DOI 10.31489/2023Ec4/213-228
JEL C81 C87
ӘОЖ 339.137.22

**Т.К. Шурен[1*], Т.П. Притворова[2], Н.П. Иващенко[3], Г.А. Райханова[4]**

*[1,2,4] Академик Е.А. Бөкетов атындағы Қарағанды университеті, Қарағанды, Қазақтан;*
*[3]М.В. Ломоносов атындағы Мәскеу мемлекеттік университеті, Мәскеу, Ресей*
*[1]itoktar@gmail.com, [2]pritvorova_@mail.ru, [3]nivashenko@mail.ru, [4]gulnurraihanova@mail.ru*

# Бәсекелестік тенденцияларын салыстырмалы талдау үшін мәтінді интеллектуалды талдау технологиясын қолдану

***Аңдатпа:***

*Мақсаты:* Мақаланың негізгі мақсаты Python-да мәтінді интеллектуалды талдау бағдарламасын құру және қолдану арқылы ғалымдардың жұмысындағы бәсекелестіктің оң және теріс тенденцияларын, сондай-ақ тиісті траекторияларын анықтау.

*Әдісі:* Зерттеуде мәтінді өңдеудің заманауи әдістері, атап айтқанда мәтінді интеллектуалды талдау қолданылған. Сонымен қатар Python бағдарламалау тілі мен MS Excel бағдарламалық жасақтамасы және Wiley Online Library кітапханасының ресми сайтындағы деректер пайдаланылған.

*Қорытынды:* Заманауи технологиялардың, интернет-қосымшалардың қарқынды дамуы үлкен көлемдегі деректерді шығарумен қатар жүреді, оларды уақтылы өңдеу бүгінде өмірдің әртүрлі салаларындағы — әлеуметтік, экономикалық, саяси және басқалардағы басты мәселелердің бірі. Осы жаһандық мәселені шешуде мәтіндік ақпаратты өңдеудің заманауи әдістері, мәтінді интеллектуалды талдау технологиялары көмекке келеді. Бұл құралдар әртүрлі деңгейдегі мәселелерді шешудің тиімділігін арттыруға мүмкіндік береді. Мәтінді интеллектуалды талдау технологиясына енгізілген алгоритмдер мәтіннің негізгі ұғымдарын, оның мазмұнын және олардың арасындағы байланысты ашады.

*Тұжырымдама:* Соңғы 30 жылда жазылған еңбектерде ғалымдар негізінен нарық, баға, зерттеу, шығармашылық, өнеркәсіп және т.б. мәселелерді көтергенін анықтадық. Сондай-ақ тұтынушыларды, инновацияларды және өнімдерді зерттеу айтарлықтай өсті. Ең үлкен өсім 1993–1998 жылдардағы жинақта небәрі 11 рет кездескен платформа сөзімен байланысты болды, бірақ 2019–2023 жылдардағы жинақта дәл осы сөз 4199 рет қолданылды, яғни бұл 38,000 %–дан астам өсуді көрсетті.

***Кілт сөздер:*** мәтінді интеллектуалды талдау, бәсекелестік, платформа, бизнес, цифрландыру, оқу.

## *Кіріспе*

Мәтінді ұзақ уақытты қажет ететін қолмен талдау әдісінің процесі өткеннен қалған. Бүгінгі таңда мәтіндік деректердің үлкен массивтерін бағдарламалық жасақтаманы пайдаланбай зерттеу мүмкін емес. Заманауи ақпараттық технологиялар зерттеушілерге мәтінді компьютерлік өңдеу және интеллектуалды талдау әдістерін қолдануға мүмкіндік береді. Интернеттің қарқынды дамуы ғылыми мақалалардан, онлайн-пікірталастардан, веб-сайттардан, чаттардан, пайдаланушылардың пікірлерінен, газеттерден, әлеуметтік желілерден және басқа да ашық көздерден деректерді өңдеуге арналған ақпараттық ресурстарды алуға мүмкіндік береді. Сондықтан мәтінді интеллектуалды талдау әдістері қазіргі уақытта бизнестің, саясаттың, білімнің және т.б. түрлі салаларында, ең алдымен, тезистер, шолулар мен баяндамалар мәтіндерінен алынған ақпараттың мазмұнын визуализациялау үшін ең өзекті және сұранысқа ие болып табылады. Мысалы, АҚШ-тың Денсаулық сақтау министрлігінде word бұлттары ұйымның негізгі қызметіне жеткілікті көңіл бөлінетінін анықтау үшін құжаттардың мазмұнын талдау үшін пайдаланылды (Atenstaedt, 2017).

Сөз бұлттарын жасайтын көптеген онлайн құралдар бар. Алғашқылардың бірі Wordle.net. Бұл құралдар мәтінде жиі кездесетін сөздерді көрнекі түрде көрсетеді және зерттелетін ақпарат туралы жалпы түсінік алудың жылдам әдісі ретінде қызмет етеді (мақала мәтіні, баяндамашының сөзі, блог немесе дерекқор жазбалары, респонденттердің онлайн жауаптары, түсініктемелер және т.б.). Кейбір жағдайларда сөз бұлттары қосымша, тереңірек зерттеуге түрткі болатын деректердің ерекше ерекшеліктерін анықтай алады. Сондай-ақ, мәтінді талдауда ғылыми мақалаларды талдауда осы әдісті сирек қолданудың себебі болып табылатын кейбір кемшіліктер бар екенін атап өткен жөн.

## *Әдебиетке шолу*

Соңғы уақытта зерттеулердің көпшілігі жасанды интеллект, машиналық оқыту және т.б. салалардағы жаңа технологияларды қолдана отырып жүргізілді. Осылайша, мәтінді интеллектуалды талдау бүгінде кеңінен зерттелуде. Мәтінді интеллектуалды талдау мәтіндік құжаттардан тиісті білімді алады. Мәтінді өңдеудің әртүрлі әдістері құрылымдалмаған деректерді құрылымдалғанға айналдырады. Мәтінді талдаудың негізгі принциптерінің бірі болып табылатын мәтінді жіктеу

* Хат-хабарларға арналған автор. E-mail: *itoktar@gmail.com*

мәтінді өңдеудің бірқатар әдістерін қолдануды талап етеді, олардың ішіндегі ең маңыздысы — табиғи тілді өңдеу (NLP) (Udgave, Kulkarni, 2020).

Интернетте көптеген зерттеу жұмыстары жарияланады. Компьютерлік және ақпараттық технологиялардың дамуының өсуі пайдаланушыларға белгілі бір тақырып бойынша қызықты ғылыми мақалаларды табуды және жіктеуді қиындатады (Cai, Luo, Wang, Yang, 2018). Сондықтан ғылыми мақалаларды ұқсас тақырыптар бойынша жүйелі түрде жіктейтін механизм болған жөн. Бұл пайдаланушыларға өздерін қызықтыратын зерттеу жұмыстарын тез және оңай табуға мүмкіндік береді. Әдетте, белгілі бір тақырыптар немесе тақырыптар бойынша зерттеу жұмыстарын іздеу көп уақытты алады. Мысалы, зерттеушілер өздерін қызықтыратын мақалаларды табу үшін интернетте көп уақыт өткізеді. Мақалалар тақырыптар бойынша топтастырылмағандықтан немесе қажетті ақпаратқа қол жетімділіктің болмауына байланысты қажетті ақпарат тиімсіз алынады (Большакова және басқалар, 2017).

Бүгінгі таңда үлкен деректер технологиясының арқасында бұл мәселе толығымен шешілді. Көптеген ғылыми жұмыстарды талдаудың, жіктеудің және өңдеудің заманауи мүмкіндіктері бұл жұмысты тиімді, басқарылатын және қол жетімді етеді. Автоматтандырылған өңдеу әдістерін қолдану жыл сайын ғылыми жұмыстардың көбеюі зерттеушілерге көмекке келеді. Олар мақаланың мәнін сипаттауға, мақаланың негізгі бөлігіндегі мазмұнды оқымас бұрын зерттеу бағытын және қысқаша мазмұнын алуға мүмкіндік береді. Осыған байланысты ғылыми мақалалардың түйінді сөздері қысқаша және ақпараттық түрде жазылуы керек (Калабин, Корнеева, 2020).

Көптеген мақалаларды ұқсас тақырыптағы мақалаларға жіктеу үшін ғалымдар С. Ким және Дж. Гил (2019) *term-frequency — inverse document frequency (TF-IDF)* схемаларына және Дирихлеттің жасырын таралуына (LDA) негізделген мақалаларды жіктеу жүйесін қолдануды ұсынады. Ұсынылған жүйе алдымен пайдаланушы енгізетін кілт сөздермен және LDA шығарған тақырыптармен кілт сөздердің репрезентативті сөздігін жасайды. Екіншіден, ол TF-IDF схемасын кілт сөздер сөздігіне негізделген мақала аннотацияларынан тақырыптық сөздерді шығару үшін пайдаланады.

Эксперименттік нәтижелер ұсынылған жүйе кілт сөз қатынасы бойынша ұқсас тақырыптағы барлық мақалаларды жақсы жіктей алатынын көрсетеді. TF-IDF және LDA схемаларына негізделген жіктеу жүйесі кеңінен қолданылады, өйткені ол өте тиімді (Нгуен, 2019).

Деректерді өңдеудің заманауи әдістерін жаңарту осы құралды пайдалану процесін жетілдірудің тиімді жолдарын табуды талап етеді. R. Atenstaedt (2012) өзінің зерттеуінде осы технологияларды тереңірек зерттеуге ықпал ететін бұлттардың ерекшеліктері мен қолдану салаларын ашады.

