты қызметтік сөздерді есептемегенде мақаладағы ең жиі кездесетін келесі сөз автор 85 рет қолданған *сұрақ* сөзі болып табылады, содан кейін 83 рет қолданылған *адам* сөзі келеді. Жоғарыда көрсетілген *машина, сұрақ* және *адам* сөздерінің үшеуін де осы мақаланың кілттік сөздері деуге болады, бірақ шартты түрде олардың біреуін ғана термин деп есептеуге болады, ол – *машина* сөзі. Кілттік сөздер мен терминдер арасындағы айырмашылықты осындай көрнекі мысал түрінде көрсетуге болады. Кілттік сөздер құжаттың мазмұнын, ал терминдер – пәндік саланың мазмұнын көрсетеді.

4-кесте

«Есептеу машиналары және сана" корпусының құрамы

№	Құжат	Тақырыбы	Сөздер мен тыныс белгілерін қоса есептегендегі токендер саны
1	01.txt	I. Еліктеп ойнау	569
2	02.txt	II. Мәселенің жаңа қойылымының сыны	614
3	03.txt	III. Ойынға тартылатын машиналар	496
4	04.txt	IV. Сандық есептеу машиналары	1487
5	05.txt	V. Сандық есептеу машиналарының	1240

		әмбебаптылығы	
6	06.txt	VI. Негізгі сұрақ бойынша қарама-қарсы көзқарастар	6800
7	07.txt	VII. Оқытылатын машиналар	3165

Көрсетілген корпусты өңдеу үшін біз Чехиядағы Карлов Университетінің компьтерлік лингвистика саласындағы ғалымдары мен мамандары құрастырған UDPipe Natural Language Processing кітапханасын пайдаландық [17]. Құрастырушылардың өздері айтқандай, табиғи тілдегі үлкен мәтіндерді автоматты түрде өңдеу көптеген тілдер үшін қайталанатын тапсырма болып табылады: мәтіндер тіпті ең ілгері есептер үшін де алдымен токендеуден парсингке дейінгі алдын ала өңдеу кезеңдерінен өткізілуі керек. Оған жауап ретінде авторлар алдын ала өңдеу тапсырмаларын қандай да бір сыртқы мәліметтерді қолданбай-ақ шеше алатын бір бинарлы файл мен бір үлгіден (әрбір тіл үшін) тұратын өте қарапайым құралды ұсынады.

Сонымен, UDPipe – бұл көптеген тілдер жинағына, соның ішінде ағылшын, неміс, француз, чех, қытай, орыс, түрік, хинди, қазақ және тіпті африкаанс тілдері үшін келесі тапсырмаларды орындай алатын конвейер: 1) токендеу; 2) морфологиялық талдау және лемматизация; 3) сөз таптары бойынша белгілеу (POS-tagging); және 4) тәуелділіктерді талдау.

Сонымен қатар, конвейер CoNLL-U түріндегі жаттықтырушы мәліметтерге оңай оқытылады (кей жағдайда тіпті өңделмеген корпустарға да) және қолданушылардың төмен болса да, лингвистикалық білімінің болуын талап етеді. UDPipe C ++, Python, Perl, Java, C # үшін кітапхана ретінде де, веб-қызмет ретінде де қол жетімді болып табылады. Сондай-ақ, R үшін басқа құрастырушылар құрастырған udpipe кітапханасы да бар, біз берілген үлгілік мысалда дәл сол аталған кітапхананы пайдаланатын боламыз.

1-қадам. Кітапхананы қосу және белгілі бір тілге арналған үлгіні жүктеу. Үлгі http://ufal.mff.cuni.cz/udpipe#language_models парақшасынан жұмыс директорийіне алдын ала жүктелуі керек.

```
library(udpipe)
```

```
library(utf8)
library(readtext)
dl.rus <- udpipe_load_model (file="russian-syntagrus-ud-2.3-181115.udpipe")
```

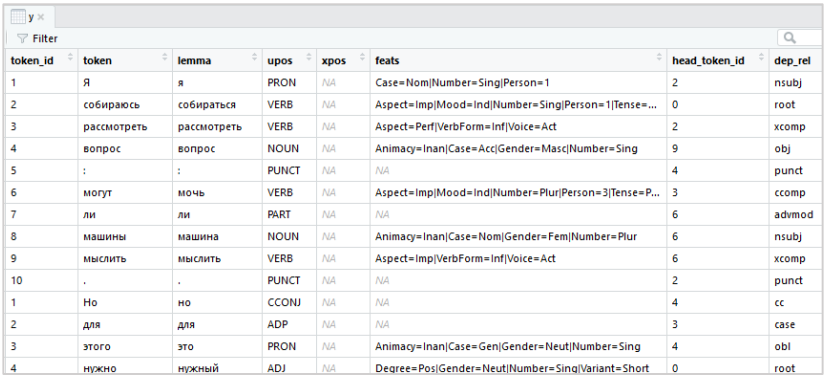
Басқа нұсқаны қолданып, файлды команда арқылы жүктеуге де болады. Мұндай жағдайда орыс тіліне арналған қарапайым үлгі көшірілетін болады, бірақ ол шектеулі болып табылады.

```
dl.file <- udpipe_download_model (language="russian")
dl.rus <- udpipe_load_model(file = dl.file$file_model)
```

2-қадам. Үлгі көмегімен мәтіндер корпусын жүктеу және оның мазмұнына сипаттама беру. Корпусты құрайтын файлдар міндетті түрде UTF8 кодында болуы керек. Егер олар басқа кодта болатын болса, онда алдын ала түрлендіріп, мазмұнын тек содан кейін ғана сипаттау керек. Мазмұндама берілген мәліметтер жиыны жолдары анықталған токендерге, ал бағандары анықталған токен белгілеріне, мысалы, морфологиялық: сөз табына, септігіне, түріне, тегіне және т.с.с. (7-суретті қараңыз) сәйкес келетін кесте болып табылады.

```
x<-readtext(paste0(getwd(), "/turing_rus/*.txt"),
    docvarsfrom = "filenames")
x$text <- enc2utf8(x$text)
y <- udpipe(x,object=dl.rus)
```

Біздің мысалымыздағы мазмұндама берілген корпус 14371 токеннен тұрады (сөздер, тыныс белгілері және мәтіннің өзге де жекеленген бірліктері), олардың әрқайсысы 18 атрибутпен (белгімен) сипатталады.



y

Filter

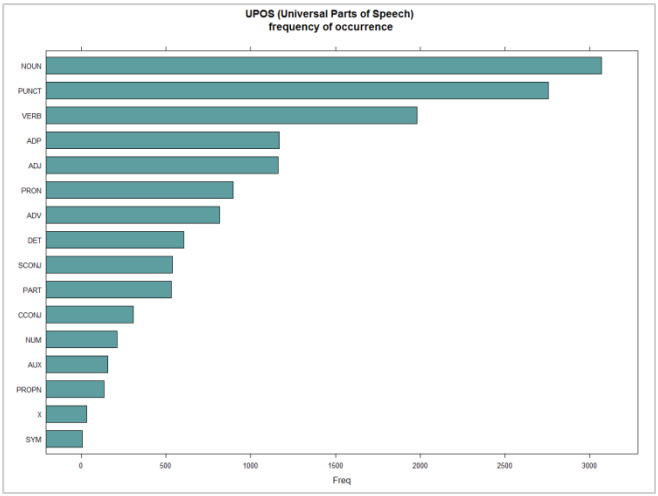
token_id	token	lemma	upos	xpos	feats	head_token_id	dep_rel
1	Я	я	PRON	NA	Case=Nom|Number=Sing|Person=1	2	nsubj
2	собираюсь	собираться	VERB	NA	Aspect=Imp|Mood=Ind|Number=Sing|Person=1|Tense=...	0	root
3	рассмотреть	рассмотреть	VERB	NA	Aspect=Perf|VerbForm=Inf|Voice=Act	2	xcomp
4	вопрос	вопрос	NOUN	NA	Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing	9	obj
5	:	:	PUNCT	NA	NA	4	punct
6	могут	мочь	VERB	NA	Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=P...	3	ccomp
7	ли	ли	PART	NA	NA	6	advmod
8	машины	машина	NOUN	NA	Animacy=Inan|Case=Nom|Gender=Fem|Number=Plur	6	nsubj
9	мыслить	мыслить	VERB	NA	Aspect=Imp|VerbForm=Inf|Voice=Act	6	xcomp
10	.	.	PUNCT	NA	NA	2	punct
1	Но	но	CCONJ	NA	NA	4	cc
2	для	для	ADP	NA	NA	3	case
3	этого	это	PRON	NA	Animacy=Inan|Case=Gen|Gender=Neut|Number=Sing	4	obl
4	нужно	нужный	ADJ	NA	Degree=Pos|Gender=Neut|Number=Sing|Variant=Short	0	root

7-сурет – Мазмұндама берілген мәтіндер корпусы

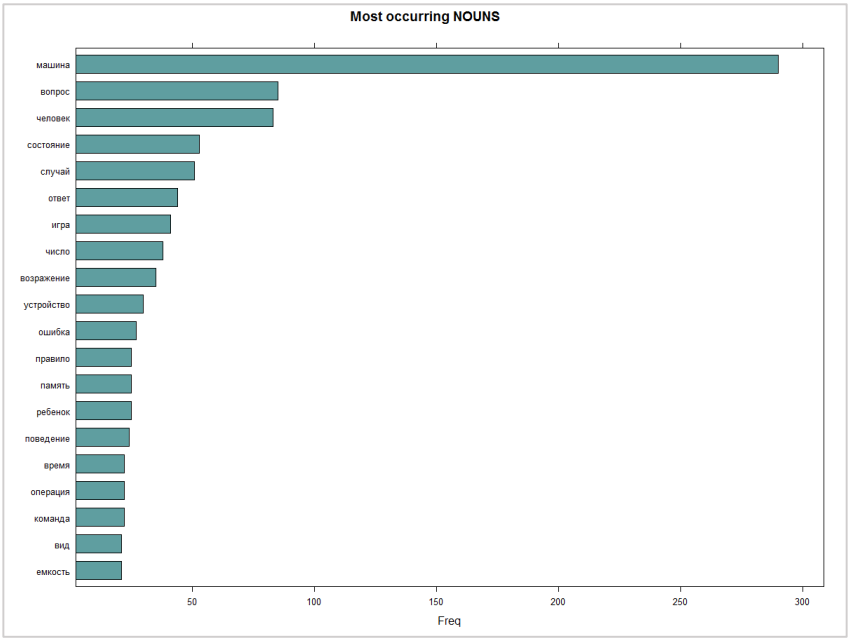
3-қадам. Корпустық статистиканы көру. udpipe кітапханасы мазмұндама берілген корпус статистикасын, мысалы, корпустағы зат есімдер, сын есімдер санын,- зат есімдер жиілігін және т.с.с. көріп отыруға мүмкіндік береді (8-10-суреттерді қараңыз). Бұл тапсырма терминдерді шығаруға тікелей қатысты болмаса да, мұндай статистика тілдік зерттеулер мен шығарманың авторлық стилі мен жанрын анықтауда пайдалы болып табылады.

```
stats <- txt_freq(y$upos)
stats$key <- factor(stats$key, levels = rev(stats$key))
barchart(key ~ freq, data = stats, col = "cadetblue",
    main = "UPOS (Universal Parts of Speech)\n frequency of occurrence",
xlab = "Freq")


stats <- subset(y, upos %in% c("NOUN"))
stats <- txt_freq(stats$lemma)
stats$key <- factor(stats$key, levels = rev(stats$key))
barchart(key ~ freq, data = head(stats, 20), col = "cadetblue",
    main = "Most occurring NOUNS", xlab = "Freq")


stats <- subset(y, upos %in% c("VERB"))
stats <- txt_freq(stats$lemma)
```

```
stats$key <- factor(stats$key, levels = rev(stats$key))
barchart(key ~ freq, data = head(stats, 20), col = "cadetblue",
   main = "Most occurring VERBS", xlab = "Freq")
```



8-сурет – Сөз таптары бөлігіндегі корпустық статистика

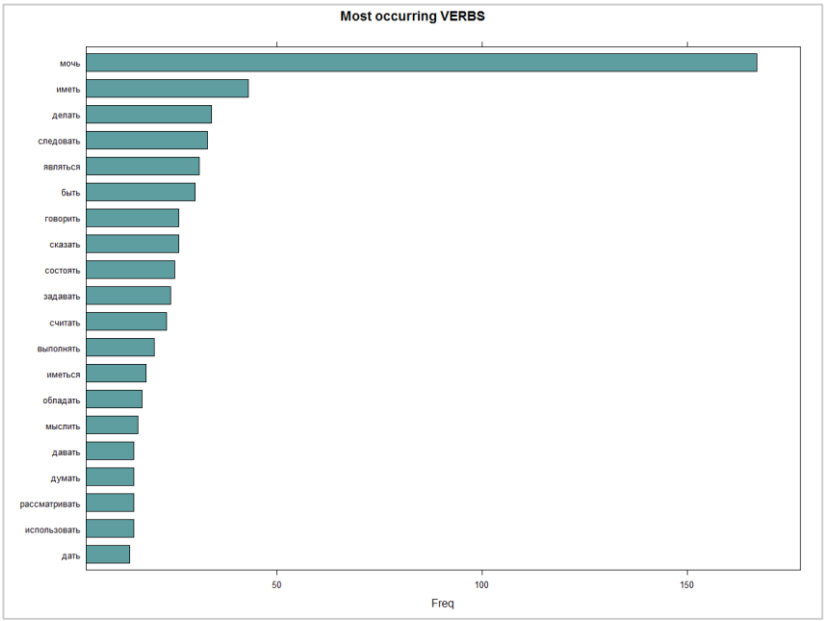


9-сурет – Зат есімдер бөлігіндегі корпустық статистика

Жоғарыда айтылғандай, терминдердің көп бөлігі шынында да күрделі, яғни көп сөзді тұрақты сөздер түрінде көрсетілуі мүмкін. Бірнеше сөзден құралған тіркесті не бірінен кейін бірі келген сөздерді анықтау арқылы, не бір сөйлемдегі немесе бір терезедегі сөздердің бірігіп кездесуінің жоғарылығын анықтау арқылы щығаруға болады. Аталған екі әдісті де udpipe кітапханасын пайдалана отырып жүзеге асыруға болады. Сондай-ақ, ізделіп отырған сөйлемдерге зат есімдер мен сын есімдер таңдалатыны көрсетілетін қосымша морфологиялық үлгіні қолдануға болады.

```
stats <- keywords_collocation(x = y, term = "token", group = c("doc_id",
"paragraph_id", "sentence_id"), ngram_max = 4)
## Co-occurrences in the same sentence, only nouns or adjectives
stats <- cooccurrence(x = subset(y, upos %in% c("NOUN", "ADJ")),
 term = "lemma", group = c("doc_id", "paragraph_id", "sentence_id"))
```

```
## How frequent do words follow one another
stats <- cooccurrence(x = y$lemma,
        relevant = y$upos %in% c("NOUN", "ADJ"))

## How frequent do words follow one another even if skip 2 words
stats <- cooccurrence(x = y$lemma,
        relevant = y$upos %in% c("NOUN", "ADJ"), skipgram = 2)
head(stats)
```



10-сурет – Етістіктер бөлігіндегі корпустық статистика

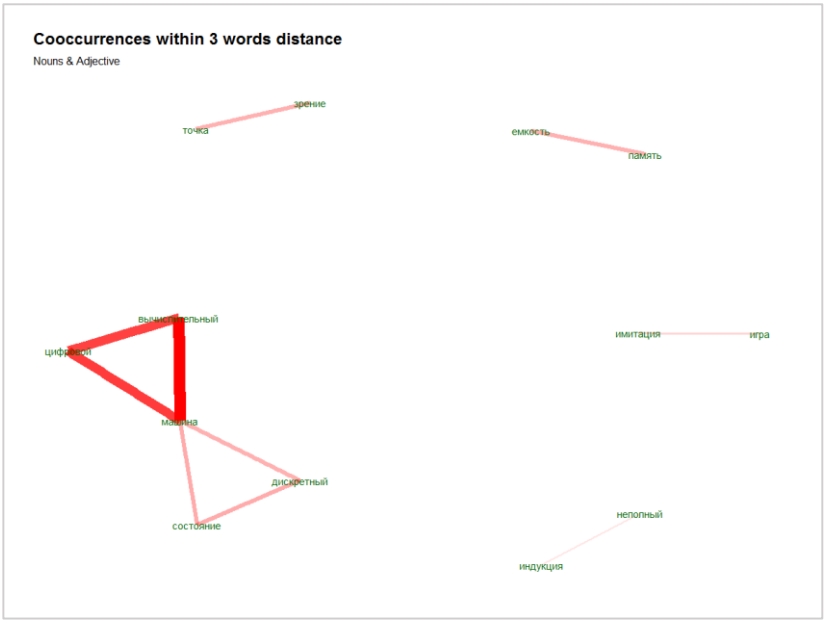
Тұрақты тіркестерді, яғни коллокацияларды граф түрінде оңай көрсетуге болады (11-суретті қараңыз).

```
library(igraph)
library(ggraph)
```

```
library(ggplot2)
wordnetwork <- head(stats, 10)
wordnetwork <- graph_from_data_frame(wordnetwork)
ggraph(wordnetwork, layout = "fr") +
 geom_edge_link(aes(width = cooc, edge_alpha = cooc), edge_colour = "red") +
 geom_node_text(aes(label = name), col = "darkgreen", size = 4) +
 theme_graph(base_family = "Arial Narrow") +
 theme(legend.position = "none") +
 labs(title = "Cooccurrences within 3 words distance", subtitle = "Nouns & Adjective")
```

Сонымен қатар, кірістірілген RAKE – кілттік сөздерді алудың ең қарапайым және жылдам әдісін пайдалануға да болады (12-суретті қараңыз).

```
stats <- keywords_rake(x = y, term = "lemma", group = "doc_id",
         relevant = y$upos %in% c("NOUN", "ADJ"))
stats$key <- factor(stats$keyword, levels = rev(stats$keyword))
barchart(key ~ rake, data = head(subset(stats, freq > 3), 20), col = "red",
    main = "Keywords identified by RAKE", xlab = "Rake")
```



