**Форма № Н-6.01у**

# Кременчуцький національний університет імені Михайла Остроградського

(повне найменування вищого навчального закладу)

# Кафедра автоматизації та інформаційних систем

(повна назва кафедри, циклової комісії)

**КУРСОВИЙ ПРОЄКТ (РОБОТА)**

# з навчальної дисципліни «Технології обробки природньої мови» \_\_\_\_

(назва дисципліни)

# на тему: «Статистична обробка текстових даних. Варіант 3»

Студента 5 курсу КН-21-2м групи  
Ступінь вищої освіти магістр

(бакалавр, магістр)

Спеціальність 122 «Комп’ютерні науки»   
Освітня програма «Інформаційні

управляючі системи та технології»

Вітюк І. В.

(прізвище та ініціали)

Керівник д. т. н., професор кафедри АІС

Шевченко І. В.

(посада, вчене звання, науковий ступінь, прізвище та ініціали)

Національна шкала \_\_\_\_\_\_\_\_\_\_\_

Кількість балів: Оцінка: ЄКТС \_\_\_

Члени комісії І. В. Шевченко

(підпис) (ініціали та прізвище)

А. М. Самойлов

(підпис) (ініціали та прізвище)

Н. В. Рилова

(підпис) (ініціали та прізвище)

# м. Кременчук 2022 рік

**Форма № Н-9.01у**

Кафедра автоматизації та інформаційних систем

Дисципліна «Технології обробки природньої мови»

Освітній ступінь «Магістр»

Спеціальність 122 «Комп’ютерні науки»

Освітня програма «Інформаційні управляючі системи та технології»\_\_\_\_\_\_\_\_\_\_

Курс 5 групи КН-22-2м семестр 2

**З А В Д А Н Н Я**

на курсовий проєкт (роботу) студенту

Вітюк Іван Володимирович

(прізвище, ім’я, по батькові)

1. Тема проєкту (роботи): «Статистична обробка текстових даних. Варіант 3»

2. Строк подання студентом проєкту (роботи): квітень 2023 р.

3. Вихідні дані до проєкту (роботи): інформація про предметну область, текстовий редактор Sublime Text, науково-технічні статті для обробки

4. Зміст пояснювальної записки (перелік питань, що підлягають розробці):\_\_\_\_\_

Дослідження предметної області, інформаційний пошук

Попередня обробка досліджуваного тексту

Статистична обробка тексту

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень)

6. Дата видачі завдання: лютий 2023 р.

**КАЛЕНДАРНИЙ ПЛАН**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| №  з/п | Назва етапів курсового проєкту | Терміни виконання етапів проєкту | Вказівки та  зауваження викладача (з зазначенням дати консультації) | Оцінювання етапів проекту | | |
| за націо- нальною шкалою | за шкалою ЄКТС | кількість балів |
| 1. | Дослідження предметної області, інформаційний пошук | 22.02.23 –  10.03.23 |  |  |  |  |
| 2. | Попередня обробка досліджуваного тексту | 10.03.23 –  25.03.23 |  |  |  |  |
| 3. | Статистична обробка тексту | 26.03.23 –  04.04.23 |  |  |  |  |
| 4. | Оформлення пояснювальної записки | 05.04.23 –  24.04.23 |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  | (усього балів не більше 60) | | |
|  | Подання проекту на захист | 25.04.23 |  | (не більше 40) | | |
|  | Разом |  |  |  |  |  |

Студент

(підпис)

Керівник \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_ І. В. Шевченко\_\_\_\_\_

(підпис) (ініціали та прізвище)

\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_ р.

**РЕФЕРАТ**

Пояснювальна записка до курсового проєкту містить: 53 сторінок, 14 рисункыв, 2 додатка, 19 використаних джерел.

Темою даного курсового проєкту є статистична обробка текстів.

Метою даного курсового проєкту є реалізація та дослідження методів виявлення ключових слів та контекстних множин.

Для досягнення мети в роботі було опрацьовано основні методи первинної та статистичної обробки текстових даних і інтерпретації результатів аналізу, а саме: зроблено аналіз проблемної області та сформульовано задачу дослідження; виконано первинну обробку тексту, а саме – видалення стоп-слів, аналіз тексту за Ципфом, виділення робочої підмножини важливих слів, стемінг деяких слів для приведення їх до єдиної словоформи.

Отримано ранжований список слів для подальшої статистичної обробки тексту; розраховано відносні частоти для підмножини важних слів по абзацах; розраховано коефіцієнти кореляції відносних частот слів вздовж абзаців; розраховано косинусну міру подібності зміни відносних частот вздовж абзаців; виділено ключові слова; сформовано контекстні підмножини слів-сателітів для окремих ключових слів, сформовані інформаційні ядра кожного абзацу, сформовано скорочений варіант тексту.

ПРИРОДНА МОВА, СТАТИСТИЧНА ОБРОБКА, КЛЮЧОВІ СЛОВА, КОНТЕКСТНА МНОЖИНА, ІНФОРМАЦІЙНЕ ЯДРО

**ЗМІСТ**

*Змн*.

Арк.

№ докум.

*Підпис*

Дата

Арк.

1

*122 – КП.2022.01.000 ПЗ*

*Розроб.*

*Луценко В. В.*

*Перевір.*

Шевченко І. В.

*Н. контр.*

Рилова Н. В.

*Затверд.*

*Статистична обробка текстів. Варіант 3*

*Пояснювальна записка*

Літ.

Аркушів

53

КрНУ Кафедра АІС

ВСТУП 2

1 АНАЛІЗ ПРОБЛЕМНОЇ ОБЛАСТІ ТА ФОРМУЛЮВАННЯ ЗАДАЧ ДОСЛІДЖЕННЯ 4

1.1 Основна проблема обробки природної мови 5

1.2 Статистична обробка текстів 7

1.3 Класифікація та кластеризація текстів 8

1.4 Автоматичне реферування та індексація 13

1.5 Змістовна близькість 14

1.6 Виділення термінологічних одиниць 15

Висновки до розділу 16

2 ПЕРВИННА ОБРОБКА ТЕКСТУ 17

2.1 Відомі статистичні закономірності у текстах та ключові слова 17

2.2 Видалення стоп-слів та стемінг 21

2.3 Формування робочої підмножини слів для статистичного аналізу 23

Висновки до розділу 25

3 СТАТИСТИЧНА ОБРОБКА ТЕКСТУ 26

3.1 Розрахунок відносних частот слів по абзацах 26

3.2 Розрахунок коефіцієнтів кореляції відносних частот слів вздовж абзаців 27

3.3 Визначення ключових слів 28

3.4 Формування контекстних підмножин для ключових слів 29

3.5 Визначення інформаційного ядра кожного абзацу 29

3.6 Складання скороченого тексту на основі ядер абзаців 35

Висновки до розділу 35

ВИСНОВКИ 36

СПИСОК ЛІТЕРАТУРИ 37

Додаток А Текст, що підлягає аналізу 39

Додаток Б Скорочений текст 45

**ВСТУП**

Дослідження мови в першу чергу викликано інтересом до природного інтелекту людино, тому така форма існування мови як письмовий текст, також є об’єктом пильної уваги. Спостерігається стійкий інтерес до робіт, що проводяться в цій області.

Лексико-статистичний аналіз текстів виконується у багатьох дослідженнях. У літературознавстві частотні списки лексики використовуються для складання енциклопедичних словників письменників та поетів. Порівняльний аналіз частотних списків лексики художніх творів різних авторів допомагає визначити особливості стилю. Одержану інформацію можна використати для визначення авторства тексту. Найінтенсивніше використання частотних списків спостерігається під час обробляння фахових текстів, які є найпоширенішою формою представлення науково-технічної (н/т) інформації в Інтернеті.

Завдання по автоматичній обробці тексту, включають в себе завдання інформаційного пошуку, пов’язані з перебуванням інформації в електронних базах знань, автоматичним вилученням даних і фактів по заданій тематику, автоматичним реферування електронних документів. Існують і інші проблеми обробки природної мови з використанням обчислювальної техніки, а саме розуміння природної мови, лінгвістичне забезпечення інформаційних систем різних типів, машинний переклад.

В умовах необхідності здійснення пошуку в постійно зростаючому масиві текстової інформації наукового характеру особливо гостро постає питання попереднього опрацювання текстових масивів. Зокрема, актуальною є проблема автоматичного аналізу повнотекстових документів, автоматичної класифікації, виявлення тематики, автоматичного реферування документів та установлення їх семантики.

Слід зазначити, що сьогодні жодна із систем аналізу текстових даних не забезпечує всіх поставлених до неї вимог стосовно повноти, точності та рівня автоматизованості. Дана ситуація пов’язана з тим, що наразі немає достатньо адекватних моделей таких систем. Це у сфері опрацювання текстових даних зумовлює постійне виникнення нових більш досконалих моделей подання текстів для використання їх у подальшому в інформаційно-пошукових системах (ІПС).

Основними проблемами, які ставляться перед ІПС при опрацюванні текстових документів, є: зменшення обсягу тексту при збереженні його семантики; автоматичне реферування; категоризація; побудова семантичних зв’язків; створення статистичного портрета. Нині серед традиційних методів аналізу можна виділити статистичні та семантичні. З огляду на доволі велику складність алгоритмів проведення семантичного аналізу, поки що ці методи не здобули значного розповсюдження і використовуються переважно для дослідницьких цілей. На практиці ж статистичні методи достатньо популярні в багатьох ІПС завдяки можливості опрацьовувати великі масиви даних і вирішувати різноманітні завдання щодо опрацювання отриманих за результатами аналізу да-них. Завданням статистичної обробки тексту є математичний опис його мовних фактів та явищ, зв’язків між ними, отримання набору моделей для вирішення визначених лінгвістичних задач.

Метою курсового проєкту є опрацювання основних методів первинної та статистичної обробки текстових даних і інтерпретації результатів аналізу. Виконання проєкту передбачає експерименти із науково-технічним текстом.

**1 АНАЛІЗ ПРОБЛЕМНОЇ ОБЛАСТІ ТА ФОРМУЛЮВАННЯ ЗАДАЧ ДОСЛІДЖЕННЯ**

Обробка природної мови (ПМ) – це формулювання і дослідження комп’ютерно-ефективних механізмів для забезпечення комунікації з ЕОМ на ПМ. Об’єктами досліджень є:

– власне природна мова;

– використання ПМ як у комунікації між людьми, так і в комунікації людини з ЕОМ;

– машинне розуміння природної мови.

Завдання досліджень – створення комп’ютерно-ефективних моделей комунікації на ПМ. Саме така постановка задачі відрізняє NLP від задач традиційної лінгвістики та інших дисциплін, які вивчають ПМ, і дозволяє віднести її до області ІІ. Проблемою NLP займаються дві дисципліни: лінгвістика і когнітивна психологія [1].