Мәтінді іздеу — бұл мәліметтерден ұғымдарды, үлгілерді, тақырыптарды, кілт сөздерді және басқа атрибуттарды таба алатын бағдарламалық жасақтама арқылы құрылымдалмаған мәтіндік деректердің үлкен көлемін зерттеу және талдау процесі.

Мәтінді іздеу деректерді өндіруге ұқсас, бірақ деректердің құрылымдық формаларына емес, мәтінге бағытталған.

Бұрын NLP алгоритмдері негізінен статистикалық модельдерге немесе компьютерлерге деректер жиынтығында не іздеу керектігін көрсететін ережелерге негізделген модельдерге негізделген. Алайда, аз бақылаумен жұмыс істейтін терең оқыту модельдері мәтінді талдауға және үлкен деректер жиынтығын қолданатын басқа жетілдірілген аналитикалық тапсырмаларға танымал балама бола бастады.

Терең оқыту дәстүрлі машиналық оқытуға қарағанда икемді және интуитивті деректерді итеративті талдау үшін нейрондық желілерді пайдаланады. Осының арқасында мәтінді өңдеу құралдары енді мәтіндік деректерде ұқсастықтар мен байланыстарды таба алады, тіпті деректер мамандары жобаның басында не табатынын білмесе де. Мысалы, бақыланбайтын модель талдаушының көмегінсіз мәтіндік құжаттардан немесе электрондық пошталардан тақырыптар тобын біріктіре алады. Sentiment analysis деп аталатын танымал мәтіндік интеллектуалды бағдарлама компания туралы пікір алуға мүмкіндік береді. Пікірді талдау деп те аталатын көңіл-күйді талдау онлайн шолулардан, әлеуметтік желілерден, электрондық пошталардан, байланыс орталығымен өзара әрекеттесуден және басқа деректер көздерінен клиенттердің не жақсы немесе жаман сезінетінін көрсететін жалпы қауіптерді табу үшін мәтінді шығарады.

Мәтінді талдау қиын болуы мүмкін, өйткені ақпарат жиі түсініксіз, сәйкес келмейді немесе тіпті қарама-қайшы. Синтаксис пен мағынадағы айырмашылықтарға, сондай-ақ жаргонды, сарказмды, аймақтық диалектілерді және әрбір тік салаға тән техникалық тілді қолдануға байланысты бұл нені білдіретінін түсінуге тырысу одан да қиын.

***Әдістері***

Бұл зерттеу үшін біз мәтінді өңдеудің заманауи әдістерін, атап айтқанда мәтінді интеллектуалды талдауды қолдандық. Біз nltk пакетімен және MS Excel бағдарламалық жасақтамасымен Python бағдарламалау тілін қолдандық.

Деректер жиынтығына тән біржақтылық бұл тереңдетіп оқыту құралдардың дұрыс емес нәтиже беруіне әкелуі мүмкін тағы бір мәселе, егер деректер мамандары модельдерді құру кезінде біржақтылықты қабылдамаса. Мәтінді өңдеуге арналған көптеген бағдарламалар бар. Бағдарламалық жасақтаманың ірі компаниялары болып табылатын IBM, oracle, SAS, SAP және TIBCO құралдарын, сондай-ақ Google Collab-ты қоса алғанда, қолдануға болатын ондаған ашық бастапқы коммерциялық технологиялар бар. Алайда, бұл бағдарламалар қарапайым зерттеушіге қол жетімді емес, ал ашық бастапқы бағдарламалық жасақтама үлкен көлемдегі деректерді өңдей алмайды және оларды нақты қажеттіліктері үшін қайта бағдарламалау мүмкін емес. Дәлірек және кеңірек нәтижеге қол жеткізу үшін NLTK бумасы бар Python бағдарламалау тіліне негізделген мәтінді өңдеу бағдарламасын құру туралы шешім қабылданды.

Natural language toolkit немесе NLTK — бұл пайдаланушыларға табиғи тілді өңдеудің көптеген әдістеріне қол жеткізуге мүмкіндік беретін сенімді Python пакеті. Бұл құралдың артықшылығы — ол тиімді, ашық дерек көзі бар, пайдалану оңай, үлкен қоғамдастыққа ие және жақсы құжатталған. Nltk ең көп қолданылатын алгоритмдерді қамтиды, соның ішінде токенизация, сөйлеу бөліктерін белгілеу, стемминг, көңіл-күйді талдау, тақырыптар бойынша сегментация және аталған нысандарды анықтау.

***Нәтижелер***

Мәтінді интеллектуалды талдаудың алғашқы қадамдарының бірі **—** деректерді сапалы және сандық түрде талдауға болатындай етіп ұйымдастыру және құрылымдау. Бастапқы жұмыс мәтінді топтастыруды, кластерлеуді және белгілеуді, деректер жиынтығын жалпылауды, таксономияларды құруды және сөз жиілігі мен деректер объектілері арасындағы қатынастар, бизнес стратегиялары және операциялық әрекеттер сияқты нәрселер туралы ақпарат алуды қамтиды. Бұл әдіс салыстырмалы түрде жаңа болғандықтан, қазіргі уақытта мәтінді іздеу үшін белгіленген ережелер жоқ. Дегенмен, төменде көрсетілген бірнеше жалпы қадамдар бар:

Мәтінді өңдеудің алғашқы қадамы **—** талдау үшін тиісті мәтіндік деректер жиынтығын анықтау және шығару. «Корпус» деп аталатын жиынтық жасалады. Корпус бұл барлық талданатын мәтіндерді қамтитын объект. Мәтін корпусымен әр түрлі операцияларды орындауға болады, мысалы, барлық сөздерді бас әріптермен ұсыну (tolower), тыныс белгілерін жою (remove Punctuation), артық бос орындарды жою және басқалар (Кабакоф, 2015).

Мәтіндік деректерді өңдеудің негізгі кезеңдері: деректерді тазарту; лемматизация; элементтерді бөлектеу.

Деректерді тазарту сандық деректерді, бос орындарды жоюды, бас әріптерді кіші әріптермен ауыстыруды қамтиды. Сонымен қатар, 1-кезеңде «тоқтату сөздері» жойылады немесе оларды «шу сөздері» деп те атайды. Яғни, өздері ешқандай семантикалық жүктемені көтермейтін сөздер (тым жиі, тым сирек, тым қысқа, зат есімдер емес, жалқы есімдер). Оларға предлогтар, жұрнақтар, жіктік жалғаулар, шылаулар, сандар, бөлшектер, жалғаулар жатады. Мысалы, «жоқ», «сондай-ақ», «бұл», «не», «арасында», «әрқашан» және тағы басқалар.

Стемминг **—** берілген бастапқы сөз үшін сөздің негізін табу процесі (сөзді негізге кесу). Негізді қалыптастыру процесінде сөздердің аяқталуы алынып тасталады. Шығу тегі тіл морфологиясының ережелеріне негізделген. Осылайша, stemming сөздің барлық грамматикалық формалары үшін қалған бөлігі бірдей болатындай етіп сөздің аяқталуы мен жұрнақтарын кесіп тастайды.

Лемматизация сөздің леммасын анықтау процесі. Лемма **—** бұл сөздің бастапқы, негізгі формасы. Зат есімдер мен сын есімдер үшін бұл номинативті сингулярлық форма, ал етістіктер үшін инфинитив.

Мақалада біз 1993–2013 жылдар аралығында PDF форматында «бәсекелестік» және «бизнес» кілт сөздерін қамтитын «Wiley Онлайн кітапханасындағы» мақалаларды талдадық. Бұл кітапханадағы мақалалар ағылшын тілінде болғандықтан, зерттеуде ағылшын сөздері және ағылшын грамматикасы қолданды. Оны біз алты кезеңге бөлдік: 1993–1998, 1999–2003, 2004–2008, 2009–2013, 2014–2018 және 2019–2023 жылдар, олардың әрқайсысында 100 мақала бар, барлығы 600 мақала. Біз сондай-ақ мақала санын теңдестірілген деп санаймыз, өйткені ол енгізілген деректердің жеткілікті көлемін қамтиды, ал қосымша мақалаларды қосу үшін өте күрделі компьютерлер қажет болады. Алайда, бұл

мақалалар PDF форматында, сондықтан біз оларды қарапайым мәтінге айналдыруымыз керек. Ол үшін Python бағдарламалау тілінің pdfminer бумасын қолдана аламыз. Авторлар жазған келесі код барлық мақалаларды PDF форматында тезірек түрлендіруге көмектеседі (1-сурет):

```
import re
from pdfminer.high_level import extract_text

data = ""

for i in range(1, 101):
    text = extract_text(f"1/{i}.pdf")
    data = str(data) + str(text) + "/n"
    print(f"{i}/100")

f = open('6.txt', 'w')
f.write(data + '/n')
f.close()
```

1-сурет. Pdf файлдарын txt файлдарына түрлендіру

*Ескерту: Visual Studio коды негізінде авторлар құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/.*

Біз алты қалта (папка) жасадық және оларды 1993–1998 аралығындағы кезең үшін — 1, 1999–2003 аралығындағы кезең үшін — 2 және т.б. деп атап, сол қалталардың ішіне сәйкес кезеңдегі мақалалар салынды, осылайша әр қалтада 100 мақала болды. Жоғарыда аталған бағдарлама жұмыс істеуі үшін біз әр қалтадағы барлық мақалалардың атын «n.pdf» форматына өзгертуіміз керек, мұндағы «n» — олардың қалтадағы орны. Авторлар бұл қадамды қолмен жасады, бірақ мұнда арнайы бағдарламаларды да қолдануға болады. Осыдан кейін бағдарламаны іске қосып, жұмыс істейміз. Нәтижесінде, жоғарыдағы код NLTK-де пайдалануға дайын алты «txt» файлын жасайды.