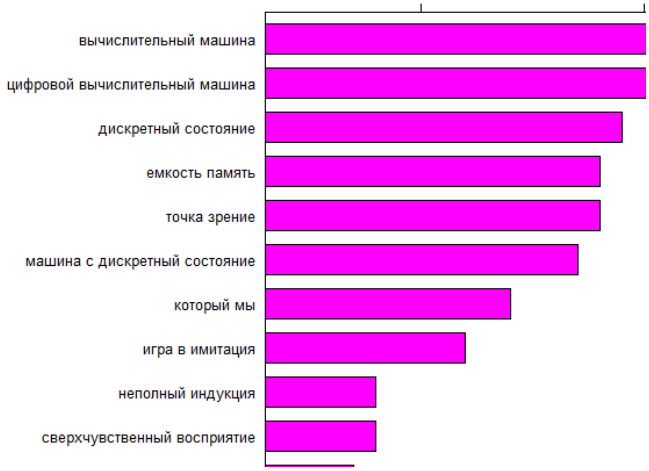
11-сурет – Тұрақты тіркестер графы

Көзқарас, жады мөлшер, есептеу машина, аса сезімтал қабылдау, кездейсоқ элемент, оқыту үдеріс, қазіргі сәт, дискретті күй

12-сурет – RAKE әдісі бойынша алынған кілттік сөздер мен тұрақты тіркестер

Күрделірек коллокацияларды тұрақты тіркестер паттерні арқылы шығаруға болады (13-суретті қараңыз). Мысалы, "(A|N)*N(P+D*(A|N)*N)*" паттерні көмегімен *еліктеп ойнау* немесе *дискретті күйі бар машина* сияқты тіркестерді шығаруға болады. Дегенмен, біз *көзқарас* сияқты көпшілік қолданатын тіркестер үлесінің бұл жерде де жоғары екендігін байқаймыз.

```
y$phrase_tag <- as_phrasemachine(y$upos, type = "upos")
stats <- keywords_phrases(x = y$phrase_tag, term = tolower(y$token),
          pattern = "(A|N)*N(P+D*(A|N)*N)*",
          is_regex = TRUE, detailed = FALSE)
stats <- subset(stats, ngram > 1 & freq > 2)

stats$key <- factor(stats$keyword, levels = rev(stats$keyword))
barchart(key ~ freq, data = head(stats, 20), col = "magenta",
   main = "Keywords - simple noun phrases", xlab = "Frequency")
```



есептеу машина, сандық есептеу машина, дискретті күй, жады мөлшер, көзқарас, дискретті күйі бар машина, біз, еліктеу ойнау, толық емес индукция, аса сезімтал қабылдау

13-сурет – Тұрақты тіркестер паттерндері көмегімен шығарылған кілттік сөздер мен тұрақты тіркестер

Соңында айтарымыз, терминдерді алдыңғы тарауда қарастырылған TextRank әдісін арқылы шығару осы аттас кітапхана арқылы іске асырып, wordcloud кітапханасы көмегімен визуалдауға болады (15-суретті қараңыз).

```
library(textrank)
stats <- textrank_keywords(y$lemma,
          relevant = y$upos %in% c("NOUN", "ADJ"),
          ngram_max = 8, sep = " ")
stats <- subset(stats$keywords, ngram > 1 & freq >= 3)

library(wordcloud)
wordcloud(words = stats$keyword, freq = stats$freq, colors = brewer.pal(5,"Pastel1"))
```

Көріп тұрғанымыздай, кілттік сөздерді тек лингвистикалық және статистикалық әдістерді пайдалану арқылы ғана

шығаруға болады, ал терминдерді шығару тапсырмасы басқаша. Терминдер – пәндік сала түсініктерінің сөздік өрнектері ретінде – пәндік сала мәнмәтінінде ғана қарастырылуы керек және сәйкесінше оларды шығару үшін сыртқы білім көздері қажет, мысалы берілген терминнің тек берілген пәндік салаға ғана қатысты екенін және қарама-қарсы корпустарда кездеспейтінін растайтын қарама-қарсы әдістер қажет. Монографияның келесі тарауы терминдерді шығарудың қарама-қарсы әдістеріне арналады.

14-сурет – TextRank әдісі көмегімен шығарылған кілттік сөздер және тұрақты тіркестер

2. ТЕРМИНДЕРДІ АВТОМАТТЫ ТҮРДЕ ТАНУ ӘДІСТЕРІ

2.1. Терминдерді автоматты түрде танудың қарама-қарсы әдістері

Қарама-қарсы әдіс – бұл терминдерді олардың пәндік сала ішінде және оның шеңберінен тыс әртүрлі кездесуіне байланысты анықтайтын әдістердің жалпы атауы [18,19,20]. Аталған әдістердің барлығын сөздің терминологиялығын олардың мақсатты (пәндік) және балама жинақтардағы үлестірімдерін салыстыру арқылы анықтауға қатысты ортақ идея біріктіреді. Балама жинақ ретінде не қарама-қарсы, яғни басқа пәндік сала мәтіндерінен құрастырылған жинақ, не жалпы, яғни ешбір пәндік салаға жатпайтын мәтіндерден құрастырылған жинақ қолданылуы мүмкін [21].

Терминдерді қарама-қарсы жолмен алуға арналған жұмыстардың бірі ретінде [22] атауға болады. Оның авторлары терминологиялықты бағалау үшін оғаштық (ағылш. Weirdness) деп аталатын жаңа, интуитивті түсінікті өлшемді енгізеді. Оғаштық әрбір кандидат-термин үшін есептеледі және ол мақсатты жинақта қолдану жиілігінің жалпы жинақтағы қолданылу жиілігіне қатынасы ретінде анықталады. Әдетте жинақтар көлемі жағынан теңдестірілмейтіндіктен, салыстырмалы жиіліктер қолданылады.

Әдеттегі сөздер үшін оғаштық формуласы 1-ге жуық, ал терминдер үшін 1-ден әлдеқайда үлкен мәндерді қайтарады, себебі мұндай жағдайда формуланың бөлімі 0-ге жуықтайды:

$$Weirdness = \frac{F_{SL}/N_{SL}}{F_{GL}/N_{GL}} = \frac{F_{SL} \cdot N_{GL}}{F_{GL} \cdot N_{SL}} \quad (5),$$

мұндағы F_{SL}, F_{GL} – бұл сөзді сәйкесінше мақсатты (SL) және жалпы (GL) жинақтарда қолдану жиіліктері; N_{SL}, N_{GL} – бұл сәйкесінше мақсатты және жалпы жинақтардағы барлық сөздер саны.

Авторлар өздерінің кейінгі, мысалы [23], жұмыстарында (1) формуланың түрлендірілген нұсқасын ұсынады, себебі олардың айтуынша, бөлшек бөлімі 0-ге айналғанда бастапқы формула сингулярлық танытады. Бұл сөздің жалпы жинақтағы

қолдану жиілігі 0-ге тең болып, нәтижесінде шексіздікке алып келетін жағдайларда орындалады. Оғаштықтың түрлендірілген, тегістелген формуласы бастапқы формуладан сөздің жалпы жинақтағы қолданылу жиілігіне 1-ді қосумен ерекшеленеді:

$$Weirdness = \frac{F_{SL} \cdot N_{GL}}{(1+F_{GL}) \cdot N_{SL}} \quad (6)$$

Әрі қарай [22] және [23] жұмыстарды салыстыру авторлардың бірінші жұмыста терминдер тізімін тек оғаштықтың жоғары мәндері негізінде ғана құрғандығын, ал екінші жұмыста олардың жиілігі жоғары жоғары оғаштықтар үйлесімін пайдаланғанын көрсетеді. Олар осылай ету арқылы мақсатты жинаққа кездейсоқ кіріп кеткен, яғни пәндік салаға жатпайтын "оғаш" сөздерден құтылуға тырысады. Мұндай әдіс терминдерді қамту дәлдігінің жоғарылауына кепілдік береді, бірақ жоғарыда атап өткеніміздей, толықтық азаяды, себебі сирек кездесетін терминдер қарастырылмай қалады.

[24] жұмыста терминдерді қарама-қарсы бағалау идеясы бір емес, екі тұжырым түрінде қалыптасады. Біріншіден, мақсатты жинақта сирек қолданылатын сөздердің бағасы төменірек болуы керек. Екіншіден, мақсатты жинақта жиі қолданылатын сөздердің бағасы жоғарырақ болуы керек, бірақ олар қарама-қарсы жинақта немесе мақсатты жинақ мәтіндерінің шектеулі жинағында жиі кездеспеуі керек. Авторлар бұл тұжырымдарды релеванттық (relevance) деп аталған өлшем түрінде операционалдандырады:

$$Relevance = \frac{1}{log_2\left(2+\frac{f_{SL} \cdot N_{SL}^t}{f_{GL}}\right)} \quad (7),$$

мұндағы f_{SL} және f_{GL} – бұл t сөзін сәйкесінше мақсатты және қарама-қарсы жинақтарда қолданудың салыстырмалы жиіліктері; N_{SL}^t – мақсатты жинақтың берілген сөз кездесетін мәтіндерінің салыстырмалы саны. Келтірілген өлшем репрезентативті терминдерді шығаруда жақсы нәтижелер көрсетеді, бірақ сирек кездесетін терминдердің бағасын жасанды түрде түсіреді, ал бұл, жоғарыда атап өткеніміздей, терминдерді қамту толықтығына кері әсерін тигізеді.

[25] жұмыста терминологиялықты қарама-қарсы әдіс негізінде бағалаудың өзгеше тәсілі ұсынылады. Бұл әдіс сөздерді өлшеудің белгілі TF-IDF формуласын пайдаланады, оған сәйкес сөздің құжатта қолданылу жиілігі жоғары болған сайын және бүкіл жинақ бойынша үлестірілуі төмен болған сайын оның сол құжаттағы салмағы жоғары болады. Формуланың авторлар «term frequency – inverse domain frequency» деп атаған жаңа нұсқасында сөздің құжаттағы емес, мақсатты жинақтағы салмағы бағаланады. Жаңа формула бойынша сөздің мақсатты жинақта қолданылуының салыстырмалы жиілігі жоғары және барлық жинақтар бойынша салыстырмалы үлестірілуі төмен болған сайын оның салмағы жоғары болады:

$$TF \cdot IDF = TF(t, D) \cdot IDF(t) = \frac{n_{t,D}}{\sum_k n_{k,D}} \cdot log\left(\frac{|TS|}{|\{d: t \in d\}|}\right)$$

мұндағы $n_{t,D}$ – бұл t сөзінің D мақсатты жинағына кіру саны, $\sum_k n_{k,D}$ – бұл D мақсатты жинағына барлық сөздердің кіру саны, $|TS|$ – бұл барлық қолданылатын жинақтардағы құжаттар саны, $|\{d: t \in d\}|$ – бұл t сөзі кем дегенде бір рет кіретін барлық құжаттар саны. Сонымен, авторлар құжаттардың санаулы ішкі жиындары шеңберінде көп шоғырланатын барлық сөздерді терминдер деп атайды. Әрине, бұл терминдердің белгілі бір бөлігі үшін өте жақсы әдіс, бірақ сирек кездесетін терминдер үшін аса пайдалы деуге келмейді.

[26] авторлары сөздердің терминологиялығын TF-IDF формуласы негізінде бағалауды да ұсынады. Олар өздері ұсынған формула нұсқасын қарама-қарсы салмақ (contrastive weight) деп атайды және оны сөздің мақсатты жинақта қолданылу жиілігі жоғары және оны қарама-қарсы жинақтарда қолданудың салыстырмалы жиілігі төмен болған сайын жоғары болатын өлшем ретінде анықтайды:

$$Contrastive\ Weight = TF(t, D) \cdot IDF(t) = \log\left(f_t{}^D\right) \cdot \log\left(\frac{F_{TC}}{\sum_j f_t^j}\right)$$

мұндағы $f_t{}^D$ – сөзді мақсатты жинақта қолдану жиілігі, $\sum_j f_t^j$ – сөздің қарама-қарсы жинақтардағы барлық қолдануларының қосындысы, $F_{TC} = \sum_{i,j} f_i^j$ – мақсатты жинақты қоса алғанда барлық жинақтардағы барлық сөздердің қолдану жиіліктерінің

қосындысы. Авторлар өздері айтып кеткендей, қарама-қарсы салмақ сөздердің терминологиялығын таза жиіліктерге қарағанда жақсы бағалайды, бірақ олардың сөзінше, әдістің F-өлшем бойынша анықталған жалпы тиімділігі көзге аса түсе қоймайды.

[27] жұмыста қарама-қарсы салмақ формуласына сыни баға беріледі. Берілген жұмыстың авторлары көрсеткендей, қарама-қарсы салмақ пен оған ұқсас өлшемдер шын мәнінде терминдердің пәндік салаға жататындығын емес, олардың үлестірілуін бағалайды. Авторлар аталған кемшілікті түзету үшін терминологиялықты екі көрсеткіш негізінде бағалауды ұсынады: сөздің мақсатты DP (ағылш. domain prevalence) жинағында көп болу өлшемі және сөздің мақсатты DT (ағылш. domain tendency) жинағына тартылу өлшемі. DP мәнінің жоғары болуы мақсатты топта басқа сөздермен салыстырғанда берілген сөздің көп екендігін көрсетеді. DT мәнінің жоғары болуы сөздің қарама-қарсы жинақпен салыстырғанда мақсатты жинақта көп кездесетінін білдіреді. DP-ны есептеуге арналған формула шын мәнінде қарама-қарсы салмақ формуласының (9) тегістелген нұсқасы болып табылады:

$$DP(t) = log_{10}({f_t}^D + 10) \cdot log_{10}\left(\frac{F_{TC}}{{f_t}^D + {f_t}^{\overline{D}}} + 10\right)$$

мұндағы ${f_t}^D$ және ${f_t}^{\overline{D}}$ – берілген сөздің сәйкесінше мақсатты және қарама-қарсы жинақтарда қолданылу жиіліктері, $F_{TC} = \sum_j f_j^D + \sum_j f_j^{\overline{D}}$ – барлық кандидат-терминдердің сәйкесінше мақсатты және қарама-қарсы жинақтарда қолданылу жиіліктерінің қосындысы.

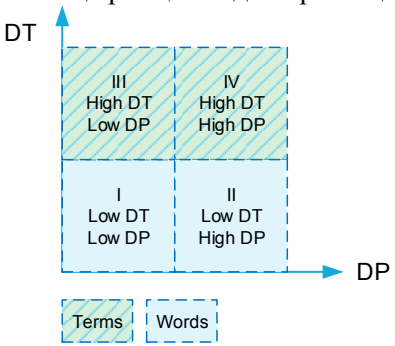
DT-ны есептеуге арналған формула оғаштық формуласының (5) тегістелген нұсқасы болып табылады, яғни қарама-қарсы жинақта жиі кездесетін сөздерге айыппұл салады:

$$DT(t) = log_2\left(\frac{{f_t}^D + 1}{{f_t}^{\overline{D}} + 1} + 1\right) \quad (11)$$

DP және DT өлшемдері дискриминациялық салмақ (discriminative weight) деп аталатын ортақ бір көрсеткішке біріктіріледі. Авторлардың ойынша бұл көрсеткіш жоғары дифференциалдаушы қасиетке ие:

$$DW(t) = DP(t) \cdot DT(t) \quad (12)$$

DT және DP көрсеткіштерінің бір-бірімен қатты корреляцияланатындығын айта кету керек. Мысалы, біздің тәжірибелерімізде аталған көрсеткіштердің корреляциясы 0,71-ден 0,82-ге дейін болды. Корреляция табиғатын түсіну үшін біз барлық кандидат терминдерді DT мен DP мәндеріне байланысты қиылыспайтын 4 топқа бөлдік: 1) DT мен DP мәндері орташадан төмен; 2) DT мәндері орташадан төмен, ал DP мәндері орташадан төмен емес; 3) DT мәндері орташадан төмен емес, ал DP мәндері орташадан төмен; 4) DT мен DP мәндері орташадан төмен емес. Сараптамалық бағалар да, (8) формула негізіндегі бағалар да бір нәтиже көрсетті: аздаған нәрселерді ескермейтін болсақ, тек 3 және 4 топтағы кандидаттар ғана термин бола алады, ал бұл DT көрсеткішінің үлкен мәндеріне сәйкес келеді (15-суретті қараңыз). Берілген нәтиже DT көрсеткішінің ақпараттылығының жоғарылығын және DP көрсеткішінің артық екендігін растайды.



15-сурет – DP және DT белгілерін салыстыру

Сондай-ақ, "аздаған нәрселерді ескермеу" тіркесі кездейсоқ емес. терминдерді бағалаудың ұсынылған әдісінің сенімділігін растау терминологиялықтың төменгі мәндері шеңберінде әдеттегі сөздер арасында біз сирек деп атайтын терминдер кішігірім ерекшеліктер түрінде кездеседі. Бұлар мақсатты жинақта 1-2 рет кездесетін және қарама-қарсы жинақта мүлдем кездеспейтін терминдер. Мұндай терминдерді шығару

дифференциалдаудың таңдамалы құралдарын қолдануды талап етеді.

Терминологиялықты бағалау үшін бірден бірнеше көрсеткішті пайдалану тек [27] жұмыста ғана кездеспейді. [28] жұмыста аталған мақсат үшін бірден 3 көрсеткіш пайдаланылады: DR сәйкестілік өлшемі (ағылш. domain pertinence), DC келісушілік өлшемі (ағылш. Domain consensus) және көп сөзді терминдердің когезиясын бағалауға арналған LC лексикалық когезия (ағылш. Lexical cohesion).

Нәтижесінде Di мақсатты жинағындағы t сөзінің терминологиялығының ақырғы бағасы аталған үш өлшемнің сызықты комбинациясынан құралады:

$$w(t, Di) = \alpha \cdot DR + \beta \cdot DC + \gamma \cdot LC \qquad (13),$$

мұндағы α, β, γ – бұл іріктеу параметрлері, үнсіздік бойынша $\alpha = \beta = \gamma = 1/3$.