Традиційно лінгвісти займалися створенням формальних, загальних, структурних моделей ПМ, і тому віддавали перевагу тим з них, які дозволяли отримувати якомога більше мовних закономірностей і робити узагальнення. Практично ніякої уваги не приділялося питанню про придатність моделей з точки зору комп’ютерної ефективності їх застосування. Таким чином, виявилося, що лінгвістичні моделі, характеризуючи власне мову, не розглядали механізми його породження і розпізнавання. Хорошим прикладом тому служить породжує граматика Хомського, яка виявилася абсолютно непридатною на практиці в якості основи для комп’ютерного розпізнавання ПМ.

Завданням же когнітивної психології є моделювання не структури мови, а його використання. Фахівці в цій галузі також ні додавали великого значення питанню про комп’ютерну ефективності [1].

Розрізняються загальна і прикладна NLP. Завданням загальної NLP є розробка моделей використання мови людиною, що є при цьому комп’ютерно-ефективними. Основою для цього є загальне розуміння текстів, як це мається на увазі в роботах Чарняка, Шенка, Карбонелл і ін. Безсумнівно, загальна NLP вимагає величезних знань про реальний світ, і велика частина робіт зосереджена на поданні таких знань і їх застосування при розпізнаванні надходить повідомлення на ПМ. На сьогоднішній день штучний інтелект ще не досяг того рівня розвитку, коли для вирішення подібних завдань у великому обсязі використовувалися б знання про реальний світ, і існуючі системи можна називати лише експериментальними, оскільки вони працюють з обмеженою кількістю ретельно відібраних шаблонів на ЕЯ.

Прикладна NLP займається звичайно не моделюванням, а безпосередньої комунікації людини з ЕОМ на ПМ. В цьому випадку не так важливо, як введена фраза буде зрозуміла з точки зору знань про реальному світі, а важливо вилучення інформації про те, чим і як ЕОМ може бути корисною користувачеві (прикладом може служити інтерфейс експертних систем). Крім розуміння ПМ, в таких системах важливо також і розпізнавання помилок і їх корекція.

**1.1 Основна проблема обробки природної мови**

Основною проблемою NLP є мовна неоднозначність. Існують різні види неоднозначності:

– Синтаксична (структурна) неоднозначність: у фразі Time flies like an arrow для ЕОМ неясно, чи йде мова про час, який летить, або про комах, тобто чи є слово flies дієсловом або іменником.

– Смислова неоднозначність: у фразі The man went to the bank to get some money and jumped in слово bank може означати як банк, так і берег.

– Відмінкова неоднозначність: привід in в пропозиціях He ran the mile in four minutes / He ran the mile in the Olympics позначає або час, або місце, тобто представлені абсолютно різні відносини.

– Референціальние неоднозначність: для системи, що не володіє знанями про реальний світ, буде важко визначити, з яким словом – table або cake – спів-відноситься займенник it у фразі I took the cake from the table and ate it.

– Літерації (Literalness): в діалозі Can you open the door? – I feel cold ні прохання, ні відповідь виражені нестандартним способом. В інших йдутьтельство на питання може бути отриманий пряму відповідь yes / no, але в даному випадку в питанні імпліцитно висловлено прохання відкрити двері.

Центральна проблема як для загальної, так і для прикладної NLP – дозвіл такого роду неоднозначностей – вирішується за допомогою перекладу зовнішнього подання на ПМ в якусь внутрішню структуру. Для загальної NLP таке перетворення вимагає набору знань про реальний світ. Так, для аналізу фрази Jack took the bread from the supermarket shelf, paid for it, and left і для коректної відповіді на такі питання, як What did Jack pay for ?, What did Jack leave? і Did Jack have the bread with him when he left? необхідні знання про супермаркетах, процесах купівлі та продажу і деякі інші [1].

Прикладні системи NLP мають перевагу перед загальними, тому що працюють у вузьких предметних областях. Наприклад, системі, використовуваної продавцями в магазинах з продажу комп’ютерів, не потрібно «роздумувати» над неоднозначністю слова terminals в питанні How many terminals are there in the order. Проте, створення систем, що мають можливість спілкування на ПМ в широких областях, можливо, хоча поки результати далекі від задовільних.

Технології аналізу природної мови. Під технологією аналізу ПМ мається на увазі переклад деякого виразу на ПМ у внутрішнє представлення. Фактично всі системи аналізу ПМ можуть бути розподілені на такі категорії: підбір шаблону (Pattern Matching), синтаксичний аналіз, семантичні граматики, аналіз за допомогою відмінкових фреймів, «чекай і дивися» (Wait And See), словниковий експертний (Word Expert), коннекціоністскій, «ковзний» (Skimming) аналіз. Нижче піде мова про деякі найбільш поширених методах, описаних в статті [1].

Підбір шаблону. Сутність даного підходу полягає в інтерпретації введення в цілому, а не в інтерпретації сенсу і структури його окремих становить на більш низькому рівні. При використанні цього методу виходить порівняння вже наявних в системі шаблонів – зразків з текстом, що надійшли на вхід. За-звичай шаблони представлені у вигляді простого списку відповідностей між класами висловлювань і інтерпретаціями. Іноді вони доповнені семантичними елементами або іншими компонентами вищого рівня. За таким принципом працює система Еліза, імітує діалог з психотерапевтом [2]. Насправді система нічого не розуміє, а лише підтримує діалог, порівнюючи репліки пацієнта з шаблонами і присвоєними їм відповідними репліками.

**1.2 Статистична обробка текстів**

Статистичні дані можуть збиратися для окремих слів, фраз, абзаців та текста в цілому. Фрази надають більше семантичної інформації, ніж окремі слова, тому що вони дають загальне уявлення і про контекст. Слово характеризує своїм оточенням і через багатозначності більшості слів, дійсно необхідно знати хоча б одну фразу, яка містить певне слово, щоб визначити його семантичне значення з більшою впевненістю [2].

Часто аналіз деякої множини документів призводить до необхідності опису їх внутрішньої структури, створення категорій для подальшої класифікації документів. Завдання організації неструктурованої множини документів називають текстової класифікацією. Її рішення дозволяє полегшити зберігання, пошук і перегляд документів. Складнощі побудови текстової анотації і труднощі в отриманні навчальних даних, привели дослідників до альтернативних контрольованих методів, які використовують невеликі набори маркованих даних. Алгоритми класифікації текстів можуть бути застосовані для багатьох завдань: тематична розмітка текстового корпусу, класифікація всіляких оглядів.

Безперервне накопичення текстових даних привело до необхідності розробки методів Text Mining для забезпечення ефективної роботи з великими корпусами текстів. Text Mining – це область досліджень, що розвивається. Перед застосуванням того чи іншого методу Text Mining потрібно чітко сформулювати мету дослідження, тому що кожна задача має свою специфіку. Важливо взяти до уваги мову документа, так як від цього залежить необхідний морфологічний і синтаксичний аналіз. Чітке визначення предметної області текстового корпусу допомагає визначити набір спеціальних термінів, які можуть бути присутніми в тексті, і які слова можна розглядати як надлишкові [3].

**1.3 Класифікація та кластеризація текстів**

При автоматичній обробці текстів серед інших завдань виникають дві близькі, а саме класифікації та кластеризації текстів.

Перша полягає в тому, щоб віднести текст до певної категорії.

Друга – розбити набір текстів на групи, близькі в певному сенсі.

Загальна постановка задачі класифікації може бути записана наступним чином. Задано множину документів D = {dl, d2, ..., dD] і множину категорій С = {с1, с2 ... СN}. Існує деяка невідома функція F, яка визначає приналежність документа до певної категорії F: DxC{0,1}. Необхідно знайти функцію F’ максимально близьку до F, яка і буде класифікатором.

Побудова класифікатора F’ проводиться на підмножині документів D1, яку називають навчальною вибіркою, а перевіряється з використанням підмножини документів D2, яке називають тестової вибіркою. Ефективність побудованого класифікатора можна оцінити за кількістю отриманих на тестовій вибір-ці збігів.

При кластеризації текстів відбувається спроба з’ясувати природне виділення в групах текстів певних груп з використанням відповідних алгоритмів і візуалізації. Для кожного тексту в будь-якому випадку будується набір ознак.

Завдання автоматичної класифікації і кластеризації текстів має велике практичне значення. Відповідні процедури застосовуються при обробці інформаційних потоків, таких як електронна пошта та новини, рекламні оголошення, створення каталогів в Інтернеті, при автоматичному реферування і анотування.

Тісно пов’язане з цим завдання атрибуції текстів. Атрибуція (від лат. Attributio – приписування) – визначення атрибутів. Існують методи, що дозволяють проводити атрибуцію тексту. А саме – віднесення його до певного жанру, стилю, часу написання. Але ймовірно найбільш важливою є задача атрибуції авторства твору. Для неї можуть бути використані будь-які методи класифікації і кластеризації текстів, але існують і значні відмінності при проведенні цих процедур. В першу чергу це пов’язано з визначенням авторства літературних творів, коли під сумнів ставиться приналежність автору відомого твору, або відбувається спроба відновити історичну справедливість, повернувши світу ім’я невідомого письменника.

Автоматичне встановлення авторства письмових текстів, крім літературознавства, застосовуються в сфері безпеки і при захисті авторських прав, кримінальному та цивільному діловодстві, криміналістиці. Завдання досить актуальне, тому що все більше текстів доводиться аналізувати в електронному вигляді, позбавленому, тих характерних атрибутів, які відрізняють тексти, написані від руки.

Наприклад:

– послано електронного листа від імені людини, який відмовляється від послання, сказавши, що лист було написано з його робочого місця під час його відсутності;

– повідомлення в блогах пишуться одним автором під декількома вигаданими іменами;

– при підготовці електронних видань було проведено зміну змісту.

З такими ситуаціями доводиться стикатися все частіше, так як доступ до робочого місця за комп’ютером часто не захищений. До сих пір при розслідуванні таких випадків використовували такі методи, як зняття показань, аналіз почерку, відбитків пальців, ритм друкування тексту на клавіатурі та інші. Зараз шукають можливість отримати інформацію про авторів тексту з самого тексту.

Особливу важливість має робота по атрибуції (встановлення авторства) анонімного або псевдонімного літературного твору, так як це пов’язано з етичними проблемами.

У загальному випадку в атрибуції літературного твору може бути виділено два етапи:

– етап формування гіпотези;

– етап перевірки гіпотези і інтерпретації результатів.

Етап формування гіпотези виконується за допомогою традиційних філологічних методів атрибуції з урахуванням як суб’єктивних, так і об’єктивним даних. На цьому етапі може бути сформульована гіпотеза про приналежності тексту Mi письменникові А. Може висуватися гіпотеза про час написання тексту Mi на основі його мовних і стилістичних зв’язків з текстами Мг, М3, ... Мп письменника А, датування яких відома.

Перевірка гіпотези виконується з урахуванням як суб’єктивних, так і об’єктивних даних, а саме наявності свідоцтв в документах архівів, листах, аналізі стилю автора та інших.

Етап перевірки гіпотези і інтерпретації результатів може виконуватися і методами прикладної лінгвістики з використанням процедур с статистико імовірнісного аналізу, теорії класифікації, теорії розпізнавання образів та інших математичних методів. Але при цьому треба розуміти, що статистико імовірнісного, що застосовуються при атрибуції, використовувати тільки для перевірки гіпотези і ніякої самостійної ролі не грають.