Екінші қадам — деректерді санаттау, жалпылау және ұйымдастыру үшін алгоритмдерді іске қосу. Мұнда, жоғарыда айтылғандай, біз Python бағдарламалау тілін, соның ішінде NLTK бумасын қолдандық.

Мәтінді интеллектуалды талдаудың үшінші кезеңінде аналитикалық модельдер идентификаторлар, шаблондар және басқа атрибуттар ұғымдарын анықтау үшін пайдаланылады.

Төртінші қадам нәтижелерді қолдануға, ал бесінші қадам нәтижелерді визуализациялау және бөлісу үшін деректерді дайындауға қатысты.

Nltk кітапханасы компьютерге жазбаша мәтінді талдауға, алдын ала өңдеуге және түсінуге көмектеседі. Біз осы кітапхананы пайдаланып мәтінді өңдеу тұжырымдамаларын жүзеге асырдық.

Мәтінді іздеу үшін Visual Code Studio бағдарламалау ортасын қолдандық. Visual Code Studio-ны пайдаланудың артықшылықтарының бірі — бұл ортаның ыңғайлы интерфейсі бар және оны тегін жүктеп алуға болады. Осы мақсатта Google Collab бағдарламалық жасақтамасын да қолдануға болады, өйткені онда NLTK пакеті орнатылған. Алайда, Google Collab бағдарламалық жасақтамасы бізге үлкен деректерді талдауға мүмкіндік бермеді. Осылайша, біз Visual Code Studio бағдарламасын пайдалануды және Python бағдарламалау тілін 3.7 немесе одан жоғары нұсқада және nltk бумасын қолмен орнатуды шештік. NLTK кітапханасын импорттағаннан кейін келесі тапсырма punkt, тоқтату сөздері және wordnet жүктеу болды.

Punkt сөйлем токенизаторы мәтінді сөйлемдер тізіміне бөледі, бақыланбайтын алгоритмді қолдана отырып, сөйлемдер басталатын қысқартулар мен сөздер бар сөз тіркестерінің моделін жасайды. Nltk-дегі тоқтату сөздері деректердегі ең көп таралған сөздер болып табылады. Бұл сіздің мазмұныңыздың тақырыбын сипаттау үшін қолданғыңыз келмейтін сөздер. Олар алдын-ала анықталған және оларды жою мүмкін емес. Дегенмен тоқтату сөздерінің тізіміне қосымша сөздерді қосуға рұқсат етіледі. Wordnet — бұл табиғи тіл құралдарының бөлігі болып табылатын ағылшын тілінің сөздігі.

Осы кітапханаларды жүктеу үшін келесі командаларды пайдалануға болады:

nltk.download(‘punkt’)

nltk.download(‘stopwords’)

nltk.download(‘wordnet’)

Олар жүктеліп, орнатылғаннан кейін Python мәтінді өңдеу кодына дайын болады.

1. Мәтінді интеллектуалды талдаудың бастапқы кезеңі токенизация деп аталады. Токенизация сөйлемдер немесе сөздер сияқты мәтіннің үлкен бөліктерін сөздер немесе сөз тіркестері сияқты басқарылатын бірліктерге бөлу процесі. Фраза немесе абзац құрылыс блоктары ретінде жұмыс істейтін жеке субъектілер болып табылатын таңбалауыштар арқылы құрылады. Токенизацияның екі түрі бар. Сөйлемдер және сөздерді токенизациялау. Осы мақаланың мақсаттары үшін біз сөздерді белгілеуіміз керек.

Ең алдымен, мұны істеу үшін «word_tokenize» модулін импорттау керек. Осыдан кейін «text» деп аталатын айнымалы жасалды, содан кейін оған ұсыныстар жиынтығын сақталды. Бұл талдау үшін сәйкес мәтіндік деректер жиынын анықтау және шығару кезінде алған «txt» файлдары. Әрі қарай салыстыру үшін біз әр «txt» файлын бөлек талдаймыз. Әрі қарай, біз «tokenized_text» деп аталатын айнымалы жасаймыз, содан кейін «word_tokenize» функциясын қолданамыз және «text» айнымалысын осы функцияның мәні ретінде орналастырамыз (2-сурет).

```
1   import re
2   import nltk
3   from nltk.tokenize import word_tokenize
4
5   text = r'''
6
7   '''
8
9   tokenized_text = word_tokenize(text)
10  print(tokenized_text)
```

2-сурет. Nltk пакетін импорттау және айнымалылар құру

*Ескерту: Visual Studio коды негізінде авторлармен құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/*

2. Бұл код үзіндісі мәтінді сөз токендеріне бөледі. Токенизация аяқталғаннан кейін біз токендерден құнды және пайдалы ақпаратты біле аламыз. Жиіліктің таралуы солардың бірі. Дегенмен, бұған дейін біз тоқтату сөздерін тауып, жоюымыз керек. Тоқтату сөздері мәтінді интеллектуалды талдауда ешқандай рөл атқармайтын сөздер мен сөйлемдер. Әдетте ағылшын тілінде «am, is, are, this, a, an» тоқтату сөздері ретінде қарастырылады. Алайда, мәтінді интеллектуалды талдаудың мақсатына байланысты әр түрлі сөздер тоқтау сөздері ретінде қарастырылуы мүмкін.

Тоқтату сөздерін жою үшін біз келесі код жолдарын қолдануымыз керек (3-сурет):

```
1   import nltk
2   from nltk.corpus import stopwords
3
4   stop_words = set(stopwords.words("english"))
5   stop_words.add("1")
6   print(stop_words)
```

3-сурет. Тоқтату сөздерінің функциясын жүзеге асыру

*Ескерту: Visual Studio коды негізінде авторлармен құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/*

Мұнда nltk, nltk.corpus кітапханалары импортталды. Содан кейін stop_words деп аталатын айнымалы жасалды және set функциясын қолданды. Бұл функцияның мәні ретінде stopwords.words болады. Ол үшін тіл болып табылатын тағы бір мән қажет. Мұнда біз ағылшын тілін қолданамыз. Дегенмен, басқа тілдерді, соның ішінде қазақ немесе орыс тілін де қолдануға болады. Егер біз тоқтату сөздерін басып шығарсақ, онда біз негізгі сөздерді көре аламыз. Қосымша тоқтату сөздерін қосу үшін «stopwords.add» функциясын енгізіп оның мәніне қосымша тоқтату сөздерін қосуға болады. Ағымдағы жұмыста 67 қосымша тоқтату сөздері қосылды. Міне, олардың кейбіреулері (4-сурет):

```
stop_words.add("e")
stop_words.add("may")
stop_words.add("also")
stop_words.add("n")
stop_words.add("c")
stop_words.add("one")
stop_words.add("h")
stop_words.add("httpsonlinelibrarywileycomdoi")
stop_words.add("journal")
stop_words.add("applicable")
stop_words.add("′")
stop_words.add("–")
stop_words.add("′")
stop_words.add("l")
stop_words.add("would")
stop_words.add("ltd")
stop_words.add("two")
stop_words.add("model")
stop_words.add("httpsonlinelibrarywileycomtermsandconditions")
stop_words.add("governed")
stop_words.add("however")
stop_words.add("level")
stop_words.add("care")
stop_words.add("health")
stop_words.add("f")
stop_words.add("number")
stop_words.add("results")
stop_words.add("g")
stop_words.add("table")
stop_words.add("pp")
```

4-сурет. Бағдарламаға қосымша тоқтату сөздерін қосу

*Ескерту: Visual Studio коды негізінде авторлармен құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/.*

Тоқтату сөздерін табудың мақсаты оларды жою. Бұл қадам маңызды ақпаратты жойылып қалмау үшін мұқият ойластырылуы керек. Маңызды емес сөздер жұмыс барысында кейінгі кезеңдерде шығуы мүмкін, сондықтан мәтінді өңдеу процесінде тоқтату сөздерін өзгерту әдеттегі тәжірибе болып табылады. Тоқтату сөздерін тапқаннан кейін оларды жою үшін келесі кодты іске қосуға болады (5-сурет):

```
filtered_words = []

for w in tokenized_text:
    if w not in stop_words:
        filtered_words.append(w)
print(filtered_words)
```

5-сурет. Тоқтату сөздерінің функциясын жүзеге асыру

*Ескерту: Visual Studio коды негізінде авторлармен құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/*

3. Нәтижесінде тоқтату сөздері жойылған барлық сөздік бірліктерді алынды және олар «filtered_words» айнымалысында сақталды. Әрі қарай, біз барлық алты кезеңдегі ең кең таралған 30 сөздің таралуын таба аламыз. Жиіліктің таралуын білу үшін «freqdist» модулі импортталды. Осыдан кейін «frequency» деп аталатын айнымалы анықталды.

Содан кейін FreqDist функциясы қолданды. Бұл функция үшін мән қажет. Жоғарғы кодта бұл «filtered_words». FreqDist жалпы үлестірімді көрсетеді. Кейде бізге барлық сөздердің жиілік таралуы қажет емес. Оның орнына біз ең көп таралған сөздердің жиілігін білуіміз керек. Егер бізге ең көп таралған 30 сөздің жиілігі қажет болса, frequency.most_common функциясын қолданып, осы әдістің мәні ретінде 30 санын пайдалану керек.

Мәтіннен құнды ақпаратты алу үшін кейде визуалды бейнелеудің көмегіне жүгінуге тура келеді. Жиіліктің таралуының графикалық көрінісі сандық нәтижеге қарағанда көбірек ақпаратты ашады.