DR сәйкестілік өлшемі қарама-қарсы жинақтар жиындары үшін жалпыландырылған оғаштық өлшемі болып табылады. Ол сөздің мақсатты жинақтағы жиілігінің қолдағы бар барлық қарама-қарсы жинақтардағы ең үлкен жиілігіне қатынасы арқылы анықталады:

$$DR(t, Di) = \frac{freq(t,Di)}{\max\limits_{j}(freq(t,Dj)} \qquad (14)$$

DC келісушілік өлшемі сөздердің жеке құжаттардағы үлестірімін ескеруге мүмкіндік береді. Ол t сөзінің Di мақсатты жинақ құжаттарында кездесуінің қалыпқа келтірілген ϕ_k жиіліктері арқылы анықталады және сөздердің аталған құжаттардағы үлестірімі бірқалыпты болған сайын жоғары келеді:

$$DC(t, Di) = -\sum_{d_k \in Di} \phi_k log\, \phi_k \qquad (15)$$

Авторлар келісушілік өлшемін енгізу арқылы оның маңыздылығын мақсатты жинақтың көп құжаттарында жиі кездесетін терминдер шектелген құжаттарда жиі кездесетін терминдермен салыстырғанда жоғары бағалануы керектігін айтып дәлелдейді. Мұндай тұжырым [18] жұмыста қолданылған эвристикаға толықтай антагонистік болып табылады. Терминологиялық белгілерін таңдауда кездесетін

қарама-қайшылықты көрсететін мұндай қызықты деректі [29] жұмыстың авторлары да атап көрсетеді.

[30] жұмыста терминологиялық сөздердің мақсатты және қарама-қарсы жинақтардағы қолданылу жиіліктерінің қатынасы емес, олардың рангтерінің айырмасы түрінде анықталады. Жинақтағы сөз рангі ретінде оның жинақтағы барлық сөздерден құралған және оларды пайдалану жиіліктері бойынша іріктелген тізімдегі орны түсіндіріледі. Берілген жинақта кездеспейтін сөздердің рангі 0-ге тең. 1 рангі жинақтағы ең сирек кездесетін сөзге тән.

Сонымен, t сөзінің мақсатты D және қарама-қарсы G жинақтарындағы рангтерінің айырмасы түрінде өрнектелген терминологиялық индексі келесідей болады:

$$thd(t,D) = \frac{rank(t,D)}{|V(D)|} - \frac{rank(t,G)}{|V(G)|} \quad (16),$$

мұндағы $V(D)$ және $V(G)$ – бұл сәйкес топтамалардың сөздіктері. Индекс -1 мен 1 аралығындағы мәндерді қабылдайды, 1 мәні сөз мақсатты жинақта ең жоғары және қарама-қарсы жинақта нөлдік рангке ие болатын жағдайға сәйкес келеді. Сөздерді индекстің кемуі бойынша саралау мақсатты жинақтың репрезентативті сөздерін жоғарыға шығаруға мүмкіндік береді (олардың ішінде терминдер де болады). Алайда, авторлардың айтуы бойынша, ұсынылған саралау тек терминдерді ғана анықтауға мүмкіндік бермейді.

Осы қатарда көрсеткіміз келген соңғы жұмыс [31]. Ол да TF-IDF формуласының негізгі құрылымындағы айыппұлдар мен сыйақылар идеясын көрсетеді және аталған формуланың “term frequency – disjoint corpora frequency” деп аталған жаңа нұсқасын ұсынады. Сыйақы ретінде сөзді мақсатты жинақта пайдаланудың абсолютті жиілігі, ал айыппұл ретінде сөздің қарама-қарсы жинақтарда қолданылуының абсолютті жиіліктерінің көбейтіндісі қолданылады:

$$TF \cdot DCF = \frac{{f_t}^D}{\prod_{g \in G} 1 + log(1 + {f_t}^g)} \quad (17),$$

мұндағы ${f_t}^D$ және ${f_t}^g$ – берілген t сөзін сәйкесінше мақсатты және қарама-қарсы жинақтарда пайдалану жиіліктері, G – барлық қарама-қарсы жинақтар жиыны.

Авторлар өздері ұсынған формуланың терминдерді шығару дәлдігі бойынша [25,30] жұмыстарда ұсынылған және басқа да жұмыстардағы бағалармен салыстырғанда жақсы екендігін тәжіриебелер арқылы дәлелдейді. Олар формула бөлшегінде көбейтіндіні пайдалануды сөзді әрбір кезекті қарама-қарсы жинақта пайдаланған сайын айыппұлдың геометриялық прогрессия бойынша өсуі керектігімен түсіндіреді. Авторлардың пікірінше, көп көлемдегі қарама-қарсы жинақтарда аз қолданылатын сөздердің терминологиялығы аздаған қарама-қарсы жинақтарда көп кездесетін сөздердің терминологиялығымен салыстырғанда төмен бағалануы керек. Бір қарама-қарсы жинақ болған жағдайда формуланың нәтижелері жоғары жиілікті терминдердің бір жағына қарай ығысады.

Сонымен, біз берілген шолуда терминологиялық бағасын операционалдаудың ең қызықты 8 қарама-қарсы әдісін қарастырдық (5-кестені қараңыз).

5-кесте

Терминдерді шығарудың қарама-қарсы әдісін операционалдаудың маңызды әдістері

№	Авторлар және дереккөзге сілтеме	Өлшем немесе индикатор атауы	Жылы
1	Ahmad et al [22, 23]	Weirdness	1999, 2005
2	Peñas A. et al [24]	Relevance	2001
3	Kim et al [25]	Term frequency- inverse domain frequency	2009
4	Basili et al [26]	Contrastive weight	2001
5	Wong et al [27]	Domain prevalence, Domain tendency	2007
6	Sclano et al [28]	Domain pertinence, Domain consensus	2007
7	Kit C., Liu X [30]	Termhood index	2008
8	Lopes et al [31]	Term frequency- disjoint corpora frequency	2016

Бұл әдістердің барлығы эвристикалық болып табылады, яғни терминдердің мақсатты және қарама-қарсы жинақтарда үлестірілу сипатына қатысты болжамдарға негізделеді. Бұл тұжырымдардың салыстырмалы талдауы әр түрлі авторлар ұстанымдарының сәйкес келетіндігін де, айтарлықтай

айырмашылықтар болатындығын да көрсетеді, ал бұл өз кезегінде берілген салада шешілмеген мәселелердің бар екендігін растайды.

2.2. Хи-квадрат, ақпараттық пайда және өзара ақпарат өлшемдері

Біз бұл тарауда 3 белгілі қарама-қарсы өлшемдерді қарастырамыз: өзара ақпарат өлшемі, ақпараттық пайда және хи-квадрат өлшемі [32-34]. Аталған өлшемдерді сипаттамас бұрын оларға ортақ белгілеулерді 6-кестеде көрсетілгендей етіп енгіземіз.

6-кесте

Белгілерді таңдау өлшемдері үшін қолданылатын белгілеулер

Символ	Мағынасы
TS	Құжаттардың барлық жиыны
A	*t* терминін қамтитын позитивті жиын құжаттарының саны
B	*t* терминін **қамтымайтын** позитивті жиын құжаттарының саны
C	*t* терминін қамтитын негативті жиын құжаттарының саны
D	*t* терминін **қамтымайтын** негативті жиын құжаттарының саны

Статистикада бір кездейсоқ шамада екіншісіне қатысты болатын ақпарат мөлшерін сипаттауға арналған екі кездейсоқ шаманың функциясын өзара ақпарат деп атайды. Біздің жағдайымызда өзара ақпарат өлшемін құжаттарды дұрыс жіктеу үшін терминнің берілген санатта болуын енгізетін ақпарат мөлшері ретінде көрсетуге болады.

$$MI(t,c) = \log_2 \frac{A*|TS|}{(A+C)*(A+B)} \qquad (18)$$

Термин тек пәндік сала құжаттарында ғана болып, басқа құжаттарда кездеспеген жағдайда өзара ақпарат өлшемі ең жоғарғы шегіне жетеді. (18) формуладан берілген өлшемнің сирек, пәндік салаға тән терминдерді таңдауға арналғандығын көруге болады. Егер екі терминнің пәндік салаға тән құжаттардағы қамтылуы бірдей (A) болса, онда есептеу барысында бөлімі кіші болатын, яғни ол кездесетін *(A+B)* құжаттарының жалпы саны кіші болатын термин өлшемінің мәні жоғары болады.

Өзара ақпарат сияқты ақпараттық пайда да t терминінің болуы пәндік саланың аталған терминді қамтитын құжаты жайлы ақпарат саны болып табылады.

$$IG(t,c) = \frac{A}{|TS|} * \log_2(\frac{A*|TS|}{(A+C)*(A+B)}) + \frac{C}{|TS|} * \log_2(\frac{C*|TS|}{(D+C)*(A+C)}) +$$
$$+ \frac{B}{|TS|} * \log_2(\frac{B*|TS|}{(A+B)*(D+B)}) + \frac{D}{|TS|} * \log_2(\frac{D*|TS|}{(D+C)*(D+B)})$$

Өзара ақпарат сияқты ақпараттық пайда да берілген термин кездесетін басқа санат құжаттары аз болған сайын жоғары болып келеді. Алайда өзара ақпарат өлшемінен өзгешелігі ақпараттық пайда симметриялы болып табылады, себебі ол берілген пәндік салаға "өте" тән терминдер үшін де, басқа салаға "өте" тән терминдер үшін де бірдей жоғары мән тағайындайды. Осы қасиеттің арқасында ақпараттық пайда өлшемі барлық пәндік сала бойынша бірқалыпты үлестірілген "стоп-сөздерді", қызметтік сөздерді тазалауды жақсы орындайды.

Хи-квадрат өлшемі – бұл статистикада үлгілер мен мәліметтердің үйлесімі жайлы гипотезаны тексеру үшін ең жиі қолданылатын өлшем. Біздің тапсырмамыз үшін оның формуласы келесідей болады:

$$CHI(t,c) = \frac{|TS|*(A*D-C*B)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \qquad (20)$$

Егер термин тек позитивті жиын құжаттарына ғана кірсе, онда өлшем таңдама құжаттарының санына тең болатын ең үлкен мәніне жетеді, ал егер термин мен санат тәуелсіз болса, онда 0-ге тең болатын ең кіші мәніне жетеді.

Кілттік сөздерді алудың жоғарыда келтірілген өлшемдерінің әрқайсысының өзіндік артықшылықтары мен кемшіліктері бар. Үш өлшем де жалпы қолданыстағы лексика сөздерін, соның ішінде стоп-сөздерден құтылуға мүмкіндік береді. Сонымен қатар, өзара ақпарат өлшемі пәндік салаға тән өте сирек кездесетін терминдерді іріктеп алуға мүмкіндік береді. Мұнан өзгешелік ретінде ақпараттық пайда мен хи-квадрат өлшемі пәндік сала құжаттарында жиі кездесетін терминдерді ("ашық" концепт) таңдап алуға мүмкіндік береді. Екі өлшем де симметриялы, олар позитивті де, негативті де жиындардың терминдерін шығарады, сондықтан терминнің сарапшыны қызықтыратын пәндік сала терминіне жататындығын не жатпайтындығын қосымша тексеру қажет болады. Тексерістің айқын әдісі A және C шамаларын

салыстыру болып табылады (егер A>C, онда термин позитивті жиынға тиесілі болады).

Төменде хи-квадрат өлшемі негізінде пәндік сала мәтіндеріндегі кілттік сөздерді алу алгоритмі көрсетілген.

1-алгоритм: Мәтіндік жинақтағы кілттік сөздерді алу
Жинақ сөздігі бойынша жүретін айналым
Жинақ сөздігінен кезекті *сөзді* шығару
A:=0; B:=0; C:=0; D:=0;
Жинақ құжаттары бойынша жүретін айналым
Жинақтағы кезекті *құжатты* шығару
Егер *құжат пәндік салаға* тиісті болса және *сөзді* қамтыса, онда A:=A+1
Егер *құжат пәндік салаға* тиісті болса және *сөзді* қамтымаса, онда B:=B+1
Егер *құжат пәндік салаға* тиісті болса және *сөзді* қамтымаса, онда C:=C+1
Егер *құжат пәндік салаға* тиісті болса және *сөзді* қамтымаса, онда D:=D+1
Айналым соңы
Chi2:=(A+B+C+D)*(A*D-B*C)^2/((A+B)*(A+C)*(B+D)*(C+D))
Егер (Chi2>6.6) ЖӘНЕ (A>C) болса, онда *сөзді кілттік сөздер тізіміне* қосу
Айналым соңы

2.3. Жеке құжаттардағы терминдерді автоматты түрде тану әдістері

Терминдерді қарама-қарсы түрде алу әдісінің кемшіліктерінің бірі міндетті түрде мәтіндердің балама жинағының қажет етілетіндігі болып табылады. Кей кезде оны таңдау бейтривиал болып кетеді, себебі бір уақытта барлық пәндік салаларға "оппозициялық" болатын бірегей әмбебап жинақ болмайды. Сондықтан заманауи зерттеушілер қарама-қарсы әдістермен бірге терминдерді шығарудың мақсатты жинақтың немесе мұндай жинақ болмаған жағдайда мақсатты

құжаттың ішкі байланыстарының талдауына негізделетін әдістерін де белсенді зерттеу үстінде.

Қарастыру мақсатында таңдалған бірінші әдісті жапондық ғалымдар Matsuo мен Ishizuka жеке құжаттардағы кілттік сөздерді алу үшін құрастырған [35]. Бұл әдіс егер сөз термин болса, онда оны жоғары жиілікті терминдер тобындағы сөздермен қатар қолданған кезде белгілі бір ішкі топқа қарай ығысу байқалатын болады деген болжамға негізделеді. Ығысу дәрежесі хи-квадрат өлшемі негізінде келесі формула бойынша бағаланады:

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w,g) - n_w p_g)^2}{n_w p_g}, \qquad (21)$$

мұндағы w – ағымдағы сөз, G – құжаттың жоғары жиілікті терминдерінің жиыны, $freq(w,g)$ – w сөзі мен жоғары жиілікті $g \in G$ терминінің шектелген мәнмәтіндерде бірге кездесу жиілігі, p_g – g терминінің құжатты кездесуінің шартсыз ықтималдығы, n_w – w сөзі мен G-дағы терминдердің құжатта бірге кездесулерінің жалпы саны.

Бақыланатын жиілік (бірге кездесу) пен күтілетін жиілік (шартсыз) арасындағы айырмашылық жоғары болған сайын нөл-гипотезаның ықтималдығы төмен келеді, яғни хи-квадрат мәндерінің жоғары болуы берілген терминнің жиіліктік терминдер маңайында кездейсоқ пайда болмағандығын білдіреді.

Сонымен, бұл әдісті іске асырған кезде алдымен жиі кездесетін терминдер шығарылады, содан кейін әр сөздің жиі кездесетін терминдермен бірге кездесуі есептеледі (сөйлем ішінде). Егер берілген терминнің жиі кездесетін терминдер маңайында кездесуі кездейсоқ болмай, ығысқан болса, онда берілген терминнің кілттік болуы әбден мүмкін. Ығысу дәрежесі хи-квадрат арқылы өлшенеді. Бұл әдіс TF-IDF өлшемімен салыстыруға болатын сапаны көрсетеді.

Жиі кездесетін терминдер олардың құжатқа кіру жиіліктерін санау арқылы шығарылады. Салыстырмалы жиіліктер алынады, яғни барлық жиіліктердің қосындысы 1 болатындай етіп қалыпқа келтіріледі. Сөздердің жиі кездесетін терминдермен қатар кездесуін есептеу үшін құжат

сөйлемдерге бөлінеді. Құжаттың, тараудың атаулары мен сурет астындағы сөздер де сөйлем деп есептеледі. Сөйлемде кездескен екі термин бірігіп кездесудің бір жағдайы ретінде қарастырылады. Егер құжаттағы әртүрлі терминдер санын N арқылы белгілейтін болсақ, онда бірге кездесу матрицасы N × N симметриялық матрица болады. Авторлар алгоритмнің жұмысын жақсарту үшін 2 әдісті пайдаланады: сөйлем ұзындықтарын қалыпқа келтіру және хи-квадрат өлшемінің тұрақтылығын арттыру.

Бірінші әдіс құжаттың ұзындықтары әртүрлі сөйлемнен тұратындығымен түсіндіріледі. Егер термин ұзын сөйлемде кездессе, онда ол көп терминдермен кездесуі мүмкін. Егер термин қысқа сөйлемде кездессе, онда басқа терминдермен бірге қолдану ықтималдығы аз болады. Сондықтан сөйлемдерді қалыпқа келтіру үшін жаңа шамалар енгізіледі:

- $Pg - g$ термині кездесетін сөйлемдердің құжаттың жалпы ұзындығына бөлінген ұзындығы;
- $nw - w$ термині кездесетін сөйлемдердің ұзындығы.

Екінші әдіс $g \in G$ жиіліктік пулындағы белгілі бір терминмен сәйкес келетін терминнің $\chi 2$ мәні жоғары болатындығына негізделеді. Алайда кей жағдайда аталған терминдер өздігінен маңызды болмай, g терминін толықтырады. Мысалы, internal терминінің жиі кездесетін state терминімен байланысы жоғары болып табылады, себебі аталған терминдер internal state тұрақты тіркесінде қолданылады. state жиі кездеспейтін термин деп болжау арқылы internal терминінің $\chi 2$ мәнін айтарлықтай азайтуға болады.

$$\chi^{2'}(w) = \chi^2(w) - max_{g \in G} \frac{\left(freq(w,g) - n_w p_g\right)^2}{n_w p_g},$$

Тұрақтылықты терминдерді кластерлеу арқылы да арттыруға болады. Бірге кездесу матрицасы бастапқыда есептеу үшін жиі кездесетін терминдерге сәйкес келетін бағандар алынатын $N \times N$ матрицасы болып табылады. Қалған бағандар ескерілмейді, яғни терминнің төменгі жиілікті терминдермен қатар кездесуі есепке алынбайды, себебі төменгі жиілікті терминдердің пайда болуының дәл

ықтималдығын бағалау қиынға соғады. Ақырғы алгоритм 16-суреттегідей болады.

Біз Matsuo & Ishizuka әдісінің баламасы ретінде терминдерді шығарудың пәндік жинақтың ішкі байланыстарының талдауына негізделетін өзіміздің екі жеке әдісімізді құрастырдық. Бұл мақсат үшін біз теріс емес матрицалық факторландыру (NMF) және Дирихленің латенттік орналастыруы негізінде (LDA) құжаттарды тақырыптық үлгілеуді пайдаландық, ол жайлы[36] жұмыста толығырақ сипатталған. Тақырыптық үлгілеу сөздер кеңістігінен тақырыптар кеңістігіне ауысуды, яғни жинақ немесе құжат ішінде әрқайсысы өзінің жеке сөздер жинағымен сипатталатын жергілікті тақырыптар жинағын ерекшелеуді білдіреді [37].