Таким чином, позитивне рішення з питання атрибуції тексту може бути прийнято тільки у разі збігу результатів філологічного і математико-статистичного аналізу. Цей підхід використовується в даній роботі при розгляді результатів тих чи інших методів при проведенні атрибуції текстів.

Можливість застосування математичних методів при атрибуції авторства забезпечує прийняте в стилістиці поняття стилю, яке досліджується зокрема академіком В. В. Виноградовим, на його думку «стиль письменника повинен вивчатися як єдина, внутрішньо цілісна система функціональних узгоджених засобів словесного вираження», в тому числі і з використанням математичних методів [4].

У вісімдесяті роки XIX століття було зроблено безліч спроб опису особливостей літературного стилю за допомогою таблиць і графіків, найпомітнішим став твір американського фізика Томаса Мандельхолл «Характеристичні криві композиції», в якому він розглядає розподіл довжини слів і частоти букв в текстах Бекона , Шекспіра, Марлоу [4].

В опублікованій в 1944 році роботу шотландського статистика Удні Єля «Статистичне дослідження літературного словникового запасу» розглядалося число народження іменників як ознака, можна судити про авторство творів [4].

В якості основних робіт, які були присвячені стилеметрії (stylometry), в Росії в докомп’ютерний період для вирішення задач атрибуції, можна привести дослідження початку XX століття Миколи Олександровича Морозова (1854 – 1946). У статті, що вийшла в 1915 році і названої «Лінгвістичні спектри», Н. А. Морозов характеризує свій метод як «засіб для відрізнення плагіату від справжніх творів того чи іншого відомого автора» [4]. Ідею методу Морозов запозичує у німецьких дослідників XIX століття В. Діттербергера і К. Ріттера, які вивчали спірні тексти (серед них тексти, приписувані Платону, Гете) методами статистичного аналізу вживаності окремих мовних форм, слів, виразів, фразеологічних зворотів і синонімів. Причому в якості рахункових одиниць вибиралися найбільш рухливі і легко замінюються синонімами елементи мови. Цікава реакція на цю статтю найбільшого вітчизняного вченого А. А. Маркова, який, посилаючись на свій досвід, вивчення тексту «Євгенія Онєгіна», а також на результати повторених ним досліджень тих же текстів, що служили матеріалом аналізу для Морозова, показав, що висновки останнього не достовірні, так як були отримані на текстах недостатнього обсягу [4].

Як відомо, для визнання належності твору якогось конкретного автору необхідно, щоб аргументи характеризували його з трьох сторін: біографічної, ідеологічної та стилістичної, при цьому опис і аналіз індивідуально авторських стилів є необхідним етапом будь-якої атрибуції [4]. Головне завдання при аналізі художніх творів в процесі атрибуції полягає у виявленні специфічних ознак мови автора і зіставленні, їх зі специфічними ознаками всіх передбачуваних авторів. До 70-х років в практиці атрибуції панували історико-документальні та філологічні методи дослідження, багато в чому базуються на суб’єктивному сприйнятті аналізі літературного твору дослідником. В.В. Виноградов виділив 11 принципів атрибуції анонімних і псевдонімних творів, розділивши їх на групи об’єктивних і суб’єктивних. У число об’єктивних принципів їм були включені лінгвостатичні [4]. Математичний апарат і, зокрема, лінгвостатичні принципи стали широко застосовуватися при атрибуції літературних творів з 80 – х родів. Історія розвитку методів атрибуції привела до розуміння того, що ефективний метод стилістичного аналізу з метою визначення авторства повинен охоплювати різні рівні мовної системи і включати в себе застосування багато-вимірних класифікацій. Опис індивідуальних авторських стилів з метою пере-ведення інформації, що міститься в неформализованном вигляді в тексті, на формалізований штучний мова на основі математичної статистики полягає у виявленні специфічних мовних ознак тексту і використанні строгих правил параметризації [4].

Можна виділити наступні характеристики, які можуть бути отримані для письмового тексту і доступні для подальшого аналізу:

– легко обчислюються за текстом: довжина пропозицій і слів, середня довжина речень і слів;

– однорідність тексту (розподіл по тексту складових одиниць тексту);

– граматичні конструкції мови;

– морфологічні конструкції мови;

– синтаксичні конструкції мови;

– лексика (багатство лексики, частотні словники, наявність визначено-них слів);

– переходи між складовими одиницями тексту;

– аналіз додаткових ознак тексту (скорочень, пунктуації, «смайликів», помилок).

Отримані лінгвистичні характеристики піддаються аналізу з використанням різних математичних методів, серед яких можна виділити наступні:

– статистичні методи;

– вивчення переходів між складовими одиницями тексту;

– методи розпізнавання образів і штучного інтелекту.

Незважаючи на велику кількість робіт присвячених методам автоматичного визначення авторства, до них можна знайти досить насторожене ставлення, як з боку математиків, так і філологів, що не залишається без уваги фахівців в області атрибуції, які розглядають критику як вимога до систематизації вже накопичених матеріалів і створення бази для проведення подальших досліджень [4].

Вивчення структури літературного тексту з використанням різних математичних методів має багату історію, а поява обчислювальних машин розширило можливості проведення різних експериментів.

Проведення подібних досліджень передбачає наявність літературного матеріалу, відібраного і підготовленого фахівцями.

**1.4 Автоматичне реферування та індексація**

Необхідність дослідження та розробки систем автоматичного реферування зумовлено збільшенням кількості та обсягу електронних документів, які потребують обробки, оскільки більшість таких документів має вигляді неструктурованих текстів, складених природною мовою, а більшість програмного забезпечення орієнтовано на роботу зі структурованими даними. Крім того, весь час зростає кількість новинних інтернет-сайтів, і для однієї і тієї ж події різні сайти надають різні інтерпретації. Разом з поширенням мобільного Інтернету та пристроїв класу смартфон, це створює попит на системи, які, зібравши дані з різних джерел, можуть дати користувачу короткий, проте достатній за охопленням звіт – реферат про поточні новини. Ще один спосіб використання автоматичного реферування пов’язаний з системами підтримки прийняття рішень.

Експертам для виконання швидкого огляду необхідно аналізувати велику кількість документів, і вдалі системи реферування скорочують час, необхідний для читання. Замінити систему реферування пошуковою системою не вдається, тому що пошукова система буде шукати те, про що експерт вже знає або здогадується.

Автоматичне індексування тексту – вибір машиною всіх слів, які є істотними для розуміння тексту. Є два способи побудови тематичного представлення – з фіксованими темами та з динамічними темами. Індексація з фіксованими темами спирається на фіксовані тематичні словники. Найпростіше скласти теми з чітко вираженою термінологією: біологія, хімія, комп’ютерна тематика, фінанси, геологія та географія, право, лінгвістика, математика, атомна енергетика, фізика тощо. У таких темах до цих ключових слів входять як деталізовані, так і більш загальні поняття.

Для задачі мультиреферування, яке в першу чергу орієнтується на тексти новин, використання фіксованих списків ключових слів не є раціональним. Окремо використовується список слів, які при тематичному аналізі тексту будуть ігноруватися.

**1.5 Змістовна близькість**

Змістовна близькість є одним з найбільш важливих елементів зв’язності тексту. Розглянемо такий фрагмент тексту: «Розвиток обчислювальної техніки відбувався одночасно з розвитком технологій. На зміну електронним лампам прийшли напівпровідники. На їх основі було створено інтегральні схеми. Щодалі, то більше і більше логічних елементів стали розміщувати на одному кристалі».

У даному тексті виділено два типи змістовної близькості:

– пов’язаної за значеннями термінів;

– пов’язаної за структурою тексту.

Деякі словосполучення позначають поняття, пов’язані значеннями відповідних слів або термінів. Зв’язок першого типу будується на основі знань про те, що два елементи (поняття) пов’язані між собою в базі знань. Інші слова позначають поняття, пов’язані з іншими поняттями за рахунок того, що зустрічаються поблизу. Цей тип зв’язку задається автором тексту і є відображенням того, що саме хотів повідомити автор цим текстом.

Зв’язки можуть бути реалізовані як термінами безпосередньо, так і мовними вказівниками (у лінгвістиці це називається анафора). Задача встановлення елементів, на які вказують мовні вказівники, називається розв’язанням анафори [5].

**1.6 Виділення термінологічних одиниць**

Задача виділення термінологічних одиниць виникає, коли в тексті зустрічаються терміни, які не відомі базі знань. Особливо це важливо, коли терміни складаються з декількох слів. Звичайно використовують такі методи: підрахунок кількості пар, t–критерій Стьюдента, критерій узгодженості Пірсона. Методи простого частотного підрахунку та t–критерій також є досить ефективними і можуть бути використані для складання списку термінів – кандидатів у системах напівавтоматичного формування термінів. Основний тип помилок обох методів – виділення стійких загальновживаних словосполучень, які задовольняють шаблонам – обмеженням.

Найбільша проблема всіх частотних методів наступна: при збільшенні довжини терміну падає його частота, навіть у спеціалізованому корпусі термін може зустрічатися один – два рази.

Метод максимальної довжини. Виділення максимальних ланцюжків, які містять терміни. Ці ланцюжки визначаються через негативний відбір: складається список слів і знаків, які не можуть входити в термін. У нашій реалізації в якості таких роздільників ми розглядаємо розділові знаки, стоп – слова, дієслова, дієприслівники; послідовності слів між цими роздільниками розглядаються як кандидати в терміни.

Метод контролю у вікні. До частоти сумісного входження включається також частота входжень слів в одне вікно певного розміру. Вважається, що якщо пари елементів зустрічаються як безпосередні сусіди більш ніж у полови-ні випадків їх появи в тому самому текстовому вікні, те ця пара являє собою термін або фрагмент терміну. Відбувається склейка пари у єдиний елемент, таблиці перераховуються так, ніби цей елемент був відомий із самого початку, до початку обробки тексту, що дає можливість і далі нарощувати термін.

**Висновки до розділу**

Аналіз літературних джерел показує, що статистична обробка текстів є важливою складовою загальної методології комп’ютерної обробки текстових даних. У подальших розділах проекту реалізовано основні етапи статистичної обробки тексту, запозиченого з технічних статей.

**2 ПЕРВИННА ОБРОБКА ТЕКСТУ**

**2.1 Відомі статистичні закономірності у текстах та ключові слова**

В усіх текстових документах, створених людиною, можна спостерігати статистичні закономірності. У будь-якій мові є слова, які зустрічаються частіше, ніж інші, але не мають самостійного значення. Є слова, які зустрічаються рідше, але мають набагато більше смислове значення.

У 1949 році гарвардський професор-лінгвіст і філолог Джордж Ципф, працюючи над принципом найменшого зусилля, сформулював кілька закономірностей, отриманих не на основі математичних висновків, а на основі аналізу статистики частоти слів в текстах на багатьох мовах, тобто емпірично [14].