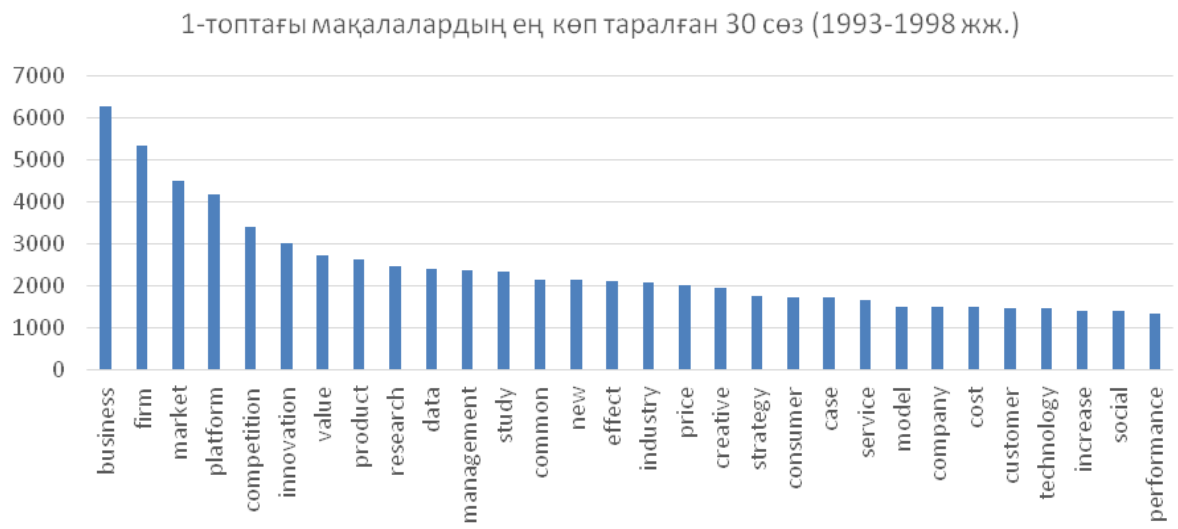
Жиіліктің таралуын графикалық түрде көрсету үшін matplotlib кітапханасының plt. функциясын пайдалану керек (6-сурет):

```
frequency = FreqDist(filtered_words)
frequency.plot(30, cumulative=False)
plt.show()
print("...nearly finished...")
print(frequency.most_common(30))
print("...Precessing finished. Thanks for waiting...")
```

6-сурет. FreqDist функциясы арқылы ең танымал 30 сөздің таралуын іздеу

*Ескерту: Visual Studio коды негізінде авторлармен құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/.*

Енді осы кодты іске қосқаннан кейін келесі нәтижені көре аламыз (7-сурет):

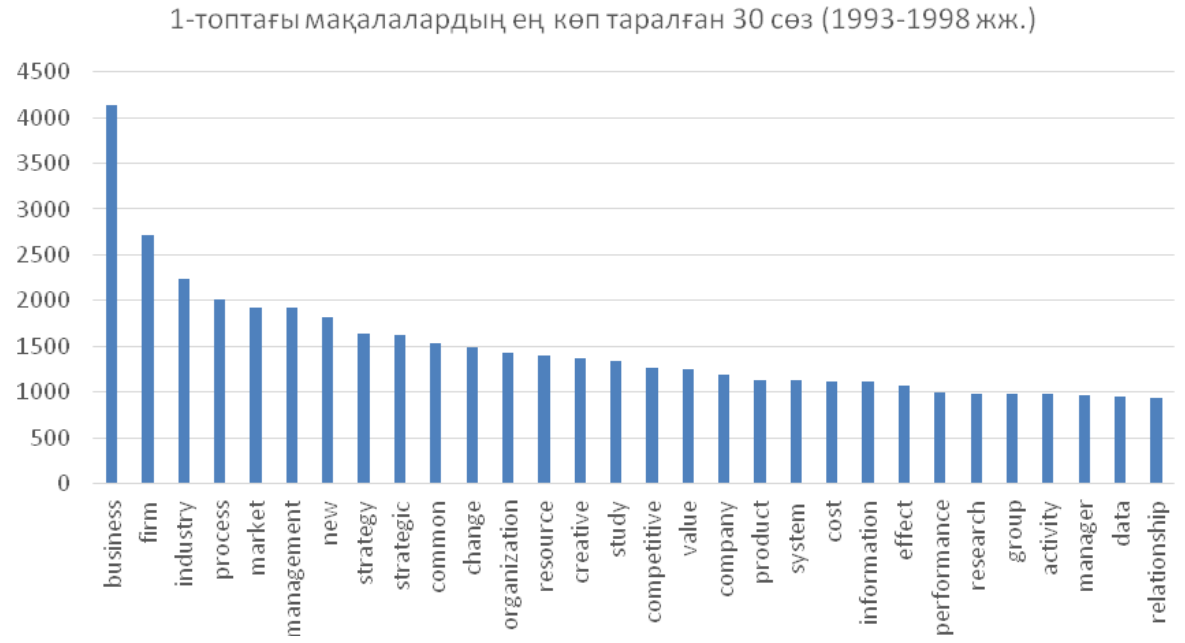



7-сурет. Лемматизация процесіне дейін 1-ші мақалалар тобындағы (1993–1998) ең жиі кездесетін 30 сөз

*Ескерту: Visual Studio коды негізінде авторлармен құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/.*

Мұнда біз «бизнес» сөзінің жиі кездесетінін көреміз. «Менеджмент» және «өнеркәсіп» сөздері қолдану жиілігі бойынша екінші және үшінші орында. Бұл тізімде тоқтату сөздері жоқ. Дегенмен, графикте бизнес (шамамен 3500) және бизнестер (шамамен 750) сөздерінің мағынасы бірдей болғанымен, бөлек есептелетіні анық. Сондай-ақ «фирмалар» сөзін (шамамен 1700) және «фирма» сөзін (1000-нан сәл артық) бірге санау керек екенін көруге болады. Мұны түзету үшін біз мәтінді өңдеуде лексиканы қалыпқа келтіру деп аталатын тағы бір маңызды операцияны қолдана аламыз. Лексиканы қалыпқа келтіру процесі мәтіндік шудың басқа түрін ескереді. Мысалы, ағылшын тіліндегі words connection, connected және connecting сөздерін бір connect сөзіне біріктіруге болады. Ол мұны сөздің барлық туынды байланысты нұсқаларын олардың жалпы негізгі терминіне дейін азайту арқылы жасайды. Әдетте лексиканы қалыпқа келтірудің екі әдісі бар. Бұл стемминг және лемматизация. Стемминг — бұл сөздерді түбір сөзіне дейін азайтатын немесе сөзжасамдық аффикстерді кесетін лингвистикалық қалыпқа келтіру әдісі. Бұл процесс сөздердің мағынасын олардың түбір сөзіне дейін төмендетеді. Лемматизация — бұл сөздерді лингвистикалық тұрғыдан дұрыс леммалар болып табылатын түбір сөзіне келтіру процесі. Ол мұны негізгі терминді өзгерту үшін сөздік және морфологиялық талдау сияқты әдістерді қолдану арқылы жасайды. Көп жағдайда лемматизация стеммингке қарағанда жетілдірілген процесс болып саналады. Стеммер қоршаған мәтінді ескермей, әр сөзді өз бетінше талдайды. Мысалы «better» сөзі оның леммасы ретінде қызмет ететін «good» сөзінен шыққан, өйткені ол үшін сөздіктен бірдеңе іздеу керек. Бұл объект лемматизация үшін негіз құру процесінен өтпейді.

Лемматизация процесінен кейін 1-топтағы ең көп кездесетін сөздердің тізімі өзгерді, бұл 8-суретте көрсетілген. «Бизнес» сөзі әлі де ең көп таралған, бірақ оның саны шамамен 3400-ден 4100 сөзге дейін өсті. «Фирмалар» сөзі жиілігі бойынша 5-ші орында (шамамен 1700 сөз), ал «фирма» сөзі 15-нші орында (1000 сөзден сәл артық). Лемматизациядан кейін бұл екі сөз «фирма» сөзіне біріктіріліп, жалпы саны 2700 сөзден тұратын 2-нші орынға ие болды. Графиктерде басқа сөздердің саны мен позицияларындағы өзгерістерді де көруге болады. Сондай-ақ, кейбір сөздердің тіркесіміне байланысты лемматизацияланған тізімде жаңалары пайда болды (8-сурет).




8-сурет. Лемматизация процесінен кейін 1-топтағы мақалалардың (1993–1998) ең көп таралған 30 сөзі

*Ескерту: Visual Studio коды негізінде авторлармен құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/*

Бұл диаграмма 1993 жылдан 1998 жылға дейінгі мақалаларда кездесетін ең танымал сөздерді көрсетеді. Бір кестеден қорытынды жасау қиын, сондықтан біз бұл операцияны барлық басқа кезеңдермен жасадық және оларды салыстыруға және қолданылатын сөздердің танымалдылық тенденциясын анықтауға мүмкіндік беретін тағы 5 кесте алдық. Біз мәтінді өңдеу технологиясын қолданбай-ақ байқау қиын болатын пайдалы және кейде күтпеген ақпаратты ала аламыз.