16-сурет – Жеке құжаттардағы терминдерді [28] жұмыс негізінде шығарудың қорытынды алгоритмі

3. АҚПАРАТТЫҚ ІЗДЕУ ТЕРМИНДЕРІН ҚАРАМА-ҚАРСЫ ӘДІСПЕН ШЫҒАРУ БОЙЫНША КЕЙС-СТАДИ

3.1. Бастапқы деректер

Бұл тарауда К. Маннинг, П. Рагхаван және Х. Шютценің ақпараттық іздеуге қатысты белгілі оқулығындағы терминдерді қарама-қарсы әдіспен алуға бағытталған тәжірибелік мысал (кейс-стади) қарастырылады. Терминдерді шығару нәтижелері авторлардың өздері құрастырған терминдердің эталондық көрсеткіші арқылы бағаланады.

BAWE (The British Academic Written English) деп аталатын академиялық жазбаша ағылшын тілінің британдық корпусы британдық үш университеттің біріккен жобасы ретінде құрастырылған: Уорик университеті, Рединг университеті және Оксфорд Брукс университеті. Жобаның мақсаты аталған университеттердің жоғары курс студенттері мен магистранттарының жазба жұмыстары ішіндегі ең жақсыларын таңдап, бір корпусқа жинау болған [38]. Осылайша, корпусқа 4 ғылым саласы бойынша 35 оқу пәніне қатысты 3000-ға жуық жұмыс енген: өнер және гуманитарлық ғылымдар, өмір жайлы ғылым, физикалық ғылымдар және әлеуметтік ғылымдар.

Қазіргі таңда корпусты Оксфордтық мәтіндер мұрағатынан 2539 нөмірлі ресурс ретінде жүктеп алуға болады [39]. Бұл күйінде ол 2761 құжаттан тұрады, олардың әрқайсында жұмыс коды, оның атауы, жазылу күні, жұмыс жанры, оқу пәні, алынған бағасы, сөздер саны сияқты мәліметтерді қамтитын толық түсініктемесі бар. Әр жұмыстың авторы жайлы ақпаратқа да түсініктеме берілген, атап айтатын болсақ, мұндай түсініктеме студенттің жынысы, оның туған жылы, ана тілі, елі, қай жақта дүниеге келгені және т.с.с. мәліметтерді қамтиды.

Бастапқыда корпус британдық жоғары оқу орындары студенттерінің жазбаша жұмыстарына тән тілдік ерекшеліктерді зерттеу үшін құрастырылған [40]. Атап айтатын болсақ, корпусқа жинақталған үлгілер бойынша академиялық жазбаша жұмыстардың стилі, лексикасы, жанр

түрлері, стиль мен жанрдың ғылым мен пән саласына тәуелділігі зерттелген. Уақыт өте келе корпусты тек тіл мамандары ғана емес, жазбаша ағылшын тілін зерттеуге қызығушылығы бар адамдардың барлығы кеңінен пайдаланатын болған.

BAWE корпусы табиғи тілді өңдеу саласында ашық қолданысқа шыққан күннен бастап құжаттардың мәтіндік жинағы ретінде қолданыла бастады. [41] жұмыста корпустың небәрі 500 құжаттан тұрған сынамалы нұсқасы құжат авторларының гендерлік сипатын автоматты түрде анықтауға қатысты тәжірибені жүргізу үшін пайдаланылған. Тәжірибе нәтижелері бойынша авторлардың 81%-ның жынысы дұрыс анықталған.

[42] жұмыста аталған корпус тақырыптық үлгілеуге қатысты тәжірибеде пайдаланылған. Авторлар өз әдістерін тексеру үшін BAWE корпусының өнер және гуманитарлық ғылымдар саласындағы мәтіндерін қолданған. [43] жұмыста авторлар ағылшын сөйлемдеріндегі тақырыпты автоматты түрде анықтау үшін корпус мәтіндерін пайдаланған. Авторлар құрастырған Theme Analyzer жүйесі әр сөйлемнің тақырыптық құрылымын ғана емес, сонымен қатар оның құрамындағы синтаксистік түйіндерді, тақырыптық рөлдерді және т.с.с. да анықтайтын болған.

BAWE корпусын пайдаланудағы қызықты тәжірибелердің бірі оны құжаттардың зерттеушіні қызықтыратын қандай да бір басқа жинақпен салыстыруға қажетті балама жинағы ретінде пайдалану болып табылады. [44] жұмыста авторлар берілген пәндік салаға қатысты кілттік сөздерді шығару үшін BAWE-ні салттық әрекеттер сипаттамасын қамтитын мәтіндер жинағымен бірге пайдаланады. Авторлар пәндік саладағы кілттік сөздерді олардың сала ішінде және одан тыс жерлерде әртүрлі кездесуі тұрғысынан анықтайтын қарама-қарсы әдісті пайдаланады. Пәндік сала ішінде жиі және оның шеңберінен тыс жерлерде өте сирек қолданылатын сөздер кілттік сөздер болады деп есептеледі. Берілген жағдайда "пәндік сала ішінде" тіркесі салттарды сипаттайтын мәтіндерде екендігін, "ал оның

шеңберінен тыс" деген балама жинақ мәтіндерінде, яғни BAWE корпусы мәтіндерінде екендігін білдіреді.

[45] жұмыста авторлар BAWE-ні басқа корпуспен салыстыру үшін пайдаланады, олардың айтуынша, бұл "корпус сипаттамалары мен мәтін түрлерін зерттеудің тікелей, дұрыс және қызықты әдісі" болып табылады. Бұл жұмыстың авторлары қарастырылып отырған әрбір корпустың топ-100 кілттік сөзін талдайды да, аталған тізімдерді өзара салыстырады.

Берілген кейстің мақсаты – қарама-қарсы әдісті біз BAWE-ні қарастырып отырғандай теңдестірілген және көрнекі балама жинақ болған жағдайда пайдаланудың мәтіндердің пәндік жинағындағы терминдерді автоматты түрде тану тапсырмасын тиімді шеше алатындығын көрсету. Бұл жұмыста мәтіндердің пәндік жинағы ретінде ақпараттық іздеу бойынша «Introduction to Information Retrieval» оқулығы алынады [46]. Оқулықтың электрондық нұсқасын Стэнфорд университетінің сайтынан алуға болады және ол тәжірибелерде терминдерді шығарудың дәлдігі мен толықтығын бағалау өлшемі ретінде қолданылатын терминдердің авторлық көрсеткішімен жабдықталған.

3.2. BAWE корпусының мазмұндық талдауы

Балама жинақ сапасының негізгі өлшемдері оның көрнекілігі мен теңгерімділігі болып табылады. Көрнекілік балама жинақтың мүмкіндігінше мақсатты пәндік саламен байланыспайтын көптеген пәндік салаларға қатысты көп мөлшердегі мәтіндерді қамтуы керектігін білдіреді. Теңгерімділік балама жинақтағы әртүрлі пәндік салалар мүмкіндігінше тең үлесте бейнелеуі керектігін білдіреді. Аталған өлшемдер тұрғысынан қарастыратын болсақ, BAWE корпусы жеткілікті көрнекі (124516 сөз қолданылады) және теңдестірілген (4 ғылым саласы мөлшері шамалас болатын мәтіндер арқылы бейнеленген) болып табылады. 7-кестеде BAWE мәтіндерінің ғылым салалары бойынша үлестірілуі, ал 17-суретте BAWE-нің теңгерімділігін көрсететін диаграмма келтірілген.

7-кесте

BAWE корпусы мәтіндерінің ғылым салалары бойынша үлестірілуі

№	Ғылым саласы	Ағылшын тіліндегі атауы	Мәтіндер саны
1	Өнер және гуманитарлық ғылымдар	Arts and Humanities (AH)	705
2	Өмір жайлы ғылым	Life Sciences (LS)	683
3	Физикалық ғылымдар	Physical Sciences (PS)	596
4	Әлеуметтік ғылымдар	Social Sciences (SS)	777
	ЖАЛПЫ		2761

[38] жұмыста корпус құрамына толық сипаттама беріледі, мәтіндердің мүмкін болатын барлық бөлімдер бойынша үлестірілуі жайлы статистикасы келтіріледі: оқу пәндері бойынша, жанрлар бойынша, жылдар бойынша, курстар бойынша және т.с.с. 18-суретте корпус мәтіндерінің әрбір ғылым саласындағы үлестірілу гистограммасын оқу пәндерін талдау арқылы көрсетеміз. Гистограммадан BAWE-дегі мәтіндердің көп бөлігі Инженерия пәніне (238) келетінін, содан кейін Биология (169) және үшінші орында Бизнес (146) екендігін көруге болады.

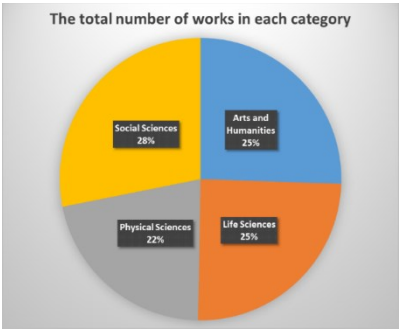
Біз сөздердің ең көрнекі үш пәндегі үлестіріміне талдау жасап, олардың бұлттарын құрастырдық (19-21-суреттерді қараңыз). Барлық сөздер саны визуалдау үшін өте үлкен болғандықтан, қолдану жиілігі 70-тен жоғары сөздер ғана пайдаланылды. Бұлттарды құрастырмас бұрын мәтіндерге алдын ала өңдеу жасалды: алдымен токендеу (мәтіндерді сөздерге және басқа да токендерге бөлу), содан кейін лемматизация (сөздерді қалыпты түрге келтіру) жүргізілді, содан кейін сандар, тыныс белгілері мен стоп-сөздер өшірілді. 8-кестеде үш оқу пәнінің әрқайсысы үшін топ-100 кілттік сөздердің жұптық қиылысулары келтірілген.

8-кесте

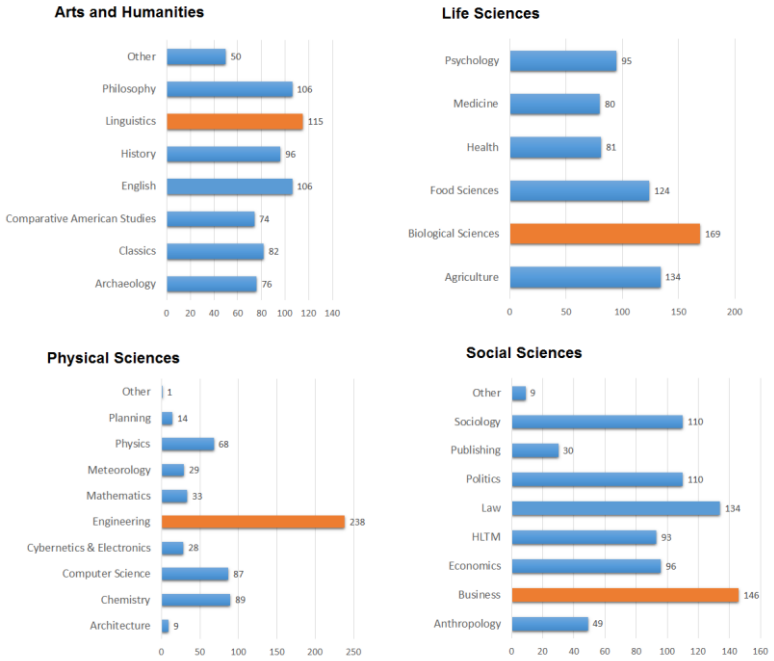
«Инженерия», «Биология» және «Бизнес» пәндерінің топ-сөздерінің қиылысулары

Инженерия және	Инженерия және Биз-	Биология және Бизнес

Биология – 26 сөз	нес – 47 сөз	– 35 сөз
activity, change, control, development, factor, figure, form, group, high, important, increase, level, need, number, order, process, product, production, quality, rate, result, role, study, system, table, time, year	analysis, based, business, case, change, company, control, cost, current, customer, development, factor, figure, financial, good, group, high, important, increase, information, level, management, market, model, need, number, order, performance, point, power, price, problem, process, product, profit, project, rate, result, service, strategy, system, table, team, term, time, work, year	area, change, control, data, development, effect, energy, experiment, factor, figure, formula, group, high, higher, important, increase, level, method, need, number, order, picture, process, product, rate, required, result, small, stage, system, table, temperature, time, type, year



17-сурет – BAWE корпусы мәтіндерінің ғылым салалары бойынша үлестірілу диаграммасы



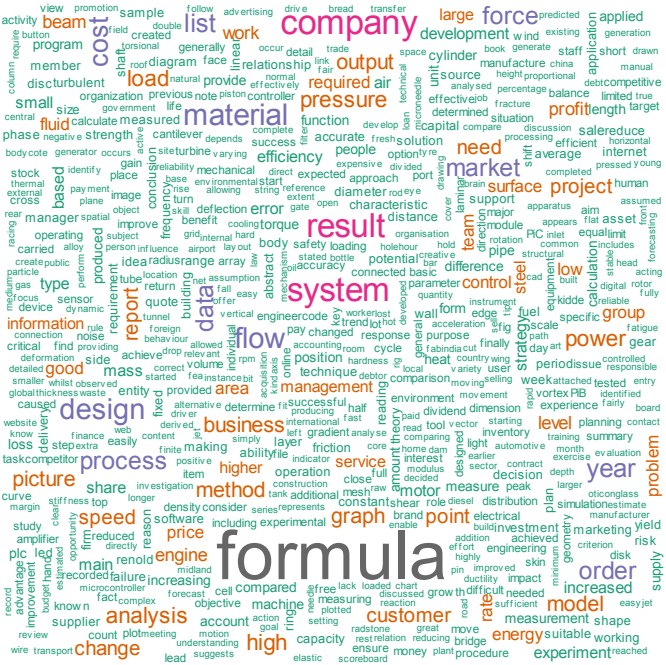
18-сурет – BAWE корпусы мәтіндерінің оқу пәндері бойынша үлестірілуі

Сонымен қатар, алынған топ-100 сөз ішінен барлық үш пәнге ортақ топ-сөздер анықталды (8-кестені қараңыз). Бұлар ғылыми мәтіндерде кең таралған нәтиже (result), жүйе (system), фактор (factor), үдеріс (process), кесте (table) және т.с.с. сөздер. Егер бұл әдісті BAWE корпусындағы барлық пәндерге тарататын болсақ, онда жалпығылыми және салааралық лексика сөздігін BAWE корпусы негізінде автоматты түрде құрастыру келешегі жайлы сөз қозғауға болады. Біз ғылыми мәтіндер корпустары негізінде автоматты немесе жартылай автоматты түрде құрастырылған жалпығылыми лексика сөздіктерін білеміз, мысалы, ағылшын тілі үшін [48] жұмысын көрсетуге болады.

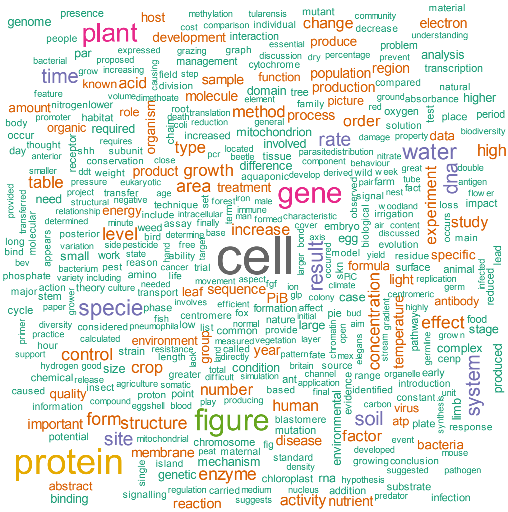
9-кесте

«Инженерия», «Биология» және «Бизнес» пәндерінің топ-сөздерінің қиылысуынан алынған жалпы академиялық топ-сөздер

	Сөз		Сөз		Сөз
	change		important	5	product
	control		increase	6	rate
	development	0	level	7	result
	factor	1	need	8	system
	figure	2	number	9	table
	group	3	order	0	time
	high	4	process	1	year



19-сурет – «Инженерия» пәні мәтіндері негізінде алынған сөздердің бұлты



20-сурет – «Биология» пәні мәтіндері негізінде алынған сөздердің бұлты

21-сурет – «Бизнес» пәні мәтіндері негізінде алынған сөздердің бұлты

3.3. Терминдерді шығару

Біз бұл жұмыста бір, екі және үш сөзді терминдерді шығарамыз, содан кейін алынған терминдер тізімін эталондық авторлық көрсеткішпен салыстыратын боламыз. Эталондық көрсеткіште 603 термин бар, оның ішінде бір сөзді 174, екі сөзді 335, үш сөзді 78, төрт сөзді 14 және алты сөзді 1 термин бар. Мысалдар 9-кестеде келтірілген.

10-кесте

«Introduction to Information Retrieval» оқулығындағы авторлық көрсеткіште кездесетін эталондық терминдер мысалы (кездейсоқ ретте алынған)

№	Бір сөзді терминдер	Екі сөзді терминдер	Үш сөзді терминдер
1	accumulator	authority score	ad hoc retrieval
2	break-even	auxiliary index	binary independence model
3	BSBI	average-link clustering	blind relevance feedback
4	lemmatization	Bayes risk	clickthrough log analysis
5	likelihood	cumulative gain	maximum likelihood estimation
6	LSA	data-centric xml	multivariate Bernoulli model
7	NLP	support vector	natural language pro-

			cessing
8	regression	term frequency	principal left eigenvector
9	regularization	term-document matrix	unigram language model
10	Reuters-21578	word segmentation	vector space model

Терминдерді шығарудың қарастырылып отырған әдістерінің $Precision$ дәлдігі мен $Recall$ толықтығын эталондық тізімнің болуы арқасында бағалауға болады. Ол үшін 10-кестеде көрсетілген мәндерді есептеу керек. Терминдерді шығарудың дәлдігі мен толықтығының алынған бағаларын орташа гармоника арқылы F-өлшем деп аталатын бір көрсеткішке біріктіруге болады.