У той час, коли Ципф сформулював помічені їм закономірності розподілу частоти слів, законом вони не вважалися – ще не було комп'ютерів і не можна було провести точні розрахунки, що підтверджують виявлені закономірності. У подальшому були проведені масштабні чисельні дослідження, які підтвердили і уточнили помічені закономірності. При цьому провідну роль в обґрунтуванні законів зіграли роботи Б. Мандельброта [14].

Зокрема Ципф вважав, що слова з великою кількістю букв зустрічаються в тексті рідше коротких слів. Ґрунтуючись на цьому постулаті, Ципф вивів два універсальних закони [14].

Виміряємо кількість входжень кожного слова в текст і візьмемо тільки одне значення з кожної групи, що має однакову частоту. Розташуємо частоти у міру їх спадання і пронумеруємо, а порядковий номер частоти назвемо рангом частоти (позначимо ранг слова ). Слова, що зустрічаються найбільш часто матимуть ранг 1, наступні за ними – 2 і так далі.

Тоді очевидно, що ймовірність зустріти довільне, заздалегідь вибране слово буде дорівнює відношенню кількості входжень цього слова до загально-го числа слів у тексті ( n – кількість входжень слова, N – кількість слів у тексті) [14].

р=ni/N (2.1)

Ципф виявив таку закономірність: добуток ймовірності виявлення слова в тексті на ранг частоти являє собою постійне число (С) [14]. Закон показує, що поширеність слова в тексті змінюється по гіперболі, залежно від кількості входжень. Наприклад, друге зі слів, що зустрічається найчастіше, зустрічається приблизно в два рази рідше, ніж перше, третє – в три рази рідше, ніж перше і так далі [14]. Значення константи в різних мовах різне, але всередині однієї мовної групи залишається приблизно незмінною, незалежно від тексту. Для російських текстів константа Ципфа приблизно дорівнює 0,08, для англійських текстів – 0,1.

Перший закон не враховує факт того, що різні слова можуть входити в текст з однаковою частотою. Ципф встановив, що частота і кількість слів, що входять в текст з цією частотою, також мають залежність. Якщо побудувати графік, відклавши по осі абсцис частоту входження слова, а по осі ординат – кількість слів у даній частоті, то отримана крива буде зберігати свій вигляд для всіх без винятку текстів [14].

Як і для першого закону, це твердження вірне в межах однієї мови. Однак і міжмовні відмінності невеликі. Якою б мовою текст не був написаний, вигляд кривої Ципфа залишиться незмінною. Може трохи відрізнятися лише коефіцієнт гіперболи [14].

Також важливим є той факт, що і документи всередині якої-небудь галузі знань можуть розподілятися відповідно до цього закону. Дослідження показують, що найбільш значущі для тексту слова лежать у середній частині графіка кривої Ципфа. Цей факт має просте обґрунтування – слова, які трапляються дуже часто, в основному виявляються прийменниками чи займенниками. З іншого боку слова, що зустрічаються рідко, в більшості випадків, не мають вирішального смислового значення. Від установки ширини діапазону (вікна пошуку ключовиків) залежить статистична якість пошуку значущих слів. Якщо встановити велику ширину діапазону, то в ключові слова будуть потрапляти допоміжні слова; якщо встановити вузький діапазон – можна втратити значущі терміни. Тому, в кожному окремому випадку, необхідно використовувати ряд євристик для визначення ширини діапазону, а також методик, що зменшують вплив цієї ширини.

Одним із способів, наприклад, є попереднє виключення з досліджуваного тексту слів, які за визначенням не можуть бути значущими тому, що складають вербальний «шум». Такі слова називаються нейтральними або стоповими (стоп-словами). Наприклад, для російського тексту стоповими словами могли б бути всі прийменники, частки, особисті займенники.

Відомі інші способи підвищити точність оцінки значущості слів за рахунок додаткової інформації. Певні слова можуть зустрічатися майже у всіх документах деякої колекції і, відповідно, давати малий вплив на приналежність документа до тієї чи іншої категорії, тобто не бути ключовими для цього документа. Тому очевидно, що, розглядаючи всю колекцію документів, ми підвищимо інформативність пошуку ключових слів [15].

При використанні цього підходу не виключена ймовірність попадання в ключові слова випадкових спеціальних термінів, рідкісних слів і власних імен та іншого вербального «шуму». Тому необхідно в попередню обробку тексту включати алгоритм, що підвищує якість відбору. Евристики такого відбору частіше залежать від конкретно взятого випадку.

Основний зміст документа (тексту) може бути виражений за допомогою певних слів, узятих безпосередньо з цього тексту [2]. Зазвичай до кожного розгорнутого тексту можна скласти цілий набір ключових слів різного обсягу (найчастіше від 5 до 15 слів). Але взагалі кількість ключових слів може варіюватися в широких межах. Відповідно до думки Л.В. Цукрового і А.С. Штерн, найоптимальнішим є невеликий набір ключових слів і словосполучень – 7-10 слів. В окремих випадках компресія тексту може призвести до знаходження од-ного основного ключового слова, яке має найбільшу частоту [3], [4].

Ключовим словом будемо вважати таке слово в тексті, яке здатне в сукупності з іншими ключовими словами відображати зміст цього тексту [5].

Набір ключових слів близький до анотації, плану і конспекту, які теж представляють документ з меншою деталізацією, але, на відміну від ключових слів, пов’язані у синтаксичні структури [1].

Термін «ключові слова» значною мірою умовний – як ключові ознаки в тексті можуть виступати не тільки слова, а й словосполучення і, навіть, речення. Ключове (опорне) слово – це термін, що відноситься до основного змісту тексту [6] і повторюється в ньому кілька разів (з урахуванням всіх можливих синонімів). Ключове словосполучення – це поєднання слів, серед яких є одне або кілька ключових [7], [8]. Ключовим також може вважатися речення, що містить два і більше ключових слова або ключових словосполучення [9]. Ключові слова мають ряд суттєвих ознак:

– високий ступінь повторюваності даних слів у тексті, частотність їх вживання;

– здатність знака концентрувати, згортати інформацію, закладену в увесь текст, об'єднувати «його основний зміст»; ця ознака особливо яскраво проявляється у ключових словах у позиції заголовку.

Однак вибір ключових слів є дуже непростою операцією і вимагає зваженого підходу. Слід вибирати ті ключові слова, які найбільш точно відображають специфіку розглянутої теми. При цьому необхідно уникати випадкових і загальних фраз, не рекомендується повторювати кілька разів одні й ті ж ключові слова. Отже, процес пошуку ключових слів є аналітичним [9].

У процесі попереднього оброблення тексту проводиться видалення неінформативних частин [10]. Найперше до таких відносять стоп-слова – це слова, що не представляють цінності як потенційно ключові. Найчастіше це прийменники, сполучники, вигуки тощо [10].

До уваги беруться також N-грами – це термін з комп'ютерної лінгвістики, що означає послідовність з N елементів тексту (термів), наприклад, слів або їх послідовностей. Кандидати в ключові слова можна відбирати у вигляді N-грам, що не розділені знаками пунктуації (крім дефіса і лапок) і стоп-словами [11], [12]. Ключовими словами можуть бути як поодинокі слова, так і пари слів, трійки тощо.

Інший підхід ґрунтується на аналізі зв’язків між словами в реченнях і в тексті, отриманих або за допомогою тих же N-грам (підхід менш трудомісткий), або на розібраному тексті. Якщо текст пройшов більш затратну обробку на етапах аналізу (морфологічний дасть базові словоформи, синтаксичний – зв'язки між словами, семантичний – смислову карту зв’язків), можна отримати більш точну інформацію про текст, наприклад, на підставі дерев синтаксичного розбору (за наявності синтаксичної розмітки) [13]. Найважливішим етапом в задачі знаходження ключових фраз є розрахунок їх ваг інформативності, який дозволяє оцінити їх значимість по відношенню один до одного в документі. Для кожної з відібраних ключових фраз розраховуються ознаки, які дозволяють судити про важливість кандидата для даного документа. Набір відібраних ключових фраз ранжується за значеннями ознак, наприклад, відповідно до їх частотності та ваги інформативності, розрахованими за однією з методик. Після ранжування проводиться відбір кращих ключових фраз з цього списку або відбираються кандидати, що перевищують встановлений мінімальний поріг значення ознаки [1]. Отже, вибір найбільш інформативної ознаки при пошуку ключових слів та / або фраз має безпосередній вплив на якість результатів їх визначення.

**2.2 Видалення стоп-слів та стемінг**

Зазвичай, текстові документи, що проходять скрізь механізми обробки інформації не є спеціально підготовленими. Це може впливати на час, що затрачується на опрацювання інформації, а також на фінальний результат.

Здебільшого, корпус досліджуваних документів включає в себе велику кількість слів. Деякі слова можуть не містити смислового навантаження, інші – можуть бути різноманітними варіантами одного слова тощо. Видалення надлишкових слів, а також приведення схожих між собою слів до однієї форми скорочують затрати часу на аналіз інформації. Усунення описаних проблем виконується на етапі попередньої обробки тексту [3].

Класичними рішеннями для видалення неінформативних слів та уніфікації слів зі схожим значенням є:

– видалення стоп-слів. Стоп-словами називаються слова, що є допоміжними в тексті та несуть мало інформації про зміст текстового документа. Зазвичай, списки таких слів складаються перед обробкою. В процесі обробки слова, що співпадають зі словами з цього списку, видаляються. Типовим прикладом таких слів є допоміжні слова та артиклі, наприклад: «так як», «окрім того» и тому подібні. Іноді, стоп-слова можуть бути специфічними для тієї чи іншої галузі та бути підібрані для конкретних випадків;

– стемінг – це морфологічний пошук основи слова. Сутність стемінгу полягає в перетворенні кожного слова до його нормальної форми. Нормальна форма виключає відмінок слова, множинну форму, особливості усного мовлення і т. п. Наприклад, слова "стиснення" і "стислий" повинні бути перетворені в корінну форму слова "стиск".

Згідно з варіантом завдання маємо текст, якій підлягає обробці (додаток А).

Проаналізуємо цей текст та складемо список стоп-слів. Нижче представлено список стоп-слів, які виявлено у тексті:

1. то, но, не, нет, что, или, еще, для, на, по, при, тому ли, бы, чтобы;

2. он, они, она, ему, них, их;

3. в, и, с, к, о, а;

4. мы, нас, нам, все;

5. был, были, было, быть, будут;

6. которая, которые, которых, который;

7. это, этого, этому, этом, этот, этой, эти, этим, того;

8. будет, теперь, можно;

9. данный, данного, данному, данном, данная, данной, данное;

10. можем, может, могут, могли, могла;

11. такой, таким, такие, такого, такое, таких, так же;

12. другим, другой, другое;

13. является, являются;

14. первые, первое, первому;

15. как, какой, какими, какую;

16. сам, самом, сам, самого, самое, самому, само, себя, себе;

17. вполне, однако, например, также, когда, если, иметь, даже, вообще;

18. дело, делом, деле;

19. допустим, допустить, допустил, допускает;

20. представить, представляет;

21. случае, случаев, случаи;

22. ответ, ответу, ответом;

23. следовательно, следует, следующим, следующий, следующее.