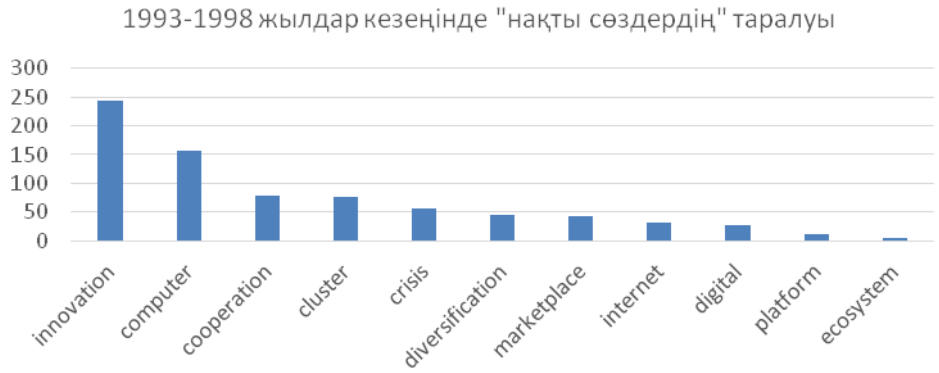
Алайда бұл бағдарламаның барлық мүмкіндіктері емес. Кейбір жағдайларда біз әр кезеңде қанша нақты сөз кездесетінін білгіміз келеді. Графикте кластер, экожүйе, әртараптандыру, платформа және т.б. сияқты сөздер жоқ. Бұл осы кезеңдегі ғалымдардың көпшілігі (1993 жылдан 1998 жылға дейін) бұл ұғымдар мен процестерді өздерінің зерттеулерінің негізгі пәндері ретінде қарастырмағанын білдіреді. Екінші жағынан, бұл ұғымдар мен процестер мүлдем қарастырылмаған дегенді білдірмейді. Осындай нақты сөздердің санын анықтау осы зерттеу саласындағы тенденцияларды түсіну үшін пайдалы болуы мүмкін. Ағымдағы мақала үшін «нақты сөздер» бөліміне келесі сөздер енгізілді: «кластер», «экожүйе», «платформа», «әртараптандыру», «дағдарыс», «интернет», «цифрлық», «компьютер», «ынтымақтастық», «сауда алаңы», «жоғары технологиялар», «инновация», «біріктіру», «сатып алу». Кодтың келесі жолдарын қолдана отырып, біз бұл сөздердің осы кезеңде қанша рет пайдаланылғанын біле аламыз (9-сурет):

```
filtered_words = []
find_words = [
        "business",
        "change",
        "common",
        "consumer",
        "cost",
        "creative",
        "customer",
        "effect",
        "group",
        "industry",
        "innovation",
        "market",
        "new",
        "platform",
        "price",
        "process",
        "product",
        "research",
        "service",
        "study",
        "value",
    ]
found_words = []
print("......Finding words is in the process........")
total_num = 0
for w in lemmed_words:
    if w in find_words:
        found_words.append(w)
        total_num += 1
print(total_num)
frequency_findwords = FreqDist(found_words)
frequency_findwords.plot(cumulative=False)
print(frequency_findwords.most_common)
print(".........Finding words has been finished.......")
```

9-сурет. Сөз іздеу функциясын құру және енгізу

*Ескерту: Visual Studio коды негізінде авторлармен құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/*

Нәтижесінде біз келесі диаграмманы аламыз (10-сурет):

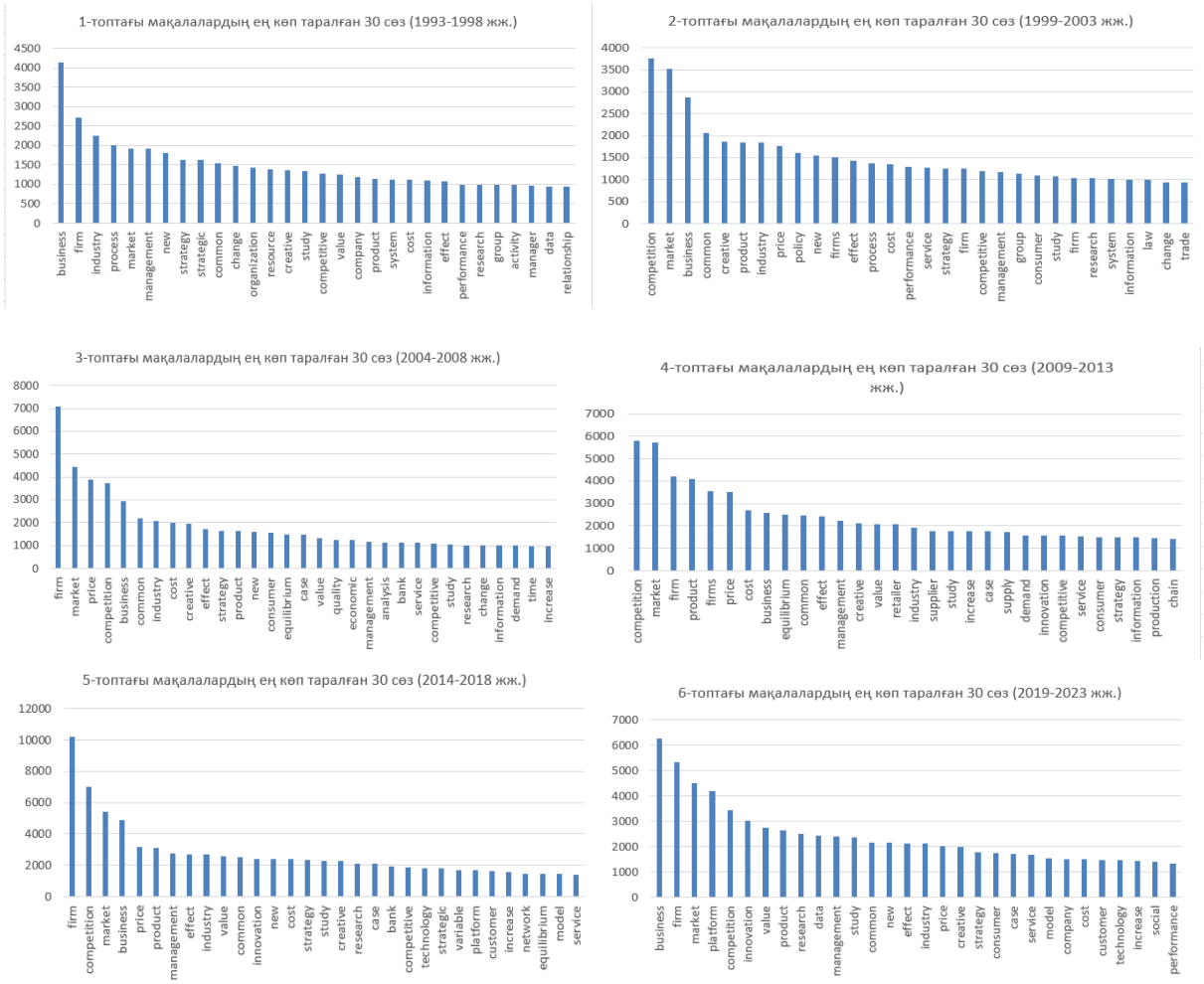



10-сурет. 1-топтағы мақалалардағы нақты сөздердің таралуы (1993–1998)

*Ескерту: Visual Studio коды негізінде авторлармен құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/.*

Жоғарыдағы графиктен 14 сөздің 11-і 1993–1998 жылдар аралығында пайда болғанын көруге болады. Ең көп таралған сөз — инновация. Сондай-ақ, ынтымақтастық пен кластер сөздердің үшінші және төртінші жиілігі болды, ал платформа мен экожүйе сөздері ең аз таралған. «Әртараптандыру» сөзі ортасында орналасқан және 50 бірліктен аз.

Келесі кезеңде біз бұл операцияны барлық басқа кезеңдерге қолдандық (2004–2008, 2009–2013, 2014–2018 және 2019–2023) және тағы 5 кесте алдық, бұл оларды салыстыруға және қолданылатын сөздердің танымалдылық тенденциясын анықтауға мүмкіндік берді (11-сурет).




11-сурет. Барлық алты топтағы мақалалардың ең көп таралған 30 сөзді салыстыру

*Ескерту: Visual Studio коды негізінде авторлармен құрастырған (талдау үшін Wiley онлайн кітапханасының мақалалары пайдаланылған), https://onlinelibrary.wiley.com/*

Диаграмма әр құжатта жиі қолданылатын терминдер тобын көрсетеді. Айта кету керек, барлық дерлік құжаттарда бизнес, фирма, бәсекелестік, нарық және өнім басым сөздер болып табылады.