11-кесте

Терминдерді шығару дәлдігі мен толықтығын есептеуге арналған тірек мәндер

Белгіленуі	Атауы	Қалай анықталады?
TP	True Positive (ақиқат анықталулар саны)	алынған терминдер ішіндегі эталондық тізімге кіретін терминдер саны
FP	False Positive (жалған анықталулар саны)	алынған терминдер ішіндегі эталондық тізімге кірмейтін терминдер саны
FN	False Negative (жалған өткізулер саны)	алынған терминдер қатарына кірмейтін эталондық тізім терминдерінің саны

3.4. Тәжірибелік жұмыс

Терминдерді шығаруға қатысты тәжірибелік жұмыстар R-де tm және quanteda кітапханаларын қолдану арқылы орындалды [49]. Екі жинақ та (кітап тараулары және BAWE корпусының мәтіндері) R-ге жүктелді, содан кейін өңдеуге ыңғайлы түрге түрлендірілді (құжаттар-терминдерге жеңілдетілген матрицасы түрінде көрсетіледі).

«Introduction to Information Retrieval» кітабы тарауларының негізінде құрастырылған жинаққа Стэнфорд университетінің сайтынан парсинг жасалды, аталған жинақ ол жерге ашық html-парақша түрінде орналастырылған. Парсинг барысында html-белгілеу өшірілді және кітаптың мазмұны тараулар саны мен кітап параграфтары бойынша 245 мәтіндік файлға экспортталды. BAWE корпусы мәтіндері негізінде құралған жинақ Оксфорд мәтіндер мұрағатының сайтынан жүктелді,

аталған жинақ ол жерге ашық мәтіндік файлдар мұрағаты түрінде орналастырылған.

Әрбір файлдың мәтіні NLTK пакетінің құрамына кіретін Wordnet Lemmatizer құралы көмегімен лемматизацияланды, NLTK – табиғи тілді символдық және статистикалық өңдеуге арналған бағдарламалардың ашық кітапханасы. Терминдерді шығару кезінде пос-таггинг қолданылған жоқ, сәйкесінше екі және үш сөзді терминдерге тән лексикалық үлгі бойынша іздеу де пайдаланылған жок. Әрине, бұл терминдерді шығару дәлдігін айтарлықтай төмендетті, себебі екі сөзді терминдерге негізінен A+N түріндегі лексикалық үлгілер тән (сын есім + зат есім), ал біз барлық екі сөзді үйлесімдерді таңдадық (биграммалар). Үш сөзді үйлесімдер де солай (триграммалар) болды. 11-кестеде BAWE-нің барлық төрт тарауын балама жинақ ретінде пайдаланған кездегі TF-DCF өлшемі көмегімен шығарылған бір және екі сөзді терминдердің алғашқы 30-ы келтірілген.

Автор нұсқасы бойынша алынған бірліктердің барлығы термин емес. Мысалы, авторлық эталондық тізімде machine learning (тіпті, біздің ойымызша бұл термин болса да) немесе set document (әрине, бұл термин емес) сияқты биграммалар жоқ. Бұл мысалдардың терминдерді шығару мәселесінің өте күрделі екендігін көрсететіні сөзсіз.

12-кесте

BAWE-нің барлық 4 тарауын балама жинақ ретінде пайдаланған кездегі TF-DCF өлшемі көмегімен шығарылған топ-30 бір және екі сөзді терминдер

Ранг	Термин	Ранг	Термин	Ранг	Термин
1	postings list	11	relevant document	21	machine learning
2	query term	12	term document	22	term frequency
3	information retrieval	13	language model	23	IDF
4	text classification	14	crawler	24	number document
5	web search	15	nonrelevant	25	Rocchio
6	document collection	16	multinomial	26	document query
7	relevance feed-	17	single-link	27	complete-link

	back				
8	training set	18	Reuters-RCV1	28	set document
9	KNN	19	SVM	29	IR system
10	inverted index	20	linear classifier	30	centroid

12-кестеде хи-квадрат өлшемі көмегімен шығарылған топ-30 бір және екі сөзді терминдерді келтіреміз. Бір қарағанда 6- және 7-кестедегі терминдер тізімі арасында аса үлкен айырмашылық жоқ сияқты көрінсе де, шын мәнінде оларды мұқият салыстыру арқылы хи-квадрат өлшемінің тиімділігі төмен екендігін байқауға болады. Мысалы, топ-30 сөздер қатарына термин емес, тек ақпараттық іздеу саласына ғана қатысты algorithm, compute, vector, Boolean сияқты сөздер енгізілген, бұл терминдер математика, компьютерлік ғылымдар, инженерия сияқты салаларда кең таралған.

13 -кесте

BAWE-нің барлық 4 тарауын балама жинақ ретінде пайдаланған кездегі хи-квадрат өлшемі көмегімен шығарылған топ-30 бір және екі сөзді терминдер

Ранг	Термин	Ранг	Термин	Ранг	Термин
1	query	11	inverted index	21	IR system
2	retrieval	12	postings list	22	term occur
3	query term	13	vector	23	centroid
4	information re-trieval	14	Boolean	24	naive
5	algorithm	15	document col-lection	25	IDF
6	posting	16	classifier	26	relevance feed-back
7	compute	17	text classifica-tion	27	relevant docu-ment
8	vector space	18	retrieval system	28	naive Bayes
9	search engine	19	term frequency	29	machine learn-ing
10	web search	20	IR	30	nonrelevant

Сонымен, дәл күтілгендей, TF-DCF әдісін қолданған кезде дәлдік, толықтық және F-өлшем көрсеткіштері хи-квадрат өлшемін қолданумен салыстырғанда жоғары болды (13-14-

кестені қараңыз). Сондықтан ақпараттық іздеу саласындағы онтологияны құрастыру кезінде TF-DCF өлшемі көмегімен таңдалған терминдер концепт құрудағы негізгі құрылымдар ретінде пайдаланылды. 22-суретте TF-DCF өлшеміне сәйкес алынған онтологиялар концептінің бұлты көрсетілген. Алынған терминдерден пәндік сала концептерінің қалай құрылғандығын сипаттау берілген жұмыс шеңберінен тыс екендігін айта кету керек.

14-сурет

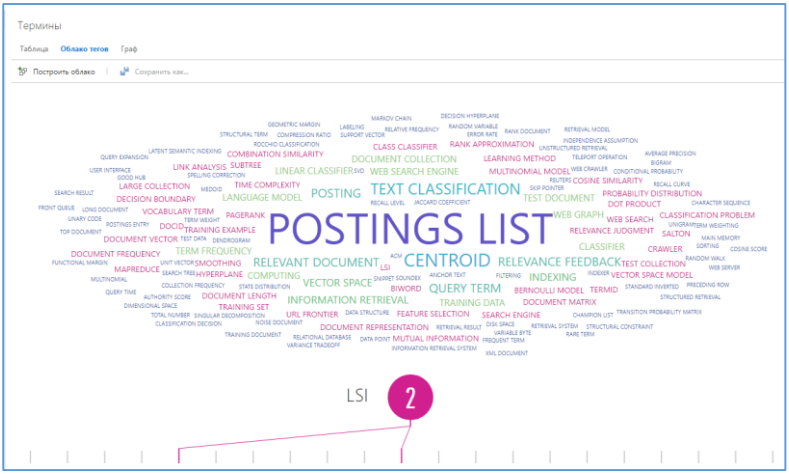
Computer Science пәнін қоспағанда төрт балама жинақты (AH+ LS+SS+PS) қолданған кездегі терминдерді шығару сапасының максималды көрсеткіштері

Көрсеткіш	Бір сөзді терминдер		Екі сөзді терминдер		Үш сөзді терминдер	
	Хи-квадрат	TF-DCF	Хи-квадрат	TF-DCF	Хи-квадрат	TF-DCF
Дәлдік	0,1818	0,24	0,1654	0,2018	0,1443	0,1271
Толықтық	0,3086	0,24	0,1284	0,2627	0,1489	0,2447
F-өлшем	**0,2288**	**0,24**	**0,1446**	**0,2283**	**0,1466**	**0,1673**

15-кесте

Computer Science пәнін қоспағанда төрт балама жинақты (AH+ LS+SS+PS) қолданған кездегі бір, екі және үш сөзді терминдерді шығару сапасының көрсеткіштері

Көрсеткіш	Хи-квадрат (шегі 24)	TF-DCF (шегі 5,5)
Дәлдік	0,1045	0,2196
Толықтық	0,2913	0,2470
F-өлшем	**0,1539**	**0,2325**



22-сурет – "Ақпараттық іздеу" пәндік саласы концептерінің бұлты

TF-DCF өлшемін қолдану кезінде терминдерді шығару сапасының көрсеткіштерін балама жинақтар санына байланысты салыстыру жақсы нәтиже көрсетті (16-кестені қараңыз). Балама пәндік салалар саны ең көп 4 болған жағдайда сапа көрсеткіштері максималды мән көрсетті. Балама пәндік салалардың мақсатты пәндік саламен қиылыспау керектігінің маңызды екендігін айта кету керек. Сонымен, балама жинақтан компьютерлік ғылымдарға (яғни ақпараттық іздеу саласына ұқсас пәндік сала) тән мәтіндерді алып тастаған кезде терминдерді шығару дәлдігі де, толықтығы да артты.

Тағы бір есте қаларлық дерек – балама мәтіндер санын арттырған кезде терминдерді шығару толықтығының төмендейтіні болып табылады. Мысалы, 3 жинақты қолданған кезде терминдер қатарына index, stemming, entropy сияқты балама жинақтарда жоқ сөздер кірген, ал 4 жинақты қолданған кезде аталған терминдер әрі мақсатты, әрі балама жинақтарда бірдей үлестірілген сөздер қатарына енген. Сәйкесінше, аталған сөздерді терминдер тізімінен шығарған кезде терминологияны қамту толықтығы төмендеген. Жалпы алғанда, үш және екі сөзді терминдерді шығару сапасы бір

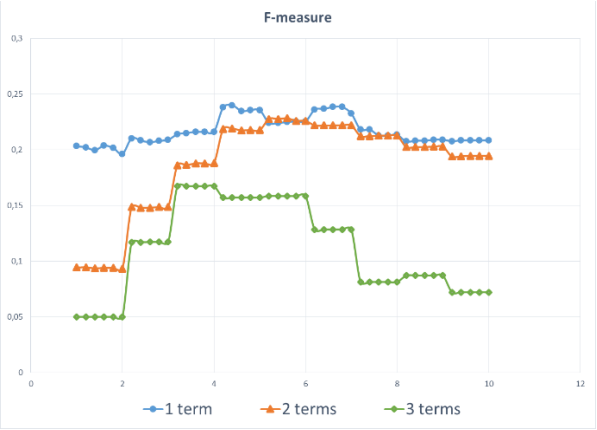
сөзді терминдермен салыстырғанда төмен болады, себебі жоғарыда атап өткеніміздей, лексикалық үлгілер қолданылған жоқ.

16-кесте

Терминдерді TF-DCF өлшемі көмегімен шығару кезіндегі дәлдік, толықтық пен F-өлшемнің балама жинақтар санына тәуелділігі (шегі 5.5)

Көрсеткіш	2 балама жинақ (LS+SS)	3 балама жинақ (АН + LS + SS)	4 балама жинақ (АН + LS + SS + PS)	Computer sciences қоспағанда 4 балама жинақ (АН + LS + SS + PS)
Дәлдік	0,2110	0,2156	0,2218	0,2196
Толықтық	0,2675	0,2623	0,2419	0,2470
F-өлшем	**0,2359**	**0,2367**	**0,2314**	**0,2325**

Хи-квадрат белгісі үшін де, TF-DCF өлшемі үшін де шектік мән таңдау мәселесі үлкен қызығушылық тудырады. 14-суретте бір, екі және үш сөзді терминдерді шығару кезінде TF-DCF өлшемінің шектік мәндерін өзгерткен кезде F-өлшем мәндерінің қалай өзгеретіндігі көрсетілген. Бұл нәтижелерге сәйкес, F-өлшемнің ең үлкен мәнін беретін тиімді шектік мән 3-тен 6-ға дейінгі аралықта жатады (4.4 – бір сөзді терминдер үшін; 5.6 – екі сөзді терминдер үшін; 3.2 – үш сөзді терминдер үшін).



23-сурет – F-өлшемнің TF-DCF өлшемінің шектік мәніне тәуелділігі

Жүргізілген тәжірибелер келесідей маңызды 3 қорытынды алуға мүмкіндік берді:

1) Балама жинақтар саны артқан кезде терминдерді шығару дәлдігі мен толықтығы өседі, сонымен қатар балама жинақтардағы терминдердің үлестірілу жиынтығын емес, олардың әрбір жеке жинақтағы жеке жиіліктерін ескеретін өлшемдерді пайдалану маңызды болып табылады;

2) Балама жинақтар қарастырылып отырған мақсатты жинаққа жақын болмауы керек;

3) Академиялық жазбаша ағылшын тілінің британдық корпусы жоғарыда келтірілген талаптарды толықтай қанағаттандырады және өзінің салыстырмалы түрде кіші көлеміне қарамастан балама жинақтар жиынтығы ретінде тиімді пайдаланыла алады.

Тәжірибе жүргізу барысында 1.3-тарауда сипатталған лексикалық-синтаксистік үлгілердің пайдаланылмағандығын айта кету керек. Терминдерді шығарудағы дәлдік мен толықтық көрсеткіштерінің аса жоғары болмауын осы арқылы түсіндіруге болады.

Қорытынды ретінде кілттік сөздер мен терминдерді тақырыптық үлгілеуде қалай қолдануға болатындығын көрсетеміз. Кіріс оқу материалдары ретінде ағылшын тіліндегі

3 дереккөз алынды. Бірінші дереккөз – Горана Целебич пен Дарио Рендуличтің "Ақпараттық-коммуникациялық технологиялардың негізгі түсініктері" атты электрондық оқулығы. Оқулық 8 тараудан тұрады, олардың әрқайсысы оқу бағдарламасының типтік мазмұнына толықтай сәйкес келеді. Екінші дереккөз – "Ақпараттық және коммуникациялық технологиялар" пәні бойынша С.Аманжолов атындағы ШҚМУ компьютерлік модельдеу және ақпараттық технологиялар кафедрасының оқытушылары құрастырған және ЖОО-ның әдістемелік кеңесі бекіткен дәрістер тезисінің жинағы. Үшінші дереккөз Saylor Academy интернет-алаңқайындағы компьютерлік ғылымдар негізі бойынша «CS301: Computer Architecture» атты ашық және тегін онлайн-курс материалдарынан тұрады.

R-де тақырыптық үлгілеуді орындау үшін topicmodels пакетіндегі LDA() функциясы қолданылады. Функция параметр ретінде құжаттар-терминдерде матрицасын, бөлінетін жасырын k тақырыптар санын және шындыққа ұқсастықты бағалау әдісін қабылдайды:

res <- LDA (dfm, k, method = "VEM")

Функция екі негізгі слотты қайтарады, бірінші слот сөздердің тақырыптар бойынша үлестірілуін, ал екінші слот құжаттардың тақырыптар бойынша үлестірілуін көрсетеді. 24-суретте екінші слотты кесте түрінде бейнелеуден үзінді келтірілген, яғни корпустың әр құжатындағы бөлінген әр 7 тақырыптың салмақтары көрсетілген. Екінші тақырыптың салмақтары бойынша сұрыптау жасалған және аталған тақырыптың аппараттық жабдықтаумен байланысты екені анық көрінеді. Мысалы, 1-дереккөздің 8-ші тарауы болып табылатын «1.8. Legal regulations» (Құқықтық нормалар) құжаты аппараттық жабдықтама тақырыбымен тығыз байланысты, бұл жерден шығатыны – ол АКТ саласында емес, бағдарламалық жабдықтама саласында емес, дәл аппараттық жабдықтама саласындағы құқықтық реттеу мәнін ашып көрсететіндігінде болып табылады. Бұл тұжырым «2.1. Computer Systems» құжаты үшін де орындалады.

Тақырыптық үлгілеу нәтижелерін көрсету үшін құрастырылған интерактивті интерфейс қолданушыға (сарапшыға) дереккөздер мен оның тарауларының тақырыптық бейнелеуін тереңірек түсінуге, сонымен қатар әртүрлі дереккөздер мен тақырыптар арасындағы терең пәндік байланыстарды анықтауға мүмкіндік береді.

Document clustering | Document-Topic Matrix

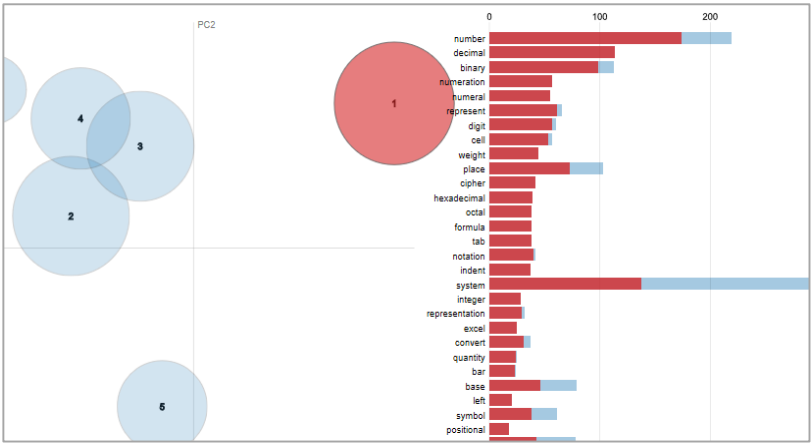
Show 10 entries Search:

Name.of.documents	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7
3.1 History of computing hardware.txt	1e-05	0.99997	0.00001	0.00001	0.00001	0.00001	0.00001
1.8 Legal regulations.txt	1e-05	0.99996	0.00001	0.00001	0.00001	0.00001	0.00001
1.1 Hardware.txt	1e-05	0.99994	0.00001	0.00001	0.00001	0.00001	0.00001
3.4 Hardware and Machine Organization.txt	6e-05	0.75693	0.24279	0.00006	0.00006	0.00006	0.00006
2.1 Computer systems.txt	5e-05	0.55091	0.00005	0.29482	0.15405	0.00005	0.00005
3.5 Parallel and Vector Architectures.txt	9e-05	0.03606	0.00009	0.00009	0.00009	0.00009	0.96350
1.2 Software.txt	5e-05	0.03005	0.00005	0.88599	0.00005	0.08377	0.00005
1.6 Environmental impact.txt	9e-05	0.00009	0.00009	0.99946	0.00009	0.00009	0.00009
1.5 Influence on health ergonomics.txt	7e-05	0.00007	0.00007	0.00007	0.00007	0.99957	0.00007
2.6 Data analysis.txt	6e-05	0.00006	0.00006	0.00006	0.00006	0.00006	0.99965
Name.of.documents	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7

24-сурет –LDA кестелік бейнелеуі

LDA тақырыптық үлгісін визуалдаудың ыңғайлы әдісі LDAvis. интерактивті құралы болып табылады, оң жақтағы шеңберді таңдаған кезде тақырыпқа сәйкес келетін кілттік сөздер көрінеді (25-суретті қараңыз). Тізім ұзындығын реттеуге болады. 25-суретте көрсетілгендей, визуалдау екі негізгі бөліктен тұрады. Визуалдаудың сол жақ панелі тақырыптар мен олардың байланысын көрсетуге жауап береді. Бұл визуалдаудағы әрбір тақырып нөмірленген шеңбер болып табылады, шеңбердің көлемі жинақтағы тақырыптың салмағымен анықталады. Ұқсас тақырыптар бір-біріне жақын көрсетіледі, кейбіреулері тіпті қиылысады. Визуалдаудың оң жақ панелі таңдалған тақырыпты түсіндіру үшін барынша

келетін терминдерді көрсетеді. Бұл сөздер әр тақырыптың мағынасын ашады.