Далі представимо абзац тексту до видалення стоп-слів та після видалення.

Абзац тексту до видалення стоп-слів:

Однако свойство алгоритмической универсальности не ограничивается констатацией того, что для всех известных алгоритмов оказывается возможной их программная реализация на ЭВМ. Содержание этого свойства имеет и характер прогноза на будущее: всякий раз, когда в будущем какое-либо предписание будет признано алгоритмом, то независимо от того, в какой форме и какими средствами это предписание будет первоначально выражено, его можно будет задать также в виде машинной программы.

Абзац тексту після видалення стоп-слів:

Однако свойство алгоритмической универсальности ограничивается констатацией, известных алгоритмов оказывается возможной программная реализация ЭВМ. Содержание свойства характер прогноза будущее: будущем предписание признано алгоритмом, независимо форме средствами предписание первоначально выражено, задать машинной программы.

**2.3 Формування робочої підмножини слів для статистичного аналізу**

Порядок виконання завдання формування робочої підмножини слів для статистичного аналізу наступний:

1. Використовуючи ресурси <https://pr-cy.ru/zypfa/text> або <https://miratext.ru/seo_analiz_text>, та аналізуємо текст без стоп-слів за Ципфом, тобто за рейтингом частотності. Ресурс видає таблицю, в якій в порядку зменшення частоти перераховані слова тексту.

2. Якщо стоп-слова видалені не повністю, всі вони й займуть верхні позиції у рейтингу частот. А важливі слова не потраплять у перші 25. Тому цей етап є ітераційним аж поки у рейтингу частих слів перши місця не займуть дійсно важливі слова.

На рис. 2.1 показано скріншот таблиці результатів аналізу тексту за Ципфом.



Рисунок 2.1 – Результати аналізу тексту за Ципфом

Далі слід провести стемінг тих слів, які мають єдиний корінь, але представлені у різних словоформах. Якщо стемінг зроблено, маємо ранжований за частотою входжень список слів, з якими надалі потрібно працювати:

1. Пошук (пошуку, пошук, пошуком, пошукових, пошуковому, пошуковой, пошуковою, пошукового).
2. Прецедент (прецедент, прецеденту, прецедентом, прецедента, прецедентів).
3. Інформац (інформаційного, інформації, інформаційно, інформацію).
4. Систем (систем, система, системами, системою, системах, системи).
5. Документ (документів, документ, документах, документа, документу, документами).
6. Метод (методики, методика, методикою, методику, методів, методах, метод, методу, методом).
7. Проблем (проблема, проблемам, проблем, проблему, проблемою).
8. Запит (запиту, запит).
9. Користувач (користувача, користувач, користувачам, користувачами).
10. Рішення.
11. Вилучен.
12. Машин (машин, машина, машинах).
13. Основі.
14. Дан (дані, даних).
15. Завдан (завданню, завдання, завданням).
16. Метрик (метрики, метрику).
17. Відстан.

**Висновки до розділу**

Отже, виконано первинну обробку тексту, а саме – видалення стоп-слів, аналіз тексту за Ципфом, виділення робочої підмножини важливих слів, стемінг деяких слів для приведення їх до єдиної словоформи. Отримано ранжований список слів для подальшої статистичної обробки тексту.

**3 СТАТИСТИЧНА ОБРОБКА ТЕКСТУ**

**3.1 Розрахунок відносних частот слів по абзацах**

Після отримання робочої підмножини слів переходимо до формування таблиць у програмі Excel. Порядок роботи наступний:

1. Відкриваємо нову книгу та аркуш Excel та робимо рядок зі слів з робочої підмножини. Таким чином робимо заголовки майбутніх стовпців першої таблиці. Залишаємо зліва листа приблизно 5 вільних стовпців.

2. Кожному абзацу обробленого тексту надається порядковий номер.

3. Для кожного абзацу визначаємо кількість слів за допомогою сервісу Статистика у редакторі Word.

4. На робочому аркуші лівіше за заготовки стовпців з частими словами формуємо два стовпці – перший – номери абзаців, другий – число слів у кожному абзаці.

5. Починаємо формувати тіло першої таблиці. Починаємо з 1-го слова.

Відкриваємо свій текст, задаємо у пошуку перше слово. Вікно пошуку показує його у якомусь абзаці. Фіксуємо номер абзацу і, використовуючи команду "Знайти далі", рахуємо, скільки разів слово зустрілося в цьому абзаці, потім у наступному абзаці і так до кінця тексту.

Результат – у першому стовпці записані абсолютні частоти входження 1-го слова у всі абзаци. Те саме робимо з усіма словами з нашого списку.

Отже, отримано таблицю абсолютних частот входжень слів у кожному абзаці тексту.

6. Формуємо другу таблицю, яка структурою повністю повторює першу. У її комірках обчислені відносні частоти входження кожного слова у кожний абзац.

Комірки форматуємо так, щоб числовий формат містив три знаки після коми. Відносна частота обчислюється так: абсолютна частота (клітина з першої таблиці) поділяється на число слів у цьому абзаці. Цю формулу записуємо для верхньої лівої клітини, переконуємося, що вона працює, і розмножуємо формулу на весь стовпець, а потім цей стовпець розмножуємо на всі інші стовпці.

Результат – заповнена таблиця відносних частот (рис. 3.1).

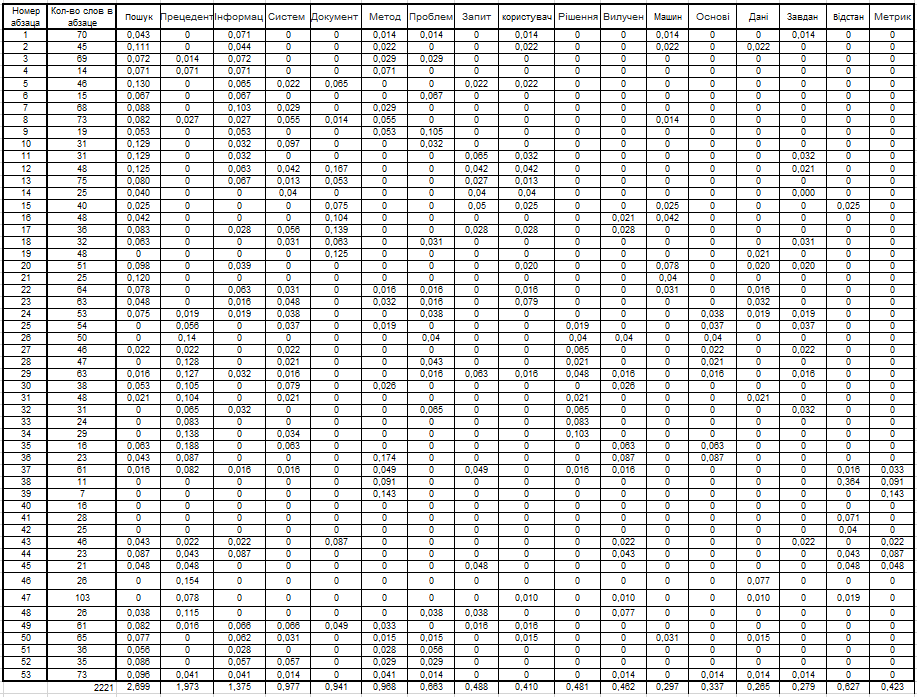


Рисунок 3.1 – Таблиця відносних частот

**3.2 Розрахунок коефіцієнтів кореляції відносних частот слів вздовж абзаців**

Порядок роботи:

1. Виділяємо робочу частину таблиці відносних частот і за допомогою майстра “Кореляція” розраховуємо коефіцієнти кореляції відносних частот слів робочої підмножини. Результат – матриця коефіцієнтів кореляції

2. Відзначаємо кольором ці комірки матриці, у яких значення коефіцієнтів кореляції перевищує 0,2. (рис. 3.2).

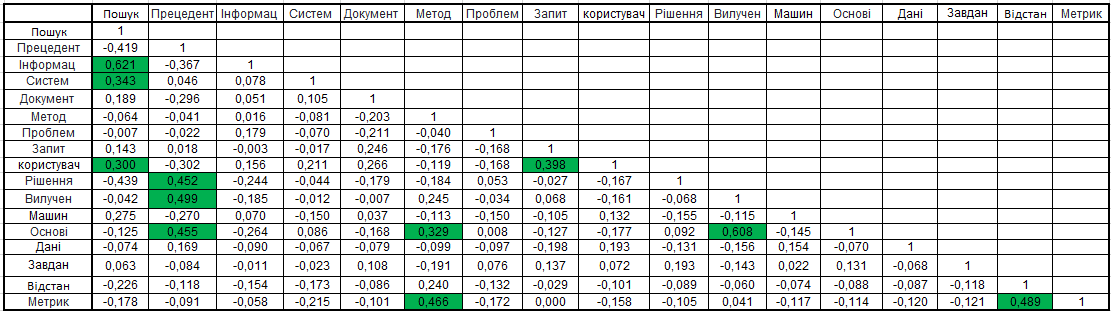


Рисунок 3.2 – Матриця коефіцієнтів кореляції

3. Складаємо ранжований ряд пар слів за убуванням значення кореляції (рис. 3.3).



Рисунок 3.3 – Таблиця ранжованих пар слів

**3.3 Визначення ключових слів**

Порядок роботи:

1. Аналізуємо таблицю ранжованих пар за коефіцієнтом кореляції. Відбираємо 8 слів, які найчастіше фігурують у цих парах, як підмножину КСА.

2. Аналізуємо таблицю ранжованих пар за косинусною мірою. Відбираємо 8 слів, які найчастіше фігурують у цих парах КСВ.

3. Визначаємо множину КС я перетин підмножин КСА та КСВ.

У результаті знайдено таки ключові слова: пошук, основи, прецедент, метод, метрик.

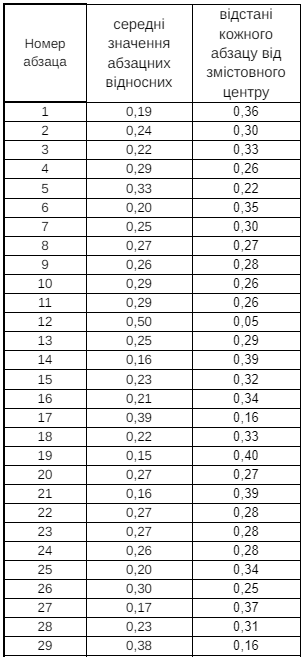
**3.4 Розрахунок відстаней абзаців відносно координат змістовного центру тексту**

Змістовним центром тексту назвемо сукупність середніх значень відносних частот ключових слів по абзацах. Алгоритм процедури містіть наступні кроки:

1. Обчислити середні значення абзацних відносних частот КС. Створити окрему стрічку з цими значеннями внизу таблиці відносних частот.

2. Застосовуючи Евклідову метрику, обчислити відстані кожного абзацу від змістовного центру. Створити окрему стрічку відстаней.

Результат розрахунку показано на рис. 3.4.