Бірінші кезеңде (1993–1998) қарым-қатынас, ресурс, қызмет, ақпарат, жүйе және процесс сияқты сөздер әдетте бизнес пен технологияның әртүрлі аспектілерін сипаттау үшін пайдаланылды. Бұл сөздер сол кездегі бизнес ландшафтын сипаттайтын тиімділікке, өнімділікке және автоматтандыруға баса назар аударады. Компаниялар тиімділік пен рентабельділікке қол жеткізу үшін өз процестерін ретке келтіру және жүйелерін оңтайландыру жолдарын іздеді. Ішкі процестер мен құрылымдарға баса назар аударылды және график мұны көрсетеді. Алайда, соңғы жылдары (2018–2023) біз ғалымдардың бизнес пен бәсекелестік туралы жазуында айтарлықтай өзгерісті байқай аламыз. «Әлеуметтік», «технология», «клиент», «инновация» және «платформа» сияқты сөздер ескі терминдерді ауыстырып, ынтымақтастыққа, креативтілікке және клиентке бағдарлануға жаңа назар аударды. Бұл ауысымның негізгі қозғаушы күштерінің бірі әлеуметтік медиа мен цифрлық платформалардың дамуы болды. Әлеуметтік медиа компаниялардың бизнесті жүргізу тәсілін өзгертті. Бүгінгі таңда компаниялар әлеуметтік тұрғыдан көбірек хабардар және жауап беруі керек, сонымен қатар клиенттермен мағыналы түрде қарым-қатынас жасауы керек. Бұл цифрлық дәуірдегі қарым-қатынас пен қарым-қатынастың маңыздылығын көрсететін әлеуметтік және клиентке бағытталған тәсілге жаңа назар аударуға әкелді. Осы ауысымда технология да маңызды рөл атқарды. Технологиялық өзгерістердің жылдам қарқыны инновациялар мен толқулар үшін жаңа мүмкіндіктер туғызды және бұл біздің

бизнес пен қоғам туралы түсінігімізді өзгертті. Бүгінгі таңда технология енді процестерді автоматтандыру және тиімділікті арттыру құралы ғана емес; бұл инновация мен шығармашылықтың қозғаушы күші. Нәтижесінде инновация және платформа сияқты сөздер бәсекелестікті зерттейтін ғалымдардың басты тақырыптарының біріне айналды, бұл технологияның бизнестің өсуі мен жетістігін ынталандырудағы маңыздылығын көрсетті. Бұл ауысымға ықпал ететін тағы бір фактор — жұмыс және жұмыспен қамтудың өзгеретін сипаты. Бүгінгі таңда көптеген адамдар экономика саласында фрилансерлер немесе тәуелсіз мердігерлер ретінде жұмыс істейді. Бұл басқалармен жұмыс істеуді және ынтымақтастықты ұйымдастырудың жаңа тәсілдеріне, платформалар мен желілерге қажеттілік туғызды. Ғалымдардың еңбектерінде қолданатын сөздер бұл өзгерісті көрсетеді: платформа және желі сияқты сөздер уақыт өте келе ғалымдар арасында кең таралуда.

Бұған дейін біз әр кезеңді жеке қарастырдық. Дегенмен, диаграммалардағы өзгерістерді көрнекі ету үшін біз барлық алты кезеңдегі ең жиі кездесетін сөздерді салыстыруымыз керек. Біз мұны MS Excel бағдарламалық жасақтамасын (1-кесте) пайдаланып кесте құру арқылы жасай аламыз. Бұл кестені құру үшін алты кезеңнің кем дегенде екі кезеңінде үздік 30-ға кіретін сөздер таңдалды. Содан кейін бұл сөздер 9-суретте берілген кодты пайдаланып алфавиттік ретпен сұрыпталды. Қарастырылып отырған барлық кезеңдерде осы сөздердің қайталану саны анықталды. Соңғы бағанда сөздердің өсу немесе құлдырау динамикасы көрсетілген.

1-кесте. Барлық алты құжатта жиі кездесетін сөздерді салыстыру

| Words | 1 | 2 | 3 | 4 | 5 | 6 | Graph |
|---|---|---|---|---|---|---|---|
| business | 4142 | 2866 | 2936 | 2581 | 4873 | 6264 | |
| change | 1485 | 947 | 996 | 990 | 1287 | 1170 | |
| common | 1536 | 2064 | 2179 | 2475 | 2514 | 2167 | |
| consumer | 164 | 1098 | 1543 | 1507 | 1308 | 1729 | |
| cost | 1114 | 1354 | 1977 | 2696 | 2389 | 1499 | |
| creative | 1376 | 1860 | 1952 | 2108 | 2264 | 1973 | |
| customer | 695 | 602 | 750 | 967 | 1616 | 1484 | |
| effect | 1071 | 1432 | 1703 | 2410 | 2692 | 2117 | |
| group | 982 | 1126 | 575 | 1235 | 1044 | 1326 | |
| industry | 2242 | 1835 | 2076 | 1913 | 2672 | 2107 | |
| innovation | 243 | 635 | 381 | 1574 | 2427 | 3025 | |
| market | 1926 | 3517 | 4434 | 5732 | 5427 | 4502 | |
| new | 1812 | 1557 | 1585 | 1360 | 2406 | 2161 | |
| platform | 11 | 11 | 84 | 336 | 1678 | 4199 | |
| price | 774 | 1772 | 3869 | 3514 | 3181 | 2021 | |
| process | 2019 | 1366 | 783 | 769 | 1213 | 1063 | |
| product | 1133 | 1841 | 1652 | 4102 | 3136 | 2648 | |
| research | 989 | 1038 | 1024 | 1346 | 2131 | 2492 | |
| service | 797 | 1279 | 1120 | 1516 | 1411 | 1671 | |
| study | 1337 | 1073 | 1058 | 1766 | 2282 | 2364 | |
| value | 1248 | 871 | 1310 | 2076 | 2589 | 2738 | |

Ескерту-MS Excel негізінде авторлар құрастырған (Wiley онлайн кітапханасындағы мақалалар талдау үшін пайдаланылды ), https://onlinelibrary .wiley.com/

Осы кестеге сүйене отырып, «бизнес», «жалпы», «тұтынушы», «шығармашылық», «клиент», «инновация», «нарық», «зерттеу», «оқу» және «құндылық» сияқты сөздер оларды қолданудың тұрақты өсу тенденциясын көрсетеді, яғни олардың пәндік саладағы маңыздылығы артып келеді. Бұл тенденцияның негізгі қозғаушы күштерінің бірі — тұтынушылардың мінез-құлқы мен қалауына көбірек көңіл бөлу. Компаниялар клиентке көбірек көңіл бөлгендіктен, олар өз клиенттерін жақсы түсіну үшін зерттеулерге қомақты қаражат сала бастады. Бұл тұтынушы, тапсырыс беруші және нарық сияқты сөздерді қолданудың күрт өсуіне әкелді, өйткені зерттеушілер тұтынушылардың мінез-құлқының негізгі факторларын анықтауға және жаңа нарықтық мүмкіндіктерді анықтауға тырысады.

Бұл үрдіске ықпал ететін тағы бір фактор — технологиялық өзгерістер мен инновациялардың жылдам қарқыны. Жаңа технологиялардың пайда болуымен бизнестен бәсекеге қабілетті болу үшін шығармашылықпен айналысу және жаңашылдық қажет (2-кесте).

2-кесте. Әр кезеңде жиі кездесетін сөздердің пайызы

| Words | 1 | 2 | 3 | 4 | 5 | 6 | Sum |
|---|---|---|---|---|---|---|---|
| business | 17,50 | 12,11 | 12,41 | 10,91 | 20,59 | 26,47 | 23662 |
| change | 21,60 | 13,77 | 14,49 | 14,40 | 18,72 | 17,02 | 6875 |
| common | 11,87 | 15,96 | 16,85 | 19,13 | 19,44 | 16,75 | 12935 |
| consumer | 2,23 | 14,94 | 21,00 | 20,51 | 17,80 | 23,53 | 7349 |
| cost | 10,10 | 12,28 | 17,93 | 24,44 | 21,66 | 13,59 | 11029 |
| creative | 11,93 | 16,13 | 16,93 | 18,28 | 19,63 | 17,11 | 11533 |
| customer | 11,37 | 9,85 | 12,27 | 15,82 | 26,43 | 24,27 | 6114 |
| effect | 9,37 | 12,53 | 14,91 | 21,09 | 23,56 | 18,53 | 11425 |
| group | 15,62 | 17,91 | 9,14 | 19,64 | 16,60 | 21,09 | 6288 |
| industry | 17,45 | 14,29 | 16,16 | 14,89 | 20,80 | 16,40 | 12845 |
| innovation | 2,93 | 7,66 | 4,60 | 19,00 | 29,29 | 36,51 | 8285 |
| market | 7,54 | 13,77 | 17,36 | 22,44 | 21,25 | 17,63 | 25538 |
| new | 16,65 | 14,31 | 14,57 | 12,50 | 22,11 | 19,86 | 10881 |
| platform | 0,17 | 0,17 | 1,33 | 5,32 | 26,55 | 66,45 | 6319 |
| price | 5,12 | 11,71 | 25,57 | 23,22 | 21,02 | 13,36 | 15131 |
| process | 27,99 | 18,94 | 10,86 | 10,66 | 16,82 | 14,74 | 7213 |
| product | 7,81 | 12,69 | 11,38 | 28,27 | 21,61 | 18,25 | 14512 |
| research | 10,96 | 11,51 | 11,35 | 14,92 | 23,63 | 27,63 | 9020 |
| service | 10,23 | 16,41 | 14,37 | 19,45 | 18,10 | 21,44 | 7794 |
| study | 13,53 | 10,86 | 10,71 | 17,87 | 23,10 | 23,93 | 9880 |
| value | 11,52 | 8,04 | 12,09 | 19,17 | 23,90 | 25,28 | 10832 |

*Ескерту: MS Excel негізінде авторлар құрастырған (Wiley онлайн кітапханасындағы мақалалар талдау үшін пайдаланылды), https://onlinelibrary.wiley.com/*

Нәтижесінде инновация, шығармашылық және құндылық сияқты сөздерді қолдану күрт өсті, өйткені зерттеушілер инновациялық өнімдер, қызметтер және бизнес үлгілері арқылы тұтынушыларға құндылық жасаудың жаңа жолдарын зерттеп жатыр. Бұл сөздерді таратуда нарықтардың жаһандануы да маңызды рөл атқарды. Кәсіпорындар қазіргі уақытта жаһандық контексте жұмыс істейтіндіктен, олар әртүрлі мәдениеттер мен нарықтық жағдайларды түсініп, оларды шарлай білуі керек. Бұл зерттеулерге көбірек көңіл бөлуге әкелді, өйткені компаниялар бүкіл әлем бойынша әртүрлі нарықтар мен тұтынушылардың мінез-құлқы туралы түсінік алуға тырысады. Бұл сөздерді қолданудың артуы сонымен қатар деректерге бағытталған және ғылыми негізделген шешім қабылдау тәсіліне деген кең қоғамдық тенденцияны көрсетеді. Үлкен деректер мен талдаулардың

пайда болуымен кәсіпорындар енді өз шешімдерін негіздеу үшін пайдалануға болатын деректердің үлкен көлеміне қол жеткізе алады. Бұл зерттеулерге көбірек назар аударуға әкелді, өйткені компаниялар тұтынушылардың мінез-құлқы, нарық тенденциялары және пайда болатын мүмкіндіктер туралы ақпарат алу үшін деректерді пайдалануды мақсат етеді.