25-сурет – Тақырыпты визуалдауға арналған LDAvis құралы

Кілттік сөздер үстіндегі жылжытпа қолданушыларға тақырыптың кілттік сөздерінің сарапануын өзгерте алатын λ параметрінің мәнін ауыстыруға мүмкіндік береді. Үнсіздік бойынша λ параметрі үшін 0.6 мәні тағайындалған. Егер параметр 1-ге тең болса, онда тақырыптың басқа терминдерде мүлдем кездеспейді деп айтуға болатын ерекше терминдері жоғары көтеріледі. Параметр төмен болған сайын тақырыптың жалпылай қолданылатын терминдерінің рангі жоғары болады. Терминнің қасындағы қызыл және көк сызықтар оның сәйкесінше берілген тақырыптағы және басқа тақырып мәтіндеріндегі үлесін көрсетеді. Мысалы, көрсетілген суретте кілттік сөздер панеліндегі "сан", "жүйе", "ондық", "екілік", "сегіздік", "нөмір", "цифр", "бейнелеу", "цифр", "он алтылық" сөздерінің рангтері жоғары. Бұл терминдердің тіркесіп келуі "Санау жүйесі" тақырыбы жайында айтылып жатқандығын анықтауға мүмкіндік береді, ал әр терминнің қасындағы қызыл және көк түстердің үлесі аталған терминнің берілген тақырыпқа қаншалықты қатысты екендігін көрсетеді. Мысалы, "жүйе" сөзінің қасындағы көк түс үлесінің жоғарылығы бұл терминнің әмбебаптылығын, ал "ондық" сөзі қасындағы қызыл

түстің жоғары үлесі оның қолдану аясының тар екендігін білдіреді.

Сонымен, терминдерді шығару мен тақырыптық үлгілеудің құрастырылған үлгілері қолданушыға тек пәндік саланы сипаттап қана қоймай, сонымен қатар оны визуалдауға, пәндік саланың ішкі тараулары арасындағы өзара байланысты көрсетуге мүмкіндік береді. Бұл зерттеуге, талдауға, сәйкес келетін әдебиеттерді таңдауға кететін уақытты үнемдейді, сонымен қатар оқу курсына барынша келетін контентті анықтай алады.

4. ДЕРЕКТІ МӘНДЕРДІ ШЫҒАРУ

4.1. Бастапқы деректер

Қазіргі таңда ақпаратты алуға арналған әртүрлі әдістер бар. Олар алуан түрлі болып келеді, бірі екіншісінен артық деп айту қиын, себебі әртүрлі жағдайларда екеуі де жақсы нәтиже көрсетеді. Ақпаратты алуға арналған әдістерді келесі санаттарға жіктеуге болады:

– ережелерге негізделетін әдістер. Сарапшылар белгілі бір мәліметтерді шығаруға қажетті ережелерді қолдан құрастырады.

– білімге негізделетін әдістер. Мұнда онтологияларға [50], тезаурустарға негізделген үлгілер [51] жатады.

– статистикалық әдістер. Олардың қатарына жасырын марковтық үлгілер [52-54], шартты марковтық үлгілер [55], шартты кездейсоқ өрістер [56] жатады.

– машиналық оқытуға негізделетін әдістер.

Ақпарат шығарудың ішкі тапсырмаларының бірі деректі мәндерді тану болып табылады. Мәтіннің ішінен адам аттары, ұйым, жер, геосаяси нысандардың атаулары, уақыт белгілері сияқты деректі нысандар таңдалады, кеңейтілген нұсқалар медициналық, биологиялық және т.б. сияқты белгілі бір пәндік салаға тән терминдерді қамтуы мүмкін. Деректі мәндерді танудың қазіргі таңдағы қолданыстағы әдістерін екі санатқа бөлуге болады:

– Ережелерге негізделген. Олар деректі мәндерді танудағы алғашқы жүйелердің бірі болып табылады. Ережелер белгілі бір тілге тән лексикалық-синтаксистік үлгілерге негізделеді. Сондықтан мұндай әдіске негізделетін жүйелер тиімдірек болып саналады [57]. Алайда бұл жүйелер өздері анықталған сала үшін шектеулі болады және тасымалданбайды [58, 59].

– Оқытушымен үйрену. Бұл әдіс сарапшылар қолмен белгілейтін жаттықтырушы мәліметтердің көп болуын талап етеді. Содан кейін жүйе ұсынылған мәліметтер негізінде деректі мәндерді тануға арналған ережелер шығарады. Бұл

санатқа шартты кездейсоқ өрістер (Conditional Random Fields) [60], максималды энтропия (Maximum Entropy) [61], шешімдер ағаштары (Decision trees) [62] және т.б. әдістер жатады.

– Оқытушысыз үйрену. Жүйе деректі есімдер үлгілерінің кішігірім жинағын пайдаланады, мысалы, елдер {«Канада», «Оңтүстік Корея», «Жапония», …}. Жүйе берілген жинақты зерттейді және берілген жинақтағы нысандар кездесетін сөйлемдер негізінде деректі мәндерді шығаруға арналған біршама ережелер құрастырады. Бұл ережелер жаңа мәндерді анықтау үшін қолданылады. Содан кейін ережелердің жаңа жинағы зерттеледі. Осылай жаңа ережелер табылғанға дейін оқыту жалғаса береді. Мысалы, [63] жұмыста деректі мәндерді мәліметтердің маркерленбеген мысалдарын пайдалану арқылы жіктеудің бақыланбайтын үлгісі талқыланады. [64] жұмыста деректі мәндердің бақыланбайтын жіктеуі және деректі мәндердің кішігірім сөздігі мен деректі мәндерге арналған маркерленбеген корпус пайдаланылатын ансамбль техникасы ұсынылады.

– Біріккен жүйелер. Екі немесе одан да көп машиналық оқыту немесе ережелер негізіндегі оқыту техникаларынан тұрады [65, 66].

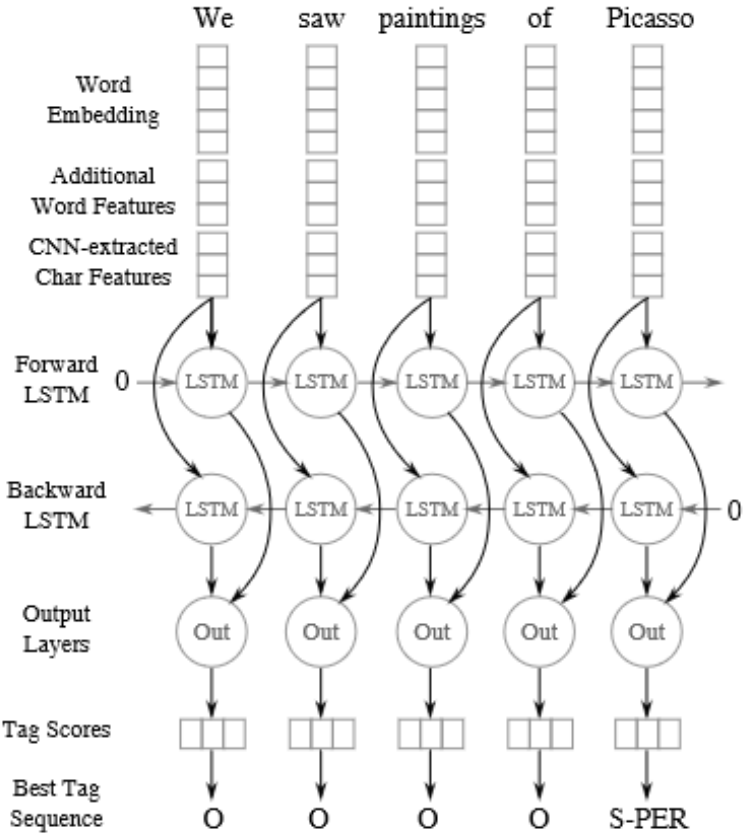
4.2. Алдыңғы жұмыстар

Деректі мәндерді шығарудағы дәстүрлі жүйелер қолмен анықталған қасиеттерді пайдаланады [67]. Кейбір бұрынғы жұмыстарда қолмен құрастырылған ережелер қолданылған [68, 69], алайда заманауи жүйелердің көп бөлігі шартты кездейсоқ өріс (CRF) [71], Жасырын Марковтық үлгі (HMM) [72], тірек векторлар әдісі (SVM) [73] сияқты машиналық оқыту үлгілеріне негізделеді. Машиналық оқытудың дәстүрлі әдістері қолмен құрастырылған ережелерге негізделмесе де, функцияларды қолмен құрастыруды талап етеді, бұл өте қымбат тұрады және домен мен тілге тәуелді болады. Соңғы уақытта нейрондық желілерді қолданылған көптеген жұмыстар дәстүрлі жүйелерден асып түсуде. Соңғы жылдары Long-Short-Term-Memory (LSTM) [74], Gated Recurrent Unit (GRU) [75] сияқты рекуррентті нейрондық желісі бар үлгілер

тізбектерді үлгілеу есептерінде сәтті қолданылып келеді, мысалы Language Modeling [76, 77], машиналық аударма [78], Dialog Act classification [79, 80]. RNN үлгілерінің мықты тұстарының бірі олардың мәтіннің негізгі құрауыштарында оқытыла алу қабілеті болып табылады (яғни сөздер мен символдарда). Мұндай жалпылау мүмкіндігі қасиеттерді бақыланбай зерттеу мен түсініктеме берілген кішігірім корпусқа негізделетін, тілден тәуелсіз NER үлгілерін құруды жеңілдетеді [81, 82].

Нейрондық үлгілерді тізбектерді таңбалау тапсырмасында қолдану алғаш рет Collobert et. al. [83] жұмысында ұсынылды. Алайда бұл үлгіге қойылатын кейбір шектеулер бар. Біріншіден, мұнда тік байланысы бар қарапайым нейрондық желі қолданылады, ал бұл сөздер маңындағы қарастырылып отырған мәнмәтіннің ауқымын шектейді. Үлгі үлкен қашықтықтағы сөздер арасындағы пайдалы қатынастарды ұмыт қалдырады. Екіншіден, сөздерді векторлауға тәуелділік салдарынан жұрнақтар мен сөз алды қосымшалары сияқты символдар деңгейінде көрсетілген қасиеттерді анықтау және пайдалану мүмкін болмайды.

Кейінірек екі бағытты LSTM немесе Stacked LSTM қолданатын түрлендірілген үлгілер ұсынылды [84, 85]. Мысалы, [84] жұмыста bi-LSTM және CRF негізіндегі архитектура қолданылады. [86] жұмыстың авторлары bi-LSTM-CNNs архитектурасын пайдаланады (26-сурет). Олар символдарды векторлау үшін оралған нейрондық желілерді қолдануды ұсынады. Енгізілген символдар ішінен ішкі сөз ақпараттарын шығару үшін CNN немесе LSTM пайдаланатын жаңа әдістер табылды, олардың нәтижелері басқа үлгілерден жақсы екендігін айта кету керек [87]. Rei et. al. [88] кіріс мәліметтері ретінде сөздер мен символдар берілетін үлгіні ұсынды.



26-сурет – Jason P.C. Chiu және E. Nichols [86] ұсынған үлгі архитектурасы

Kuru et. al. [89] символдар негізіндегі нейрондық үлгіні ұсынады. Кіріске тек қана символдарды қабылдайтын бұл үлгі ешқандай сыртқы мәліметтер қолданылмаған жағдайда жақсы көрсеткіштер көрсетеді. Аталған үлгі әр символға тегті алдын ала хабарлайды және сөздегі барлық символдардың алдын ала айтылған тегтерінің бірдей болуын тексереді.

Берілген жұмыста сөздер де, символдар да векторлық түрде көрсетілетін және мәнмәтіндік ақпаратты үлгілеу үшін екі бағытты LSTM блогының кірісіне берілетін нейрондық

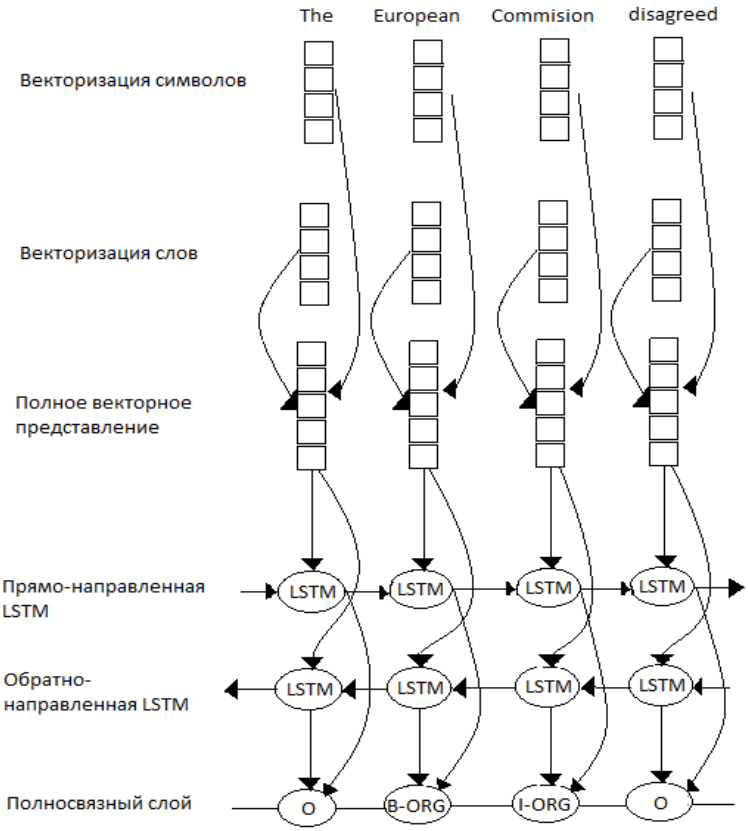
үлгі қарастырылады. Ақпаратты символдар деңгейінде кодтау үшін де екі бағытты LSTM қолданылады.

4.3. Үлгі

Деректі мәндерді шығару тапсырмасын шешу үшін символдар мен сөздерді векторлайтын bi-LSTM блокқа негізделген үлгі құрылды (27-сурет). Үлгіні құрастыруда [90] жұмыста ұсынылған әдіс қолданылды. Жұмыс авторлары нейрондық желінің сөздер мен символдар деңгейіндегі функцияны bi-LSTM және CNN біріккен екі бағытты архитектурасын қолдану арқылы автоматты түрде анықтайтын жаңа архитектурасын ұсынады.

4.3.1. LSTM

LSTM (Long Short-Term Memory) – рекурренттік нейрондық желілердің бір түрі. Рекурренттік нейрондық желілер алдыңғы қадамдардың нәтижелерін есте сақтау қабілетіне ие, бірақ оларды ұзақ есте сақтай алмайды. Градиенттің жоғалып кету мәселесі туындайды [91]. [t] LSTM желілері аталған мәселенің алдын алу үшін құрастырылған. Олар қай ақпараттың ұмытылатынын, қай ақпараттың келесі қадамдарға жіберілетінін бақылап отыратын негізгі үш бөліктен тұрады. LSTM блокты сұлба түрінде 28-суреттегідей етіп көрсетуге болады.



Символдарды векторлау, Сөздерді векторлау, Толық векторлық бейнелеу, Тура бағытталған LSTM Кері бағытталған LSTM, Байланысы толық қабат

27-сурет - Желінің негізгі архитектурасы

t уақыт мезетінде LSTM блокты жаңартуға арналған формуланы ресми түрде келесідей көрсетуге болады:

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$
$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$
$$\tilde{c}_t = tanh(W_c h_{t-1} + U_c x_t + b_c)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$

мұндағы

x_t – t уақыт мезетіндегі кіріс векторы,

h_t – t уақыт мезетіндегі жасырын күй (шығыс) векторы,

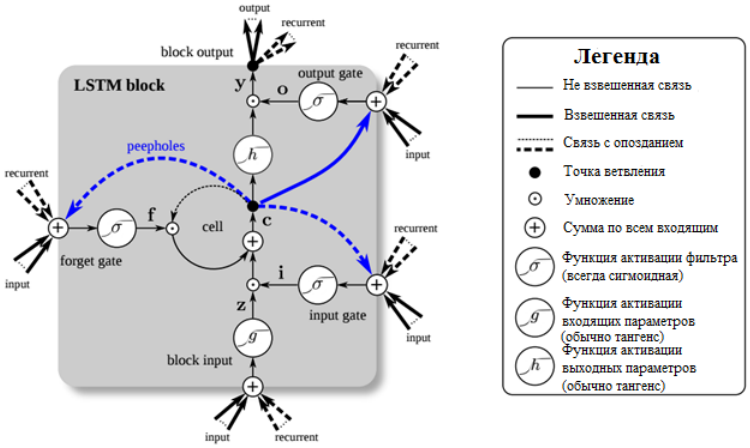
σ- сигмоидтық функция,

Ui, Uf, Uc, Uo – x_t кіріс векторына арналған әртүрлі сүзгілер салмақтарының матрицасы,

Wi,Wf,Wc,Wo – h_t жасырын күй векторына арналған салмақтар матрицасы,

$\odot$ - элемент бойынша көбейту,

b_i, b_f, b_c, b_o – ығысу векторлары.