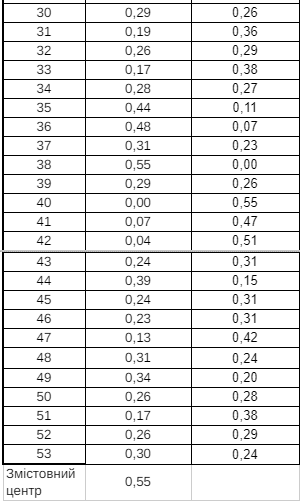


Рисунок 3.4 – Розрахунок відстаней абзаців до змістовного центру

**3.5 Розподіл абзаців за відхиленням від змістовного центру**

Позначивши відстані абзаців як *di*, розрахуємо дисперсію за формулою

, (3.1)

де N – кількість абзаців.

Потім розрахуємо середнє квадратичне відхилення:

 . (3.2)

Значення *S*, 2*S*, 3*S* розташуємо в окремих клітинках книзі.

Визначимо номери абзаців, що мають відстані у межах *S*, у межах 2*S* та 3*S*. Створимо таблицю розподілу абзаців за цими категоріями.

Здійснимо змістовний аналіз абзаців кожної категорії з метою порівняння значущості абзаців для змісту усього тексту. Перевіримо гіпотезу щодо зв’язку відстані абзацу від центру та його значущості у тексті.

// Тут зробити власні висновки щодо відповідності змісту кожного абзацу та наявності залежності цієї відповідності та знаходження абзацу у однієї з 3х груп.

У Додатку Б наведено один за одним тексти абзаців, які мають відхилення не більше *S*.

**Висновки до розділу**

Отже зроблено наступне:

1. Розраховано відносні частоти для підмножини важних слів по абзацах.

2. Розраховано коефіцієнти кореляції відносних частот слів вздовж абзаців.

3. Виділено ключові слова.

4. Користуючись переліком ключових слів, обчислено координати змістовного центру тексту.

5. Розраховано відстані кожного абзацу від змістовного центру.

6. Розраховано дисперсію та середнє квадратичне відхилення координат абзаців від центру тексту.

7. Абзаці згруповано на три категорії в порівнянні їх відстаней з величинами *S*, 2*S*, 3*S*.

8. Здійснено змістовний аналіз абзаці усіх трьох категорій для перевірки гіпотези залежності значущості абзацу від його відстані до центру.

Гіпотезу прийнято.

**ВИСНОВКИ**

У даній курсовій роботі опрацьовано основні методи первинної та статистичної обробки текстових даних і інтерпретації результатів аналізу, а саме:

1. Зроблено аналіз проблемної області та сформульовано задачу дослідження.

2. Виконано первинну обробку тексту, а саме – видалення стоп-слів, аналіз тексту за Ціпфом, виділення робочої підмножини важливих слів, стемінг деяких слів для приведення їх до єдиної словоформи. Отримано ранжований список слів для подальшої статистичної обробки тексту.

3. Розраховано відносні частоти для підмножини важних слів по абзацах.

4. Розраховано коефіцієнти кореляції відносних частот слів вздовж абзаців.

5. Виділено ключові слова.

6. Користуючись переліком ключових слів, обчислено координати змістовного центру тексту.

7. Розраховано відстані кожного абзацу від змістовного центру.

8. Розраховано дисперсію та середнє квадратичне відхилення координат абзаців від центру тексту.

9. Абзаці згруповано на три категорії в порівнянні їх відстаней з величинами *S*, 2*S*, 3*S*.

10. Здійснено змістовний аналіз абзаці усіх трьох категорій для перевірки гіпотези залежності значущості абзацу від його відстані до центру.

Гіпотезу відхилено/ прийнято.

11. На основі значущих абзаців складено скорочений варіант тексту.

**СПИСОК ЛІТЕРАТУРИ**

1. Партико З. В. Прикладна і комп’ютерна лінгвістика / З. В. Партико – Львів: Афіша, 2008. – 221 с.
2. Волошин В. Г. Комп’ютерна лінгвістика: Навчальний посібник / В. Г. Волошин – Суми: Університетська книга, 2004. – 382 с.
3. Анісімов А. В. Комп’ютерна лінгвістика для всіх: Міфи. Алгоритми. Мова / Анісімов А. В. – Киев: Наукова думка, 1988. – 223 с.
4. Стоп-слова в SEO-текстах [Електронний ресурс]. – Електрон. дані. – Київ, 2009. – Режим доступу: https://studiof1.uа/blog/seo/stop-slova-v-tekstax
5. Liddy E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. – 15 p.
6. Автоматична обробка текстів на природній мові та комп’ютерна лінгвістика : навч. посібн / Большакова Е.І., Клишинській Е.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. – К.: 2012. – 262 с.
7. Newsgroups [Електронний ресурс]. – Режим доступу до ресурсу: http://qwone.com/~jason/20Newsgroups/.https://bitbucket.org/sych\_d/text-preprocessing-investigation/src/master/.
8. Natural Language Processing with Python / Steven Bird, Ewan Klein, and Edward Loper. – O’Reilly Media, 2009. – 504 p.
9. Scikit-learn Machine Learning in Python [Електронний ресурс]. – Режим доступу до ресурсу: http://scikit-learn.org/stable/.
10. Natural Language Toolkit [Електронний ресурс]. – Режим доступу до ресурсу: https://www.nltk.org/.
11. Text preprocessing investigation code [Електронний ресурс]. – Режим доступу до ресурсу: <https://bitbucket.org/sych_d/text-preprocessing-investigation/src/master/>.
12. Text Data Management and Analysis / ChengXiang Zhai and Sean Massung. – ACM Books series, 2016. – 509 p.

**Додаток А**

Текст, що підлягає аналізу

1.АКТУАЛЬНІСТЬ РОБОТИ. У сучасних умовах розвитку інформаційних технологій та пошукових машин виникає потреба у нових методах забезпечення ефективного інформаційного пошуку. Проблема досконалого пошуку і збору інформації, яка може бути використаною при вирішенні важливих завдань в ході науково-дослідної діяльності, для підтримки прийняття рішень в науково-технічній, соціальній та інших сферах залишається відкритою впродовж десятиліть. Це зумовлено феноменом стрімкого перенасичення інформаційного простору [1]. До числа причин, які зумовили таке накопичення значних об'ємів інформації та зростання вимог до достовірності даних, що надаються користувачеві слід віднести популяризацію так званих віртуальних університетів та електронних форм навчання, збільшення важливості Internet-простору у сферах дозвілля та розвитку суспільства.

2.Відомо що сучасна пошукова машина здатна вдосконалюватися в різних напрямках: з'являються нові чинники ранжування або змінюється їх пріоритет, змінюється формат взаємодії інструментарію пошуку з користувачем, посилюються вимоги до якості побудови сайтів, а також з'являються нові сервіси, що спрощують пошук інформації. Відповідно вимоги до швидкості пошуку, актуальності інформації з кожним днем зростають, що в свою чергу впливає на розробку методів та алгоритмів пошуку і подання даних.

3.В свою чергу, розвиток комп'ютерної техніки також тягне за собою суттєве зростання обсягу інформації, що подається в електронному вигляді. Вплив цього процесу на розвиток сучасних інформаційних технологій, включаючи пошук, відзначається в більшості наукових публікацій в періодичних виданнях [2]. Хоча на сьогоднішній день й існує значна кількість методів та алгоритмів інформаційного пошуку, проте неперервний розвиток цієї галузі та вищезазначені проблеми вказують на необхідність постійного покращення існуючих методів та розробку якісно нових підходів. Тож, відповідно, проблема вдосконалення методів інформаційного пошуку є актуальною. Наукова новизна дослідження полягає в тому, що вперше було розглянуто можливість використання методики прецендентів з точки зору інформаційного пошуку та встановлено перспективність даної методики у рамках вдосконалення пошукових алгоритмів.

4.МЕТА РОБОТИ. Головною метою даної роботи є суттєве поліпшення результатів інформаційного пошуку за рахунок застосування методики прецедентів.

5.МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ. У загальному випадку під поняттям "інформаційний пошук" розуміють процес відшукання серед деякої множини текстів (документів) таких, які присвячені саме зазначеній в пошуковому запиті темі, або містять потрібні користувачеві факти чи відомості. Пошук може здійснюватися як вручну, так і за допомогою інформаційно-пошукової системи з використанням засобів автоматизації. Залежно від характеру інформації, яка міститься в документах, які містить пошукова видача, - пошук може бути документальним, в тому числі й бібліографічним чи фактографічним.

6.Проблемам організації інформаційного пошуку було присвячено багато наукових праць. Зокрема, деякі теоретичні аспекти було розглянуто в публікаціях як вітчизняних, так і закордонних дослідників: Урвачова В.А [1], Шокін Ю.І. [2], Климчук С.О. [3], Маннінг К.Д. [4] та ін.

7.Варту уваги інформацію, з точки зору дослідника, містить стаття В.А. Урвачової [1]. У роботі вченої наводиться короткий огляд сучасних методів та алгоритмів інформаційного пошуку. В огляд також включені класичні алгоритми, які покладені в основу сучасних пошукових методів. Окремої уваги заслуговують зроблені нею висновки щодо перспективності застосування технологій інтелектуальних агентів для пошуку інформації. З урахуванням давності викладеної інформації, слід зазначити, що в монографії Ю.І. Шокіна [2] детально розглянуто загальні аспекти розробки та створення інформаційно-пошукових систем. Зокрема, наводиться докладний виклад моделей, структур і алгоритмів, що описують окремі різновиди інформаційно-пошукових систем. Заслуговують уваги розкриті автором перспективи напряму інформаційного моделювання при організації пошуку.

8.У своій роботі [3] С.О. Климчуком пояснюються важливі, з точки зору дослідника, принципи організації прецедентної системи (Case-Based Reasoning System). Зокрема, проаналізовано переваги методики прецедентів у рамках створення інтелектуальних засобів підтримки прийняття рішень. Публікація заслуговує уваги, з урахуванням можливості застосування відповідної методики для побудови системи інформаційного пошуку. Незважаючи на те що роботу К.Д. Маннінга [4] задумано як вступний курс з інформаційного пошуку та написанио з точки зору інформатики; в ньому поряд з класичним пошуком розглядаються веб-пошук, принципи роботи пошукових механізмів а також класифікація та кластеризація текстів. Книга містить сучасний виклад всіх аспектів проектування та реалізації систем збору, індексування та пошуку документів, методів оцінки таких систем, а також введення в методи машинного навчання.

9.Очевидно, що проблема широко обговорюється науковим співтовариством. Однак, попри значну кількість публікацій дослідників, проблема вдосконалення методів інформаційного пошуку не розв'язана повністю та залишається актуальною.

10.Проблема пошуку, збору та оптимізації інформації з'явилася ще в період розвитку пошукових систем [1]. У той час пошукові системи надавали велике значення аспектам, якими власники сайтів могли легко маніпулювати. Це призвело до того, що у видачі багатьох пошукових систем перші кілька сторінок займали сайти, наповнення яких було нерелевантним.