Екінші жағынан, баға, процесс және өзгеріс сияқты сөздер оларды қолданудың төмендеу тенденциясын көрсетеді, бұл олардың бәсекеге қабілеттілікке маңыздылығының төмендеуін көрсетеді. «Баға» сөзін қолданудың төмендеуінің негізгі себептерінің бірі — тұтынушылық құндылыққа назар аударудың артуы. Қазіргі бизнес ортасында баға клиенттер сатып алу туралы шешім қабылдаған кезде ескеретін көптеген факторлардың бірі ғана. Компаниялар өздерін тек баға бойынша бәсекелесудің орнына клиенттерге ұсынатын құндылығымен саралауға ұмтылуда. Бұл екпіннің өзгеруі ғылыми зерттеулерде «баға» сөзіне баса назар аударудың төмендеуіне әкелді. Сол сияқты «процесс» сөзінің азаюын бизнестегі инновация мен икемділікке баса назар аударумен түсіндіруге болады. Дәстүрлі бизнес-процестер көбінесе көлемді және баяу болып саналады және инновация мен прогреске кедергі ретінде қарастырылады. Нәтижесінде, кәсіпорындар өз процестерін оңтайландыру және икемді болу жолдарын іздеуде, бұл ғылыми зерттеулердегі «процесс» сөзіне баса назар аударудың төмендеуіне әкелді. Ақырында, «өзгеріс» сөзін қолданудың төмендеуін бірқатар факторлармен түсіндіруге болады, соның ішінде бизнестің өзгеру қарқынының артуы және бизнестің өзгеретін нарықтық жағдайларға бейімделу және жауап беру қажеттілігі. Кәсіпорындар серпінді және жауап беретін бола бастағанда, өзгерістер тұжырымдамасы олардың күнделікті қызметіне көбірек еніп, осы тақырып бойынша арнайы зерттеулерге деген қажеттіліктің төмендеуіне әкеледі.

Соңында біз әр кезеңдегі сөздердің қайталану пайызын есептеуді шештік. Мұны істеу үшін біз бүкіл кезеңдегі сөздердің қосындысын есептедік және әр сөздің жалпы сомасына қатысты әр кезеңнің үлесін есептедік. Содан кейін шартты пішімдеу мүмкіндігін пайдаланып, өзгерістерді жақсырақ визуализациялау үшін кестені боядық. Сөздердің ең аз үлесі бар ұяшықтар қызыл түске боялған, ал сөздердің ең көп үлесі бар ұяшықтар жасыл түске боялған. Жалпы, кестеден «өзгеріс», «процесс», «шығындар» және «баға» сөздерінен басқа барлығының пропорционалды түрде өскенін көруге болады. Бұл олардың барлық 3 онжылдықта өзектілігін көрсетеді. «Платформа» сөзі ең күрт өсуімен ерекшеленді. Бір қызығы, бастапқы кезеңде жалпы соманың 0,17 % ғана пайдаланылды. Алайда, кестеден көріп отырғанымыздай, соңғы екі кезеңде сөздердің үлесі алдымен 26,55-ке, содан кейін 66,45 %-ға дейін өсті. Сонымен қатар, «платформа» сөзі соңғы кезеңдегі ең танымал төртінші сөз болды. Соңғы үш онжылдықта «платформа» сөзін қолданудың күрт өсуін цифрлық платформалардың өсуімен, платформалық бизнес модельдерінің маңыздылығының артуымен және терминнің икемділігі мен қолданылуымен түсіндіруге болады. Цифрлық технологиялар дамып келе жатқандықтан және кәсіпорындар платформалық бизнес үлгілерін көбірек енгізіп жатқандықтан, алдағы жылдары ғылыми зерттеулерде «платформа» сөзін қолдану кеңейе береді деп күтуге болады.

***Талқылау***

Заманауи технологиялардың қарқынды дамуымен жаңа компьютерлік және интернет қосымшалары бейне, фото, мәтін, дауыстық және әлеуметтік медиа деректері сияқты бұрын-соңды болмаған жылдамдықпен үлкен көлемдегі деректерді шығарады. Бұл деректер көбінесе жоғары өлшемді сипаттамаларға ие, бұл деректерді талдау және шешім қабылдау үшін үлкен қиындық тудырады. Әдістерді дұрыс таңдау олардың көп өлшемді деректерді өңдеудегі және аналитикалық компоненттің тиімділігін арттырудағы тиімділігін көрсетеді (Мезенцева, Коломиец, 2020).

Функциялар мен әдістерді таңдау артық және маңызды емес функцияларды жою кезінде деректерді өңдеуді қысқартуда маңызды рөл атқарады. Белгілерді таңдау әдісі талдау алгоритмдерін алдын-ала өңдеуге, сондай-ақ R бағдарламасының көмегімен нәтижелердің дәлдігін жеңілдетуге және жақсартуға мүмкіндік береді (Мастицкий, Шитиков, 2014).

Соңғы онжылдықта көптеген компаниялар мәтіндік ақпаратты өңдеуге арналған арнайы бағдарламалық жасақтаманы әзірлеуге кірісті. Біз мыналарды атап өтеміз: Google, IBM, SAS, Angoss Software Corporation және т.б. «R бағдарламасы ең қол жетімді, өйткені басқа бағдарламалардың жұмысында бірқатар кемшіліктер бар», – дейді Д.Б. Ковтун. Мысалы, Google бағдарламаларында құрылымдалмаған деректерді талдауға шектеулер бар, ал Google бағдарламалық жасақтамасы еркін қол жетімді емес (Ковтун, 2021). Тақырыптық модельдеу құжаттың терминдік матрицасына түрлендірілген мәтіндік деректерге арналған машиналық оқыту алгоритмдерін қолданудың кең класын білдіреді.

Тақырыптық модельдер — бұл «мәтіндік құжаттар корпусындағы жасырын тақырыптарды анықтауға және өлшеуге бағытталған статистикалық алгоритмдер». Тақырыптық модельдер екі топқа бөлінеді. Біріншісі тек бір тақырыпты қамтитын құжаттарды қамтиды (бір мүшелік модельдер). Екіншіден, бірнеше тақырыпты қамтитын құжаттар (аралас мүшелік модельдері). Әр құжатта тек бір тақырып болуы мүмкін деген модельдер, мысалы, кластерлік талдауды қолдану арқылы жүзеге асырылады (k-орташа, k-медианалар және т.б.). Дегенмен, әр құжатта көптеген тақырыптар болуы мүмкін деген модельдер танымал болды. Қазіргі уақытта көптеген тақырыптық модельдер бар: Дирихлеттің классикалық жасырын орналасуы (LDA), корреляцияланған тақырыптық модельдер, динамикалық тақырыптық модельдер, иерархиялық тақырыптық модельдер және құрылымдық тақырыптық модельдер (Шипунов және басқалар, 2014).

***Қорытынды***

Біз соңғы 30 жыл ішінде бәсекелестік жұмыстарды жазған ғалымдар негізінен нарық, баға, зерттеу, шығармашылық, өнеркәсіп және т.б. мәселелерін көтерді деген қорытындыға келдік. Сондай-ақ тұтынушыларды, инновацияларды және өнімдерді зерттеуде үлкен өсім байқалды. Ең жоғары өсу 1993–1998 жылдардағы таңдауда небәрі 11 рет кездескен платформа сөзімен байланысты болды, бірақ 2019–2023 жылдардағы таңдауда бұл сөз 4199 рет қолданылды, бұл 38 000 %–дан астам өсуді көрсетті.

Бүгінгі таңда біз тиімділік пен автоматтандырудан гөрі ынтымақтастыққа, креативтілікке және клиентке бағдарлануға көбірек көңіл бөлеміз. Әлеуметтік медианың, технологияның және жұмыстың жаңа түрлерінің дамуы бұл өзгеріске, сондай-ақ ұйымдастыру мен ынтымақтастықтың жаңа тәсілдеріне деген қажеттілікке ықпал етті.

Соңғы үш онжылдықтағы ғылыми мақалалардағы «бизнес», «әдеттегі іс», «тұтынушы», «креатив», «клиент», «инновация», «нарық», «зерттеу», «оқу» және «құндылық» сияқты сөздерді қолданудың өсуі бизнес пен экономиканың өзгеретін сипатын көрсетеді. Клиенттердің мінез–құлқы мен қалауына назар аударудың артуы, технологиялық өзгерістер мен инновациялар, жаһандану және деректерге негізделген шешім қабылдаудың өсуі сияқты факторлар осы тенденцияға ықпал етті.

Екінші жағынан, тұтынушылық құндылыққа көбірек көңіл бөлу, инновациялар мен икемділікке баса назар аудару және бизнес үшін икемді болу және өзгерістерге жауап беру қажеттілігі баға, процесс және ғылыми мақалалардағы өзгерістер сияқты сөздерді қолданудың төмендеуіне әкелді.