28-сурет - LSTM блок [92]

4.3.2. Bi-LSTM

bi-LSTM-ның негізгі идеясы тек алдыңғы ғана емес, сонымен қатар кейінгі символдар тізбегін де ескеру болып табылады [93]. Осылайша оның айналасы толық қарастырылады. Bi-LSTM алдымен тізбекті және кейін қарастырылып отырғанға кері бағытталған тізбекті

қарастыратын тура бағытталған LSTM-нен тұрады. Содан кейін алынған тізбектер конкатенацияланады.

4.3.3. Векторлау

Векторлық бейнелеу немесе векторлаудың мақсаты мәтіндік корпустан ақпарат шығару және оның әрбір элементіне (сөз/символ) бірегей сандық векторды сәйкестендіру болып табылады. Векторлау – тілді үлгілеу және табиғи тілді өңдеудегі белгілі бір сөздіктегі сөздерге сөздіктегі сөздердің айтарлықтай аз бөлігін сәйкестендіруге бағытталған бейнелеулерді оқыту әдістерінің бірі.

Векторлық бейнелеулердің теориялық негізі табиғи тілді өңдеудің мәтіндегі сөздердің үлестірімін (дистрибуция) бағалау арқылы тіл бірліктері (сөздер, түсініктер, құжаттар) арасындағы семантикалық жақындықты зерттеуге арналған әдістер тобы болатын дистибутивтік әдіс болып табылады. Дистибутивтік талдаудың негізгі құралдары мәнмәтіндік векторлар және бірге кездесу матрицасы болады [94].

Сөздің мәнмәтіндік векторы ретінде берілген сөз бір мәнмәтін ішінде кездесетін сөздерді көрсететін вектор түсіндіріледі [95]. Құжаттың мәнмәтіндік векторы деп берілген құжатта кездесетін сөздерді көрсететін вектор аталады. Олай болса, екі сөз немесе құжат арасындағы семантикалық қашықтық оларға сәйкес келетін мәнмәтіндік векторлар арасындағы евклиптік қашықтық немесе косинустық қашықтық ретінде анықталады.

Бірге кездесу матрицасы ретінде жолдары мен бағандары тіл бірліктері болатын және жолдар мен бағандардың қиылысатын тұсына жалпы мәнмәтіндегі тіл бірліктерінің бірге қолданылу немесе тиістілік көрсеткіштері жазылатын матрица түсіндіріледі. Мысалы, терминдер-терминдерге матрицасын терминдердің бір құжатта немесе екінші құжатта бірге кездесуі негізінде құрылуы мүмкін. Ол бинарлы болуы, яғни нөлдер мен бірліктерден тұруы мүмкін, егер екі термин бір құжатқа немесе бір сөйлемге кірсе, 1, кері жағдайда 0 көрсетіледі. Сондай-ақ, мұндай матрица жиіліктік болуы да мүмкін, яғни екі термин бірге кездесетін құжаттар немесе

сөйлемдер санын көрсетеді. Әдеттегі бірге кездесу матрицасы элементтері терминдердің пәндік жинақ құжаттарына кіру жиіліктері (салыстырмалы немесе абсолютті) болатын матрица болады.

Мысалы, терминдер арасындағы семантикалық байланыстардың күштерін бағалау үшін "құжаттар-терминдерге" дистрибутивтік матрицасын қалай пайдалануға болатындығын көрсетейік. Матрицадағы әрбір термин вектор-баған болады, осылайша, кез келген екі термин арасындағы семантикалық байланысты сәйкес векторлар арасындағы жақындық немесе қашықтық ретінде қарастыруға болады, мұндай жағдайда векторлы кеңістіктегі кез келген белгілі өлшемдерді пайдалануға болады. Мысалы, жоғарыда айтылғандай, косинустық өлшемді пайдалансақ:

$$r_{ij} = cos(\bar{T}_i, \bar{T}_j) = \frac{\bar{T}_i \cdot \bar{T}_j}{|\bar{T}_i| \cdot |\bar{T}_j|},$$

мұндағы $\bar{T}_i, \bar{T}_j$ – бұл "құжаттар-терминдерге" матрицасының сәйкесінше *i*-шы и *j*-шы терминдерге сай келетін вектор-бағандары (i мен j терминдердің толық тізімінен өтеді), r_{ij} – бұл жақындық мәні, семантикалық байланыстар матрицасының элементі.

Дистрибутивтік матрица мен мәнмәтіндік векторлардың табиғи-тілдік мәтіндерді семантикалық талдау есебіндегі алатын орнын асыра бағалау қиын ([96, 97, 98]). Табиғи тілді өңдеу және ақпараттық іздеудің дистрибутивтік әдістерге негізделетін белгілі үлгілері мен тәсілдері қатарына келесілерді жатқызуға болады:

- Bag-of-words үлгісі;
- Bag-of-related words үлгісі;
- Латенттік семантикалық талдау (LSA);
- Латенттік Дирихле орналастыруы (LDA);
- Теріс емес матрицалық факторландыру;
- Тірек векторлар машинасы;
- Word2Vec үлгісі;

– Word embedding үлгісі (сөздерді сызықты векторлық кеңістікке кіргізу).

Зерттеу жұмысы барысында авторлар барлық аталған үлгілер мен әдістерді пайдаланды. Атап айтатын болсақ, Bag-of-words үлгісі мен оның қиындатылған Bag-of-related words нұсқасы құжаттарды жіктеу, мәтіндерді автоматты түрде сегменттеу, терминдер мен құжаттар арасындағы семантикалық байланыстарды шығару үшін пайдаланылды [99]. Латенттік семантикалық талдау ассоциативті байланыстарды іздеу, құжаттарды кілттік сөздермен индекстеу кезінде таңбалық кеңістікті төмендету, дистрибутивтік матрицаларды шу мен ыдыраудан тазалау үшін қолданылды [100]. [101] жұмыста көрсетілгендей, барлық жиіліктік матрицалар ыдыраған және шулы болып келеді. Сондықтан мұндай матрицалардың сипаттмаларын өзгертудің қарапайым әдісі – оларға сингулярлық ыдырау жүргізіп, тек негізгі таңбалық құрауыштарды ғана қалдыру арқылы таңбалық кеңістікті кішірейту. Латенттік Дирихле орналастыруы мен теріс емес матрицалық факторландыру мәтіндерді автоматты түрде сегменттеу және құжаттың визуалды бейнесін тақырыптар мен тірек кілттік сөздер жинағы түрінде құру үшін қолданылды. Word2Vec әдісі мен бағдарламалық құралы деректерді шығару жүйесін құру үшін пайдаланылды.

Терминдердің бірге кездесу матрицасының оңтайлы қолданысының бірі оның таксономияларды құру үшін пайдаланылуы болып табылады. Мұндай жағдайда матрица сөздердің барлық жиындарына емес, кілттік терминдер мен олардың атрибуттарына, яғни кілттік сөздермен бір паттернде кездесетін сөздердің жиындарына құрылады. Мұндай паттерндердің мысалы ретінде келесі түрдегі лексикалық-семантикалық үлгіні келтіруге болады: Бастауыш + Анықтауыш, Бастауыш + Баяндауыш, Бастауыш + Толықтауыш. Осындай жолмен құрастырылған матрица түсініктер торын құруда негіз болатын, яғни пәндік салалар мен олардың иерархиясының түсініктері анықталатын формальды мәнмәтін ретінде қолданыла алады [102]. Аталған

әдіс медицина саласындағы түсініктердің кілттік топтарын анықтау мақсатында медициналық мәтіндерге парсинг жасалған кезде қолданылған болатын [103]. Дистрибутивтік матрицалар негізіндегі формальды түсініктерді талдау кластерлеу және машиналық оқыту әдістерімен бірге автоматты түрде онтология құрастырудың күшті құралына айналады.

Дистрибутивтік матрицаларды сөздер арасындағы байланыстарды сипаттаудың әрі қарай семантикалық граф немесе концепт-карталар арқылы визуалдауға болатын жинақы әрі ыңғайлы тәсілі ретінде қарастыруға болады [104, 105, 106]. Графтың төбелері матрицаның жолдары мен бағандары, ал бүйірлері – салмақтары матрица мәндерінен анықталатын байланыстар болады.

Бұл жұмыста one-hot embedding векторлау әдісі қолданылды [107].

4.4. Тәжірибелік бөлім

Мәліметтер жинағы

Құрастырылған үлгі ағылшын тіліндегі мәліметтер арқылы оқытылды, CoNLL-2003 мәліметтер жинағы пайдаланылды [108]. Жинақ көлемі 1629 сөйлемді құрайды. Тәжірибе Python-да жүргізілді.

17-кесте

Мәліметтер жинағындағы деректі мәндердің санаттар бойынша саны

	LOC	MISC	ORG	PER
Жаттықтырушы жинақ	7140	3438	6321	6600
Сенімділікті растаушы жинақ	1837	922	1341	1842
Тестілік жинақ	1668	702	1661	1617

Мәліметтер 9 тегті қолдану арқылы белгіленді:

1. B-PER
2. I-PER
3. B-LOC

4. I-LOC
5. B-ORG
6. I-ORG
7. B-MISC
8. I-MISC
9. O

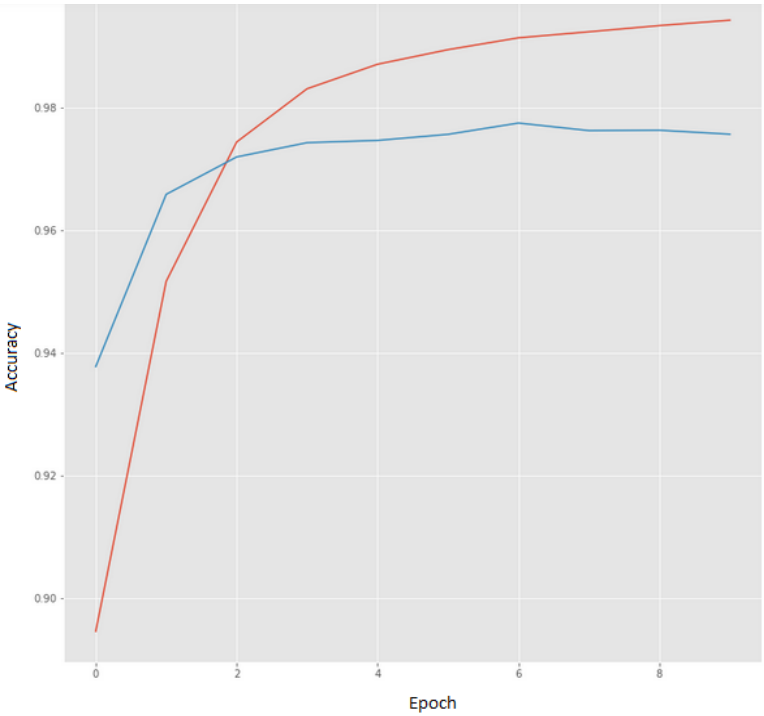
мұндағы PER – PERSON (адам, адам аттары), LOC – LOCATION (локация), ORG – ORGANIZATION (ұйым), MISC – MISCELLANEOUS (алдыңғы үш санатқа жатпайтын аралас деректі мәндердің атаулары), B – BEGINNING (деректі мәннің басы, бірінші токені), I – INSIDE (деректі мәннің әрі қарайғы (ішкі) токендері), O - OTHER, деректі мән болмайтын токен.

4.5. Нәтижелер

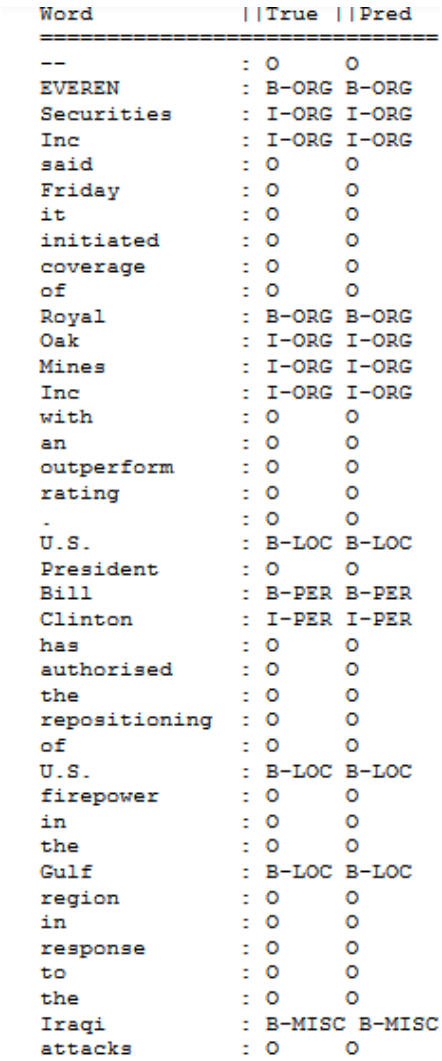
Үлгіні оқыту 10 дәуір ішінде жүргізілді. 29-суретте деректі мәндерді белгілеу дәлдігінің дәуірдің өтуі бойынша өзгерісі көрсетілген.

Сөйлемдегі деректі мәндерді тану мысалы 30-суретте келтірілген, екінші бағанда бастапқы тегтер, үшінші бағанда үлгі болжаған тегтер көрсетілген.

Құрастырылған үлгіні талдау F1-өлшем көмегімен жүргізілді. F1 мәні макро-орташаландырылған және микро-орташаландырылған өлшеммен есептелді (31-сурет).



29-сурет - Дәлдіктің (accuracy) дәуірге байланысты өзгеруі, қызыл сызықпен жаттығу кезіндегі дәлдік, көк сызықпен сенімділікті растау кезіндегі дәлдік көрсетілген.



Word	True	Pred
--	O	O
EVEREN	B-ORG	B-ORG
Securities	I-ORG	I-ORG
Inc	I-ORG	I-ORG
said	O	O
Friday	O	O
it	O	O
initiated	O	O
coverage	O	O
of	O	O
Royal	B-ORG	B-ORG
Oak	I-ORG	I-ORG
Mines	I-ORG	I-ORG
Inc	I-ORG	I-ORG
with	O	O
an	O	O
outperform	O	O
rating	O	O
.	O	O
U.S.	B-LOC	B-LOC
President	O	O
Bill	B-PER	B-PER
Clinton	I-PER	I-PER
has	O	O
authorised	O	O
the	O	O
repositioning	O	O
of	O	O
U.S.	B-LOC	B-LOC
firepower	O	O
in	O	O
the	O	O
Gulf	B-LOC	B-LOC
region	O	O
in	O	O
response	O	O
to	O	O
the	O	O
Iraqi	B-MISC	B-MISC
attacks	O	O

30-сурет - Сөйлемдегі деректі мәндерді тану мысалы, екінші бағанда бастапқы тегтер, үшінші бағанда үлгі болжаған тегтер көрсетілетін

```
f1-score with average macro: 0.8982352937221123
f1-score with average micro: 0.9799387757074335
```

31-сурет F1-өлшем мәні

Нейрондық желінің символдар мен сөздерді векторлауы бар bi-LSTM-ды қамтитын құрастырылған үлгісі деректі мәндерді тануда жақсы нәтижелер көрсетеді. Келешектегі зерттеу жұмыстарында сөздерді сызықты векторлық кеңістікке кіргізе отырып (word embedding), нейрондық желілердің біріккен әдістерін тиімді құрастыру және пайдалану жоспарлануда.

ҚОРЫТЫНДЫ

Пәндік сала мәтіндеріндегі терминдерді автоматты түрде шығару көптеген қосымшалары бар тапсырма болып табылады. Автоматты түрде шығарылатын терминдер құжаттарды рубрикациялауда жіктеуші белгілер ретінде, тезаурустар мен онтологияларды генерациялауда семантикалық концепт ретінде, БАҚ-ның контент-талдауында тірек түсініктер ретінде пайдаланылуы мүмкін. Іс жүзінде мәтінді автоматты түрде өңдеудің барлық тапсырмаларында, аннотациялауда, индекстеуде, жіктеуде, машиналық аудармада, білімдерді алуда терминология шығару қажет болады.

Аталған тапсырманы шешу үшін көптеген тиімді әдіс-тәсілдер құрастырылған, олардың ішіндегі ең қарапайым және тұрақтысы сөздерді қолдану статистикасына негізделетін әдістер болып табылады: ақпараттық пайда өлшемі, хи-квадрат өлшемі, өзара пайда өлшемі. TF-DCF, Weirdness және т.б. қарама-қарсы тәсілдер әдістердің бір үлкен тобын құрады. Терминдердің маңыздылығын құжаттың ішкі байланыстарына қарап анықтайтын қарама-қарсы емес әдістер де жақсы нәтижелер көрсетті.

Біз бұл монографияда терминология шығарудың көптеген қарама-қарсы әдістерін қарастырдық, сонымен қатар терминдерді шығару тапсырмасы мен тақырыптық үлгілеу тапсырмасының байланысын көрсеттік. Сондай-ақ, деректі мәндерді шығарудың сөздер мен символдар векторлық түрде көрсетіліп, мәнмәтіндік ақпаратты үлгілеу үшін екі бағытты LSTM блогының кірісіне берілетін үлгісі көрсетілді. Ақпаратты символдар деңгейінде кодтау үшін де екі бағытты LSTM қолданылды. Сөздердің сызықты векторлық кеңістікке кіру мәселесіне баса назар аударылады.

1-3-тараулардағы мысалдар R экожүйесінде орындалды, қазіргі таңда оның құрамында көптеген тілдерге, соның ішінде орыс және қазақ тілдеріне арналған табиғи тілді өңдеудің жетік аппараты бар. 4-тараудағы мысалдар Python тілінде орындалды. Python-да табиғи тілді өңдеуге арналған көптеген

кітапханалар бар: аса жылдам токендеу, талдау, мәтіндерді лемматизациялау және мәндерді тану. Python-ның құралдарын қолдану арқылы Word2Vec әдісінің көмегімен семантикалық сөздердің векторларын құру тапсырмасын шешуге және терең оқыту алгоритмдерін жүзеге асыруға болады.