11.У загальномі випадку під терміном "релевантність" розуміють міру відповідності отримуваного результату бажаному. В термінах пошуку - це міра відповідності результатів пошуку завданню, поставленому в пошуковому запиті [5]. Відповідно, нерелевантним називається такий, що був відібраний у результаті інформаційного пошуку, але зміст якого не відповідає запиту користувача.

12.Пошуковими системами, при вирішенні завдань збору, зберігання, обробки і видачі інформації, виконуються такі операції [2]:

1) пошук документів;

2) аналіз вмісту документів;

3) побудова пошукових образів документів (отримання з документів інформації, яка використовується системою як відомості про документ);

4) зберігання пошукових образів документів (відомостей про документи);

5) аналіз запитів користувачів (споживачів інформації);

6) пошук релевантних (відповідних) запиту документів;

7) організація пошукової видачі користувачам.

13.Ефективність інформаційного пошуку характеризується двома відносними показниками: коефіцієнтом точності (відношенням числа документів, що відповідають критеріям інформаційного запиту до загальної кількості документів, отриманих в пошуковій видачі) і коефіцієнтом повноти (відношенням числа документів, що відповідають критеріям інформаційного запиту до загальної кількості таких документів, що містяться в просторі, оброблюваному інформаційно-пошуковою системою) [1]. Необхідні значення цих показників залежать від специфіки інформаційних потреб. Наприклад, якщо відбувається пошук патентних описів з метою проведення експертизи патентної заявки на новизну необхідна 100% повнота результату пошукової видачі, а при пошуку, орієнтованому на звичайного дослідника, або випадкового користувача, то прийнятною вважається точність (релевантність) результату близько 80%, а повнота - близько 50% [4].

14.Однак, при визначенні релевантності пошукові системи в першу чергу звертають увагу на те, скільки разів на сторінці зустрічається фраза, тотожна запиту користувача. Цей параметр називається частотою ключового слова. Чим він вищий, тим релевантнішим вважається сайт.

15.Більшість пошукових машин знаходять величезну кількість "релевантних" сторінок за запитом користувача. Кожен знайдений документ ранжується за ступенем його відповідності до запиту. Релевантність кожного документа оцінюється за допомогою різних технологій: обліку частоти появи на сторінці шуканих слів, "відстані" між шуканими словами, вмісту МЕТА-тегів, просторово-часового контексту документа, популярності ресурсу в рейтингах, використання індексу цитування.

16.Типову організацію машин пошуку можна розглянути на прикладі машини WebCrawler, розробленої в університеті Вашингтон (Сіетл, США). WebCrawler починає процес пошуку нових сайтів з відомих йому документів і переходить за посиланнями на інші сторінки. Він розглядає мережевий простір як орієнтований граф і використовує алгоритм обходу графа, працюючи в такому циклі [1]:

1) знайти новий документ;

2) зазначити документ як вилучений;

3) розшифрувати посилання з цього документу;

4) проіндексувати зміст документу.

18.Пошуковий механізм працює в двох режимах: пошук документів в реальному часі та режимі індексування документів. У режимі індексування система будує індекс інформації зі знайдених документів, в режимі пошуку знаходить документи, які максимально відповідають запиту користувача. Агенти в системі WebCrawler відповідають за вилучення документів з мережі.

19.Для виконання цієї роботи пошуковий механізм знаходить вільного агента і передає йому завдання на пошук. Агент приступає до роботи і повертає або зміст документа, або пояснення, чому документ не може бути доставлений. Агенти запускаються як окремі процеси, що дозволяє ізолювати основний процес роботи системи від помилок і проблем з пам'яттю.

20.Одночасно можуть бути використані до 15 агентів [1]. У базі даних зберігаються метадані документів, зв'язки між документами і повнотекстовий індекс. База оновлюється кожного разу, коли надходить новий документ. Для відсікання семантично незначущих слів існує стоп-словник. Словам з документу приписується вага, рівна частоті їх появи в даному тексті, поділеній на частоту появи слова в посиланнях на інші документи. Такий індекс дозволяє швидко знаходити по заданому слову посилання на документи, що містять його.

21.Аналогічним чином влаштовані і інші машини пошуку. Вони не можуть налаштовуватися на переваги користувача і не мають достатніх ресурсів для аналізу інформації, а мережевим роботам стає все важче справлятися з постійним зростанням кількості Інтернет-ресурсів [5]. Головним завданням машин пошуку є індексація ресурсів глобальної мережі. Фактично в базах даних машин пошуку зберігається інформація про те, де і що лежить в мережі. Тому можна вважати, що існуючі машини пошуку забезпечують низькорівневий сервіс для клієнтсьих пошукових програм більш високого рівня.

22.У загальному випадку вважається, що будь-яка пошукова машина повинна відповідати наступним вимогам [2]:

1) простота у використанні;

2) чітко організований і оновлюваний індекс;

3) швидкий пошук за індексом і швидке реагування;

4) надійність і точність результатів пошуку.

23.В результаті проведеного дослідження було зазначено, що всі сучасні пошукові системи мають деякі серйозні недоліки:

1) стандартний механізм пошуку за ключовими словами в сучасних інформаційно-пошукових системах видає результати з великим показником інформаційного шуму;

2) велика кількість пошукових машин з різними призначеними для користувача інтерфейсами породжує проблему когнітивного перевантаження;

3) методи індексування баз даних, як правило, не пов'язані з інформаційним змістом;

4) часто видаються посилання на інформацію, якої в Інтернеті вже давно немає, а також немає можливості в реальному часі враховувати динаміку зміни змісту Інтернет-ресурсів;

5) в пошукових машинах немає розвинених засобів розуміння природних мовних конструкцій.

24.Також серйозним недоліком сучасних пошукових систем є їх централізація, що вимагає колосальних ресурсів (величезні обсяги бази даних, безліч серверів, апаратури і т.д.) [5]. Ще однією з основних проблем при створенні сучасних інформаційно-пошукових систем є недостатнє врахування думок і бажань користувачів. Існуючі механізми персоналізації володіють певними обмеженнями і не можуть враховувати в повному обсязі потреби користувачів. Явні методи персоналізації вимагають активного залучення користувача в процес накопичення персональних даних (потрібна наявність зворотного зв'язку), що не зовсім зручно для користувача пошукової системи, а неявні методи персоналізації не завжди адекватно можуть реагувати на зміни в перевагах користувачів.

25.Нині досить важливою проблемою в галузі інформаційного пошуку є проблема конструювання інтелектуальних систем, орієнтованих на відкриті і динамічні бази даних [4]. В основі таких систем лежить інтеграція здатних до адаптації, модифікації і навчання моделей пошуку, виявлення та оперування знаннями, орієнтованих на специфіку шуканої (предметної) області та відповідний тип невизначеності, що відображає їх здатність до розвитку і зміни свого стану. Організація пошуку на основі прецедентів дозволяє об'єднати в собі різні підходи до вирішення завдання інтелектуалізації та персоналізації пошуку.

26.У більшості енциклопедичних джерел термін "прецедент" визначається як випадок, що стався раніше і слугував прикладом або виправданням для наступних випадків подібного роду [3]. Міркування на основі прецедентів (Case-Based Reasoning) являється методикою здатною вирішити нове або невідоме завдання, використовуючи або адаптуючи рішення вже відомої задачі, тобто використовуючи вже накопичений досвід вирішення подібних завдань. Підхід на основі прецедентів виник в процесі розвитку досліджень в області створення експертних систем (систем, заснованих на знаннях).

27.Як правило, процес виведення на основі прецедентів підлягає декомпозиції на чотири основні етапи, що утворюють так званий цикл міркування на основі прецедентів [3]. Основними етапами такого-циклу є:

1) вилучення найбільш відповідного (подібного) прецеденту (або прецедентів) до ситуації, що склалася з бібліотеки прецедентів;

2) повторне використання вилученого прецеденту для спроби вирішення поточної проблеми;

3) перегляд і адаптація в разі необхідності отриманого рішення відповідно до проблеми;

4) збереження нового прийняття рішення як частини нового прецеденту.

28.До переваг міркувань на основі прецедентів можна віднести наступні аспекти [3]:

1) можливість безпосередньо використовувати досвід, накопичений системою без інтенсивного залучення експерта в тій чи іншій предметної області;

2) можливість скорочення часу пошуку рішення поставленої задачі за рахунок використання вже наявного рішення для такого завдання;

3) можливість виключити повторне отримання помилкового рішення;

4) відсутня необхідність повного та поглибленого розгляду знань відносно конкретної предметної області.

29.До недоліків міркувань на основі прецедентів можна віднести наступне [3]:

1) при описі прецедентів зазвичай обмежуються поверхневими знаннями про предметну область;

2) велика кількість прецедентів може привести до зниження продуктивності системи;

3) проблематичним є визначення критеріїв для індексації і порівняння прецедентів;

4) проблеми з налагодженням алгоритмів визначення подібних (аналогічних) прецедентів;

5) неможливість отримання рішення задач, для яких немає прецедентів або ступінь їх схожості (подібності) менше заданого порогового значення.

30.Основна мета використання апарату прецедентів в інформаційно-пошуковій системі полягтиме в видачі відповіді на запит користувача на основі прецедентів, які вже мали місце в минулому при виконанні подібних запитів. Інформація про новий запит використовуватиметься для вилучення з бібліотеки прецедентів найбільш підходящого прецеденту (прецедентів). Витягнутий прецедент використовується повторно для отримання рішення нової проблеми (завдання) [3]. Потім запропоноване рішення в разі необхідності може бути адаптоване до особливостей нової ситуації і застосовано на практиці. У разі успішного застосування, перевірене рішення спільно з описом запиту утворює новий прецедент, який зберігається в базі прецедентів.

31.Вибір методу отримання прецедентів безпосередньо пов'язаний зі способом уявлення прецедентів і відповідно зі способом організації бібліотек прецендентів [3]. Останні є важливою складовою бази знань інтелектуальної системи, але можуть виступати як і окремий компонент пошукової системи. Таким чином, їхня структура істотньо впливає на різні показники роботи системи і, зокрема, на час пошуку та вилучення прецедентів.

32. Існують різні способи подання та зберігання прецедентів: від простих (лінійних) до складних ієрархічних. Варто зазначити, що прості способи зберігання та подання прецедентів, що базуються на технології реляційних баз даних, вимагають значно менше витрат на реалізацію, а також підтримку і супровід бібліотек прецедентів системи на відміну від більш складних, але може знадобитися значно більше часу для здійснення пошуку рішення при простому поданні прецедентів порівняно з іншими способами представлення і збереження прецедентів.

33.Прецеденти можуть бути представлені у вигляді списку параметрів, концептуальних графів, семантичної мережі, деревовидних структур, предикатів, фреймів, малюнків і мультимедійної інформації [3]. Прецедент може включати наступні компоненти:

1) опис завдання (проблеми або проблемної ситуації);

2) рішення задачі;

3) результат застосування рішення.