***Әдебиеттер тізімі***


Atenstaedt R. Word cloud analysis of the BJGP / R. Atenstaedt // British Journal of General Practice. — 2012. — Vol. 62(596). — P. 148. DOI: https://doi.org/10.3399/bjgp12X630142.

Atenstaedt R. Word cloud analysis of the BJGP: 5 years on / R. Atenstaedt // British Journal of General Practice. — 2017. — Vol. 67(658). — P. 231–232. DOI: https://doi.org/10.3399/bjgp17X690833.

Cai J. Feature selection in machine learning: A new perspective / J. Cai, J. Luo, S. Wang, S. Yang // Neurocomputing. — 2018. — Vol. 300. — P. 70–79. DOI: 10.1016/j.neucom.2017.11.077.

Kim S. Research paper classification systems based on TF-IDF and LDA schemes / S. Kim, J. Gil // Human-centric Computing and Information Sciences. — 2019. — N 9(1). — P. 1–21. DOI: 10.1186/s13673–019–0192-.

Mezentseva O. Optimization of analysis and minimization of information losses in text mining / O. Mezentseva, A. Kolomiiets // Herald of Advanced Information Technology. — 2020. — 3(1). — P. 373–382.

Udgave A. Text Mining and Text Analytics of Research Articles / A. Udgave, P. Kulkarni // Palarch's Journal Of Archaeology Of Egypt/Egyptology. — 2020. — 17(6). — P. 1–7.

Verzani J. Getting started with RStudio / J. Verzani // O'Reilly Media. — 2017. — 98 p.

Wiley Online Library. — [Electronic resource]. — Access mode: https://onlinelibrary.wiley.com/.

Большакова Е. И. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пос. / Е. И. Большакова, К. В. Воронцов, Н. В. Лукашевич, А. С. Сапин. — М.: НИУ ВШЭ, 2017. — 268 с.

Кабаков Р. R в действии. — [Электронный ресурс] / Р. Кабаков. — Режим доступа: https://www.manning.com.

Калабин А. Л. Анализ информационных критериев отбора значимых признаков в методах *Text Mining* / А. Л. Калабин, Е. И. Корнеева // Вестн. Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. — 2020. — № 2. — С. 150–159.

Ковтун Д. Б. Исследование внутриведомственного взаимодействия органов власти РФ на основе документов стратегического планирования с помощью технологии *Text Mining* / Д. Б. Ковтун // Моск. экон. журн. — 2021. — № 2. — С. 1–10.

Мастицкий С. Э. Статистический анализ и визуализация данных с помощью R / С. Э. Мастицкий, В. К. Шитиков. — 2014. — 401 с.

Нгуен М. Т. Тестирование методов машинного обучения в задаче классификации *http* запросов с применением технологии TFIDF / М. Т. Нгуен // Вестн. Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. — 2019. — № 4. — С. 119–131.

**Т.К. Шурен, Т.П. Притворова, Н.П. Иващенко, Г.А. Райханова[4]**

## Применение технологии интеллектуального анализа текста для сравнительного анализа тенденций в области конкуренции

***Аннотация:***

*Цель:* Основной целью данной статьи является анализ конкуренции в работах ученых с целью выявления положительных и отрицательных тенденций, а также характерных траекторий, путем создания и применения программы интеллектуального анализа текста на *Python.*

*Методы:* В статье мы использовали современные методы обработки текстов, в частности интеллектуальный анализ текста, язык программирования *Python*, программное обеспечение *MS Excel*, а также данные с официального сайта онлайн-библиотеки *Wiley Online Library по адресу https://onlinelibrary.wiley.com/*.

*Результаты:* Стремительное развитие современных технологий, интернет-приложений сопровождается генерацией больших объемов данных, своевременная обработка которых сегодня является одной из главных проблем в различных сферах жизни — социальной, экономической, политической и других. В решении этой глобальной проблемы на помощь приходят современные методы обработки текстовой информации, так называемые технологии интеллектуального анализа текста. Эти инструменты позволяют повысить эффективность решения задач разного уровня. Алгоритмы, встроенные в технологию интеллектуального анализа текста, раскрывают основные понятия текста, его содержание и взаимосвязь между ними.

*Выводы:* Мы обнаружили, что за последние 30 лет ученые, написавшие работы в области конкуренции, в основном поднимали проблемы рынка, цен, исследований, творчества, промышленности и другие. Также значительно возросло количество работ, связанных с изучением потребителей, инноваций и продуктов. Наибольший рост наблюдался относительно слова *platform*, которое встречалось в сборнике 1993–1998 гг. всего 11 раз, а в сборнике 2019–2023 годов — 4199 раз, что показало увеличение более чем на 38 000 %.

***Ключевые слова:*** интеллектуальный анализ текста, конкуренция, платформа, бизнес, цифровизация, потребители.

**T.K. Shuren, T.P. Pritvorova, N.P. Ivashchenko, G.A. Raikhanova**

## Application of text mining technology for comparative analysis of trends in the field of competition

***Abstract***

*Object:* The main purpose of this article is to analyze the competition in the works of scientists to identify positive and negative trends, as well as characteristic trajectories, by creating and applying a text mining program in python.

*Methods:* For this study, we used modern methods of word processing, in particular text mining. We used python programming language and software MS Excel. We used data from the official website of the Wiley Online Library at https://onlinelibrary.wiley.com/

*Findings:* The rapid development of modern technologies, Internet applications is accompanied by the generation of large amounts of data, the timely processing of which is today one of the main problems in various spheres of life — social, economic, political, and others. In solving this global problem, modern methods of processing text information, the so-called text mining technologies, come to the rescue. These tools allow to increase the efficiency of solving problems of different levels. The algorithms embedded in the text mining technology reveal the basic concepts of the text, the content and the relationship between them.

The integration of modern text mining systems and Python programming language makes it possible to conduct research in the field of text analysis and processing. These systems, using statistical methods, process the rating of news documents, materials of scientific documents, blogs, tweets, emails, advertisements and other information. The main task of text analysis is to get a clear idea about the topics of interest, to extract important information. For the analysis of texts, 600 articles from Wiley Online Library in English in PDF format were selected, including information on trends in the business and competition for 1993–2023.

*Conclusions:* We have found that over the past 30 years, scientists who have written works in the field of competition mainly raised the problems of market, price, research, creativity, industry and others. There was also a huge increase in the study of consumers, innovations and products. The highest growth was associated with the word platform, which was found in the 1993–1998 compilation only 11 times, but in the 2019–2023 compilation the same word was used 4199 times, which showed an increase of more than 38,000 %.

***Keywords:*** text mining, competition, platform, business, digitalization, consumers.

***References***

Atenstaedt, R. (2017). Word cloud analysis of the BJGP: 5 years on. *British Journal of General Practice*, *67*(658), 231–232. DOI: https://doi.org/10.3399/bjgp17X690833.

Atenstaedt, R. (2018). Word cloud analysis of the BJGP. *British Journal of General Practice*, *62*(596), 148. DOI: https://doi.org/10.3399/bjgp12X630142.

Bolshakova, E. I., Vorontsov, K. V., Lukashevich, N. V., & Sapin, A. S. (2017). Avtomaticheskaia obrabotka tekstov na estestvennom yazyke i analiz dannykh [Automatic text processing in natural language and data analysis]. Moscow: NIU VShE [in Russian].

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79. DOI: 10.1016/j.neucom.2017.11.077.

Kabakov P. R v deistvii [R in action]. Retrieved from https://www.manning.com.

Kalabin, A. L. & Korneeva, E. I. (2020). Analiz informatsionnykh kriteriev otbora znachimykh priznakov v metodakh *Text Mining* [Analysis of information criteria for the selection of significant features in Text Mining methods]. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriia Sistemnyi analiz i informatsionnye tekhnologii — Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies*, 2, 150–159 [in Russian].

Kim, S. & Gil, J. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*, 9(1), 1–21). DOI: 10.1186/s13673–019–0192-.

Kovtun, D. B. (2021). Issledovanie vnutrivedomstvennogo vzaimodeistviia organov vlasti RF na osnove dokumentov strategicheskogo planirovaniia s pomoshchiu tekhnologii *Text Mining* [Research of interdepartmental interaction of the authorities of the Russian Federation on the basis of strategic planning documents using *Text Mining* technology]. *Moskovskii ekonomicheskii zhurnal — Moscow Economic Journal*, 2, 1–10 [in Russian].

Mastitskii, S. E. & Shitikov, V. K. (2014). Statisticheskii analiz i vizualizatsiia dannykh s pomoshchiu R [Statistical analysis and visualization of data using R]. Moscow [in Russian].

Mezentseva, O. & Kolomiiets, A. (2020). Optimization of analysis and minimization of information losses in text mining. *Herald of Advanced Information Technology*, *3*(1), 373–382).

Nguen, M. T. (2019). Testirovanie metodov mashinnogo obucheniia v zadache klassifikatsii *http* zaprosov s primeneniem tekhnologii TFIDF [Testing machine learning methods in the task of classifying *http* requests using TFIDF technology]. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriia Sistemnyi analiz i informatsionnye tekhnologii — Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies*, 4, 119–131 [in Russian].

Udgave, A. & Kulkarni, P. (2020). Text Mining and Text Analytics of Research Articles. *Palarch's Journal Of Archaeology Of Egypt*, *17*(6), 1–7.

Verzani, J. (2017). Getting started with RStudio. *O'Reilly Media*, 98.

Wiley Online Library. Retrieved from https://onlinelibrary.wiley.com/.