ӘДЕБИЕТТЕР ТІЗІМІ

1. Frantzi K. T., Ananiadou S., Tsujii J. The c-value/nc-value method of automatic recognition for multi-word terms //International Conference on Theory and Practice of Digital Libraries. – Springer, Berlin, Heidelberg, 1998. –585-604 б.
2. Heylen K., De Hertog D. Automatic term extraction //Handbook of Terminology. – 2015. – Vol. 1. – 27 pp.
3. Simpson M. S., Demner-Fushman D. Biomedical text mining: a survey of recent progress //Mining text data. – Springer, Boston, MA, 2012. – 465-517 б.
4. Гусева О.И. Критерии терминологичности и корреляция “термин–слово общего языка” //Вісник Маріупольського державного гуманітарного університету. Сер.: Філологія. – 2008. – №. 1.
5. Коршунов А. В. Извлечение ключевых терминов из сообщений микроблогов с помощью Википедии //Труды Института системного программирования РАН. – 2011. – Т. 20.
6. Большакова Е. И., Васильева Н. Э. Терминологическая вариантность и ее учет при автоматической обработке текстов //Одиннадцатая национальная конференция по искусственному интеллекту с международным участием. – 2008. – Т. 2. –174-182 б.
7. Захаров В. П., Хохлова М. В. Автоматическое извлечение терминов из специальных текстов с использованием дистрибутивно-статистического метода как инструмент создания тезаурусов //Структурная и прикладная лингвистика. – 2012. – №. 9. –222-233 б.
8. Ефремова Н. Э. и др. Терминологический анализ текста на основе лексико-синтаксических шаблонов //Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». – 2010. – 124-130 б.
9. Браславский П.И., Соколов Е. А. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста //Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. Диалог. – 2006. –88-94 б.
10. Новикова Д. С. Автоматическое выделение терминов из текстов предметных областей и установление связей между ними //Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. – 2012.
11. Бессмертный И.А., Нугуманова А.Б., Платонов А.В. "Интеллектуальные системы: учебник и практикум для академического бакалавриата." М.: Издательство Юрайт (2017).
12. Ванюшкин А.С., Гращенко Л.А. Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. – 2016. – №. 19.

13. Conrado M., Pardo T., Rezende S. A machine learning approach to automatic term extraction using a rich feature set //Proceedings of the 2013 NAACL HLT Student Research Workshop. – 2013. –16-23 б.
14. Астраханцев, Н. А. "Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии." Астраханце Н. А. Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии //Труды Института системного программирования РАН. – 2014. – Т. 26. – №. 4.
15. Mihalcea R., Tarau P. Textrank: Bringing order into text //Proceedings of the 2004 conference on empirical methods in natural language processing. – 2004.
16. Turing A. M. Computing machinery and intelligence. 1950 //The Essential Turing: The Ideas that Gave Birth to the Computer Age. Ed. B. Jack Copeland. Oxford: Oxford UP. – 2004. – 433-64 б.
17. Straka M., Hajic J., Straková J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing //LREC. – 2016.
18. Бессмертный И.А. и др. Метод контрастного извлечения редких терминов из текстов на естественном языке //Научно-технический вестник информационных технологий, механики и оптики. – 2017. – Т. 17. – №. 1.
19. Nugumanova A. et al. A New Operationalization of Contrastive Term Extraction Approach Based on Recognition of Both Representative and Specific Terms //International Conference on Knowledge Engineering and the Semantic Web. – Springer, Cham, 2016. – 103-118 б.
20. Nugumanova A. et al. A contrastive approach to term extraction: Case-study for the information retrieval domain using BAWE corpus as an alternative collection. //Eurasian Journal of Mathematical and Computer Applications. – 2017. – Vol. 5. – P. 73-86.
21. da Silva Conrado M., Pardo T. A. S., Rezende S. O. A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set //HLT-NAACL. – 2013. – 16-23 б.
22. Ahmad K. et al. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER) //TREC. – 1999.
23. Gillam L., Tariq M., Ahmad K. Terminology and the construction of ontology //Terminology. – 2005. – Т. 11. – №. 1. – 55-81 б.
24. Peñas A. et al. Corpus-based terminology extraction applied to information access //Proceedings of Corpus Linguistics. – 2001. – Т. 2001.
25. S.N. Kim, T. Baldwin, and M-Y. Kan. 2009. An unsupervised approach to domain-specific term extraction. In Australasian Language Technology Association Workshop 2009, page 94-98
26. Basili R. A contrastive approach to term extraction. In Proceedings of the 4th Terminological and Artificial Intelligence Conference (TIA2001).

27. Wong W., Liu W., Bennamoun M. Determining termhood for learning domain ontologies using domain prevalence and tendency //Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70. – Australian Computer Society, Inc., 2007. – 47-54 б.
28. Sclano F., Velardi P. Termextractor: a web application to learn the shared terminology of emergent web communities //Enterprise Interoperability II. – Springer London, 2007. – 287-290 б.
29. Astrakhantsev N.A., Fedorenko D.G., Turdakov D.Y. Methods for automatic term recognition in domain-specific text collections: A survey //Programming and Computer Software. – 2015. – Т. 41. – №. 6. – 336-349 б.
30. Kit C., Liu X. Measuring mono-word termhood by rank difference via corpus comparison //Terminology. – 2008. – Т. 14. – №. 2. – 204-229 б.
31. Lopes L., Fernandes P., Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf //Knowledge-Based Systems. – 2016.
32. Z. Zheng, X. Wu, and R. Srihari. Feature Selection for Text Categorization on Imbalanced Data. // ACM SIGKDD Explorations Newsletter, 2004. – Vol. 6. – P. 80-89.
33. G. Forman. An extensive empirical study of feature selection metrics for text classiffication. // J. Mach. Learn. Res., 2003. – P. 1289-1305,
34. Lancaster, Henry Oliver, and E. Seneta. Chi-Square Distribution. John Wiley & Sons, Ltd, 1969.
35. Matsuo Y., Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information // International Journal on Artificial Intelligence Tools. – 2004. – Vol. 13(01). – Pp. 157-169.
36. Nugumanova A. et al. Using Non-Negative Matrix Factorization for Text Segmentation. // Int. Conference on Mathematical and Information-al Technologies. – Beograd, 2016. – Pp. 233-242.
37. Blei D. M. Probabilistic topic models //Communications of the ACM. – 2012. – Vol. 55(4). – Pp. 77-84.
38. Heuboeck A., Holmes J., Nesi H. The BAWE corpus manual. – Technical report, Universities of Warwick, Coventry and Reading, 2007.
39. http://ota.ahds.ac.uk/headers/2539.xml [Электрондық ресурс]
40. Ebeling S. O., Heuboeck A. Encoding document information in a corpus of student writing: the British Academic Written English corpus //Corpora. – 2007. – Т. 2. – №. 2. – 241-256 б.
41. Doyle J., Kešelj V. Automatic categorization of author gender via n-gram analysis //The 6th Symposium on Natural Language Processing, SNLP. – 2005. – 1-5 б.
42. Allahyari M., Kochut K. Automatic topic labeling using ontology-based topic models //Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference. – IEEE, 2015. – 259-264 б.
43. Park K., Lu X. Automatic analysis of thematic structure in written English //International Journal of Corpus Linguistics. – 2015. – Т. 20. – №. 1. – 81-101 б.

44. Reiter N. et al. Adapting standard NLP tools and resources to the processing of ritual descriptions //ECAI 2010. – 2010. – 39 б.
45. Kilgarriff A. Getting to know your corpus //International Conference on Text, Speech and Dialogue. – Springer Berlin Heidelberg, 2012. – 3-15 б.
46. Manning C. D. et al. Introduction to information retrieval. – Cambridge : Cambridge university press, 2008. – 496 б.
47. https://nlp.stanford.edu/IR-book/ [Электрондық ресурс]
48. Da Sylva L. Corpus-based derivation of a “basic scientific vocabulary” for indexing purposes //Journal of Linguistics. – 2009. – Т. 45. – №. 1. – 167-201 б.
49. Feinerer I. Introduction to the tm Package Text Mining in R //2013-12-01]. http://www, dainf, ct. utfpr, edu. br/-kaestner/Mineracao/RDataMining/tm, pdf. – 2017.
50. D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddl. Ontology-based extraction and structuring of information from data-rich unstructured documents. In Conference on Information and Knowledge Management (CIKM), pages 52–59, 1998.
51. C. Cardie. A case-based approach to knowledge acquisition for domain-specific sentence analysis. In Proceedings of the Eleventh National Conference on Artificial Intelligence, pages 798–803. AAAI Press, 1993.
52. T. Scheffer, C. Decomain, and S. Wrobel. Active hidden Markov models for information extraction. In Proceedings of the International Symposium on Intelligent Data Analysis, 2001.
53. T. Scheffer, S. Wrobel, B. Popov, D. Ognianov, C. Decomain, and S. Hoche. Learning hidden Markov models for information extraction actively from partially labeled text. K¨unstliche Intelligenz, (2), 2002.
54. M. Skounakis, M. Craven, and S. Ray. Hierarchical hidden Markov models for information extraction. In IJCAI, 2003.
55. A. K. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In ICML, 2000.
56. A. K. McCallum and D. Jensen. A note on the unification of information extraction and data mining using conditional-probability, relational models. In IJCAI’03 Workshop on Learning Statistical Models from Relational Data, 2003.
57. K.Shaalan. Rule-based approach in Arabic natural language processing. Int.J. Inf.Commun.Technol.3(3). 2010.11–19.
58. Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000, September). Rule-based named entity recognition for Greek financial texts. In Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000) (pp. 75-78). Patras, Greece
59. Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., & Vaithyanathan, S. (2010, October). Domain adaptation of rule-based annotators for named-entity recognition tasks. In Proceedings of the 2010 conference on em-

pirical methods in natural language processing (pp. 1002-1012). Cambridge, Massachusetts.
60. Ekbal, A., & Bandyopadhyay, S. (2009). A conditional random field approach for named entity recognition in Bengali and Hindi. Linguistic Issues in Language Technology, 2(1), 1-44.
61. Curran, J. R., & Clark, S. (2003, May). Language independent NER using a maximum entropy tagger. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (Vol. 4, pp. 164-167). Edmonton, Canada.
62. Szarvas, G., Farkas, R., & Kocsor, A. (2006, October). A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. In Proceedings of the International Conference on Discovery Science (pp. 267-278). Berlin, Heidelberg: Springer.
63. Collins, M., & Singer, Y. (1999, June). Unsupervised models for named entity classification. In 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP (pp. 100-110). PG County, USA.
64. Kim, J. H., Kang, I. H., & Choi, K. S. (2002, August). Unsupervised named entity classification models and their ensembles. In Proceedings of the 19th International Conference on Computational Linguistics-Volume 1 (pp. 1-7). Association for Computational Linguistics, Taipei, Taiwan.
65. Saha, S. K., Sarkar, S., & Mitra, P. (2008, January). A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition. In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I (pp. 343-349). Hyderabad, India.
66. Rocktäschel, T., Weidlich, M., & Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics, 28(12), 1633-1640.
67. Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Za-iqing Nie. Joint Entity Recognition and Dis-ambiguation. InProceedings of the 2015 Confer-ence on Empirical Methods in Natural LanguageProcessing, pages 879–888, Lisbon, Portugal. As-sociation for Computational Linguistics. 2015.
68. Lisa F Rau. 1991. Extracting company names from text. In Artificial Intelligence Applications, 1991.Proceedings., Seventh IEEE Conference on. IEEE, volume 1, pages 29–32.
69. Satoshi Sekine and Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hi-erarchy. In LREC. pages 1977–1980.
70. David Nadeau and Satoshi Sekine. 2007.A sur-vey of named entity recognition and classification. Lingvisticae Investigationes 30(1):3–26.
71. Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the seventh conference on Natural

language learning at HLT-NAACL 2003-Volume4. Association for Computational Linguistics, pages188–191.
72. Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In Proceedings of the fifth conference on Applied natural language processing. Association for Computational Linguistics, pages 194–201.
73. Masayuki Asahara and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In Proceedings of the 2003Conference of the North American Chapter of the Association for Computational Linguistics on Hu-man Language Technology-Volume 1. Association for Computational Linguistics, pages 8–15.
74. Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780. 1997.
75. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho¸and Yoshua Bengio. 2014. Empirical evaluation ofgated recurrent neural networks on sequence model-ing.arXiv preprint arXiv:1412.3555.
76. Tomas Mikolov, Martin Karafi´at, Lukas Burget, Jan Cernock`y, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Interspeech. volume 2, page 3.
77. Martin Sundermeyer, Ralf Schl¨uter, and Hermann Ney.2012. Lstm neural networks for language modeling. In Interspeech. pages 194–197.
78. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprintarXiv:1409.0473.
79. Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. arXiv preprint arXiv:1306.3584.
80. Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A hierarchical neural model for learning sequences of dialogue acts. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume1, Long Papers. Association for Computational Linguistics, Valencia, Spain, pages 428–437.
81. Xuezhe Ma and Eduard Hovy. 2016.End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354.
82. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer.2016. Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pages 260–270. http://www.aclweb.org/anthology/N16-1030.
83. Ronan Collobert, Jason Weston, L´eon Bottou, MichaelKarlen, Koray Kavukcuoglu, and Pavel Kuksa.2011. Natural language processing (almost) fromscratch.Journal of Machine Learning Research,12(Aug):2493–2537.

84. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. CoRR, abs/1603.01360. 2016.
85. Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXivpreprint arXiv:1508.01991.
86. Jason P.C. Chiu, Eric Nichols. Named Entity Recognition with Bidirectional LSTM-CNNs. 2016
87. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. CoRR, abs/1603.01360. 2016.
88. Marek Rei, Gamal Crichton, and Sampo Pyysalo. 2016.Attending to characters in neural sequence label-ing models. In Proceedings of the 26th Inter-national Conference on Computational Linguistics, pages 309–318.
89. Onur Kuru, Arkan Ozan Can, and Deniz Yuret. 2016.Charner: Character-level named entity recognition. In Proceedings of the 26th International Conferenceon Computational Linguistics, pages 911–921.
90. Jason P.C. Chiu, E. Nichols. Named Entity Recognition with Bidirectional LSTM-CNNs. arXiv:1511.08308v5 [cs.CL] 2016.
91. Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty oftraining recurrent neural networks. 2012.
92. K. Greff, R. K. Srivastava, J. Koutník, Bas R. Steunebrink and J. Schmidhuber. LSTM: A Search Space Odyssey. 2017.
93. C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith.Transition-based dependency parsing withstack long short-term memory. In Proceedingsof ACL-2015 (Volume 1: Long Papers), pages334–343, Beijing, China, July. 2015.
94. Нугуманова А. Б. и др. Обогащение модели Bag-of-words семантическими связями для повышения качества классификации текстов предметной области //Программные продукты и системы. – 2016. – №. 2 (114).
95. Wei Y., Wei J., Xu H. Context vector model for document representation: a computational study //National CCF Conference on Natural Language Processing and Chinese Computing. – Springer International Publishing, 2015. – С. 194-206.
96. Mikolov T., I. Sutskever, K. Chen, G. Corrado, J. Dean. Distributed representations of words and phrases and their compositionality. NAACL HLT. 2013.
97. Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. Indexing by Latent Semantic Analysis. The American Society for Informa-tion Science.1990;41:391-407
98. J. Pennington, R. Socher, C. D. Manning. GloVe: Global vectors for word representation. 2014.
99. Nugumanova, A.,Mansurova, M.,Baiburin, Y.,Alimzhanov, Y. Using non-negative matrix factorization for text segmentation // 2017. CEUR Workshop Proceedings – P.233-242.

100. Nugumanova, A.,Mansurova, M.,Alimzhanov, E.,Zyryanov, D.,Apayev, K. An automatic construction of concept maps based on statistical text mining // Communications in Computer and Information ScienceData Management Technologies and Applications. 4th International Conference, DATA Colmar, France, July 20-22, 2015, Revised Selected Papers. 2016. P. 29-38.
101. 3. Slonim, N., Tishby, N. The power of word clusters for text classification. Proceedings of the 23rd European Colloquium on Information Retrieval Research, Vol. 1. –2001.
102. 5. Самойлов Д. Е., Семенова В. А., Смирнов С. В. Анализ неполных данных в задачах построения формальных онтологий //Онтология проектирования. – 2016. – Т. 6. – №. 3 (21).
103. Досанов Б.Б., Мансурова М.Е., Нугуманова А.Б. Статистикалық әдістер негізінде пәндік аймақ онтологиясын құру // Вестник ВКГТУ №3 Том 1, Часть 3, 2018 г. – 112-120 б.
104. Nugumanova, A., Mansurova, M.,Alimzhanov, E.,Zyryanov, D.,Apayev, K. Automatic generation of concept maps based on collection of teaching materials // DATA 2015 - 4th International Conference on Data Management Technologies and Applications, Proceedings, 2015. - P.248-254.
105. Nugumanova, A.,Mansurova, M.,Alimzhanov, E.,Zyryanov, D.,Apayev, K. An automatic construction of concept maps based on statistical text mining // Communications in Computer and Information ScienceData Management Technologies and Applications. 4th International Conference, DATA Colmar, France, July 20-22, 2015, Revised Selected Papers. 2016. P. 29-38.
106. Nugumanova A., Baiburin Y., Mansurova M., Alimzhanov Ye. An Investigation of the Educational Curriculum with Use of Formal Concept Analysis // Proc. of 10th IADIS International Conference on Information Systems, Budapest, 10-12 of March 2017. – P. 567-574.
107. https://www.tensorflow.org/beta/tutorials/text/word_embeddings [электрондық ресурс]
108. https://github.com/davidsbatista/NER-datasets/tree/master/CONLL2003 [электрондық ресурс]

МАЗМҰНЫ

Ғылыми басылым

Алия Нугуманова
Мадина Мансурова

ТАБИҒИ ТІЛ МӘТІНДЕРІНДЕГІ ТЕРМИНДЕРДІ АВТОМАТТЫ ТҮРДЕ ТАНУ

Монография

Шығарушы редактор *Г.С. Бекбердиева*
Компьютерлік беттеу *Г.К. Шаккозовой*
Мұқаба дизайны: *А. Калиева*

АҚ №
Басуға қол қойылды 15.09.2019. Пішімі 60х84/16.
Офсеттік қағаз. Сандық баспа. Б.қ. көлемі
Таралымы 500 дана. Тапсырыс № . Келісілген баға.
С. Аманжолов атындағы ШҚМУ
Өскемен қаласы