34.Опис результату може включати список виконаних дій, додаткові коментарі та посилання на інші прецеденти. прецедент може мати як позитивний, так і негативний результат застосування рішення, а також в деяких випадках може приводитися обгрунтування вибору даного рішення і можливі альтернативи.

35.Слід зазначити, що у простих задачах класифікації деякі відмінності просто ігноруються і клас рішення витягнутого прецеденту переноситися на клас рішення нового прецеденту [3]. Однак, багато систем, враховують відмінності між знайденим і наявним прецедентом, і тому рішення витягнутого прецеденту не може бути безпосередньо перенесено на нову ситуацію.

36.Для успішної реалізації пошуку за допомогою міркувань на основі прецедентів, відповідно, необхідно забезпечити коректне вилучення прецедентів з бібліотеки прецедентів системи.

37.Існують різні способи такого отримання прецедентів, наприклад [3]:

1) метод найближчого сусіда і його модифікації;

2) метод пошуку на основі дерев рішень;

3) метод вилучення на основі знань;

4) метод вилучення з урахуванням застосування прецедентів.

38.З точки зору інформаційного пошуку було прийнято рішення скористатися методом найближчого сусіда. Одним з ключових моментів даного вибору стало те, що даний метод є найпоширенішм з методів порівняння і вилучення прецедентів [3]. Він дозволить досить легко обчислити ступінь подібності поточного запиту і прецедентів з бібліотеки системи. Принцип роботи полягає у тому, що з метою визначення ступеня подібності на множині параметрів, використовуваних для опису прецедентів і поточного запиту, вводиться певна метрика. Далі відповідно до обраної метрики визначається відстань від цільової точки, відповідної поточному запитуї, до точок, що представляють прецеденти з бібліотеки прецедентів і вибирається найближча до цільової точка.

39.Безумовно, ефективність методу найближчого сусіда багато в чому залежить від вибору метрики (міри схожості).

40.Основні метрики, якы можуть бути використані в методі найближчого сусіда [3]:

1) Евклідова відстань. Евклідова відстань є геометричною відстанню в багатовимірному просторі; відстань між точками C і T в n-вимірному просторі визначається за наступною формулою (1):

(1)

де d (C, T) - відстань між C і T, x(1,2,3 ... n) та y(1,2,3 ... n) - значення ознак для прецедента та поточної ситуації, n - кількість змінних, якими описуються прецеденти та поточна ситуація.

2) Відстань Хемінга. Найпростішою мірою схожості, а точніше кажучи, відмінності між закодованими представленнями є відстань Хемінга. Хоча першочергово вона була введена для двійкового коду, вона цілком може примінятися для порівняння будь-яких упорядкованих наборів, які складаються з елементів, здатних набувати дискретних значень.

41.Розглянемо для прикладу два упорядковані набори x та y які складаються з дискретних, нечислових символів (наприклад логічних 0 та 1). Порівняння на несхожість полягатиме у кількості неспівпадаючих символів у цих наборах. Таку величину d(C, T) називають відстанню Хемінга; визначається вона лише для послідовностей однакової довжини. Тобто якщо x = (1,0,1,1,1,0), y = (1,1,0,1,0,1), то d(C, T) = 4.

3) Міра схожості Танімото. В загальному випадку може примінятися для оцінки релевантності (ступеня відповідності) документів при інформаційному пошуку. Дескрипторам в цих документах можуть бути присвоєні індивідуальні ваги.

42.Якщо aik - вага, відносна до k-го дескриптору i-го документу, то ступінь схожості двох документів визначених через C і T можна визначити за формулою (2):

(2)

при цьому вираз набуде наступного вигляду (3):

(3)

43.Як вже наголошувалося, вибір відповідної метрики творче і досить трудомістке завдання, від успішного вирішення якої безпосередньо залежить результативність пошуку та вилучення прецедентів.

44.Для вирішення поставленої задачі інформаційного пошуку [5] було обрано у якості метрики використовувати евклідову відстань. Відповідно, спираючись на дану метрику було створено та запропоновано відповідний алгоритм вилучення прецендентів для організації інформаційного пошуку.

45.Нехай заданий прецедент (C) і поточна ситуація (пошуковий запит) (T) в n-вимірному просторі ознак (властивостей), тоді ступінь подібності або близькості можна визначити, використовуючи евклідову метрику для визначення відстані між C і T (4).

(4)

46.Вхідні дані: поточна ситуація (T), бібліотека прецедентів - непорожня множина прецедентів СL, m - кількість прецедентів в бібліотеці, порогове значення ступеня схожості K.

47.Вихідні дані: тимчасовий контейнер зберігання прецедентів SC порогового значення К.

Алгоритм:

Крок 1. Для визначення значення ступеня подібності S(C,T) необхідно знайти максимальне значення відстані dmax в евклідовій метриці, використовуючи границі діапазонів параметрів для описання прецедентів (xпоч та xкін , i = 1, ..., n).

Крок 2. Поки j ? m обираємо прецедент Сj з множини CL та переходим до наступного кроку, інакше вважаємо що всі прецеденти бібліотеки розглянуті й переходимо до кроку 6.

Крок 3. Розраховуємо відстань в евклідовій метриці між обраним прецедентом Сj та поточною ситуацією Т (5) та переходимо до наступного кроку.

(5)

Крок 4. Обчислюємо ступінь подібності (6) S(Cj,T); переходимо до наступного кроку.

(6)

Крок 5. Якщо S(Cj,T) ? К , то id-номер даного прецеденту розміщуємо в SC (тимчасовому контейнері для зберігання прецедентів, які задовольняють умову) та переходимо до кроку 2.

Крок 6. Якщо після всіх ітерацій контейнер SC залишається пустим, значить прецеденти для поточної ситуації не знайдені і потрібно перейти до кроку 7 та передати користувачеві повідомлення про необхідність зниження порогового значення К; в іншому випадку вважається що прецеденти для поточної ситуації успішно вилучені із бібліотеки за id-номерами, що зберігаються в тимчасовому контейнері SC та переходимо до наступного кроку.

Крок 7. Кінець (завершення алгоритму).

48.Слід також зазначити, що при повторному використанні знайденого прецедента в контексті нової проблемної ситуації важливо звернути увагу на наступні особливості: різниця між вилученим та новим прецедентом а також те яку частину вилученого прецеденту можна застосувати до поточної ситуації (пошукового запиту).

49.Таким чином, можна зробити допущення щодо перспективності використання даної методики у рамках вдосконалення системи інформаційного пошуку. Відповідно, між побудованими таким чином інформаційно-пошуковими системами (з метою їх навчання) та користувачами повинен встановлюватися ефективно працюючий зворотний зв'язок (абонент повідомляє, якою мірою цей документ відповідає запиту і чи потрібно продовжувати пошук, вказує на ступінь відповідності цього документа його інформаційним потребам), який дозволяє уточнювати потреби абонентів, своєчасно реагувати на зміни цих потреб і оптимізувати роботу системи. Таким чином, можна підсумувати, що використовуючи методику прецедентів [3] при ретроспективному пошуку, пошуковою системою знаходитимуться в першу чергу документи, які містять необхідну інформацію.

50.ВИСНОВКИ. В результаті проведеного дослідження було зазначено, що всі сучасні пошукові системи мають деякі серйозні недоліки:

1) стандартний механізм пошуку за ключовими словами в сучасних інформаційно-пошукових системах видає результати з великим показником інформаційного шуму;

2) велика кількість пошукових машин з різними призначеними для користувача інтерфейсами породжує проблему когнітивного перевантаження;

3) методи індексування баз даних, як правило, не пов'язані з інформаційним змістом;

4) часто видаються посилання на інформацію, якої в Інтернеті вже давно немає, а також немає можливості в реальному часі враховувати динаміку зміни змісту Інтернет-ресурсів;

5) в пошукових машинах немає розвинених засобів розуміння природних мовних конструкцій.

51.Виходячи з проведеного огляду сучасного стану досліджень встановлено що все очевиднішими стають потреби в розробці розвинених засобів інтелектуалізації та персоналізації пошуку. Окрім того, встановлено, що хоча проблема широко обговорюється науковим співтовариством; попри значну кількість публікацій дослідників, проблема вдосконалення методів інформаційного пошуку не розв'язана повністю та залишається актуальною

52.На сьогоднішній день практично всі сучасні інформаційно-пошукові системи Інтернету активно працюють у сфері розробки інструментів інтелектуалізації та персоналізації пошуку [10], але більшість серйозних проблем в цій галузі поки не вирішені. Відповідно, перспективним напрямом залишається дослідження і розробка методів та програмних засобів інтелектуалізації та персоналізації в інформаційно-пошукових системах Інтернету.

53.В ході дослідження встановлено перспективність використання методики прецедентів у рамках вдосконалення пошукових методів та, зокрема, при побудові орієнтованих на розподілену структуру інформаційно-пошукових систем. Окрім того, автором виділено основні напрями для розробки алгоритму вилучення прецедентів в рамках організації інформаційного пошуку. Наголошено, що організація пошуку на основі прецедентів дозволяє об'єднати в собі різні підходи до вирішення завдання інтелектуалізації та персоналізації пошуку і знизити навантаження на індекс пошукового інструменту, а також спростити вирішення проблеми, пов'язаної із забезпеченням конфіденційності даних.

54.Висновки та пропозиції в рамках даного дослідження можуть бути використані в науково-дослідній та викладацькій діяльності. Зокрема, результати, отримані при проведенні даного дослідження можуть бути використанні при подальшому аналізуванні та вдосконаленні методів інформаційного пошуку.

**Додаток Б**

Скорочений текст

Розглянемо для прикладу два упорядковані набори x та y які складаються з дискретних, нечислових символів (наприклад логічних 0 та 1). Порівняння на несхожість полягатиме у кількості неспівпадаючих символів у цих наборах. Таку величину d(C, T) називають відстанню Хемінга; визначається вона лише для послідовностей однакової довжини. Тобто якщо x = (1,0,1,1,1,0), y = (1,1,0,1,0,1), то d(C, T) = 4.

3) Міра схожості Танімото. В загальному випадку може примінятися для оцінки релевантності (ступеня відповідності) документів при інформаційному пошуку. Дескрипторам в цих документах можуть бути присвоєні індивідуальні ваги.

Якщо aik - вага, відносна до k-го дескриптору i-го документу, то ступінь схожості двох документів визначених через C і T можна визначити за формулою (2):

при цьому вираз набуде наступного вигляду (3):

Як вже наголошувалося, вибір відповідної метрики творче і досить трудомістке завдання, від успішного вирішення якої безпосередньо залежить результативність пошуку та вилучення прецедентів.

Для вирішення поставленої задачі інформаційного пошуку [5] було обрано у якості метрики використовувати евклідову відстань. Відповідно, спираючись на дану метрику було створено та запропоновано відповідний алгоритм вилучення прецендентів для організації інформаційного пошуку.

Слід також зазначити, що при повторному використанні знайденого прецедента в контексті нової проблемної ситуації важливо звернути увагу на наступні особливості: різниця між вилученим та новим прецедентом а також те яку частину вилученого прецеденту можна застосувати до поточної ситуації (пошукового запиту).