ДОДАТОК А

Слайди презентації

АТЕСТАЦІЙНА РОБОТА МАГІСТРА

Дослідження методів семантичного аналізу для пошукових механізмів

Виконала:

Ст. гр. ПЗСм-18-1

Хікматова Д. М.

Керівник:

д. т. н. Четвериков Г. Г.

Рисунок А.1 – Титульний слайд

Рисунок А.2 – Слайд «Семантичний аналіз у пошукових механізмах»

СЕМАНТИЧНИЙ АНАЛІЗ У ПОШУКОВИХ МЕХАНІЗМАХ

Інформаційний пошук – це процес пошуку неструктурованих даних, які задовольняють запиту, тобто ті, які містять необхідні факти, відомості чи дані.

Роль семантичного аналізу:

- Вирішення синонімії та омонімії;
- Виявлення зв’язків між словами у запиті;
- «Розуміння» змісту нових слів;
- Підбір результатів пошуку;
- Тематичне моделювання для розвідувального пошуку;
- ...

2

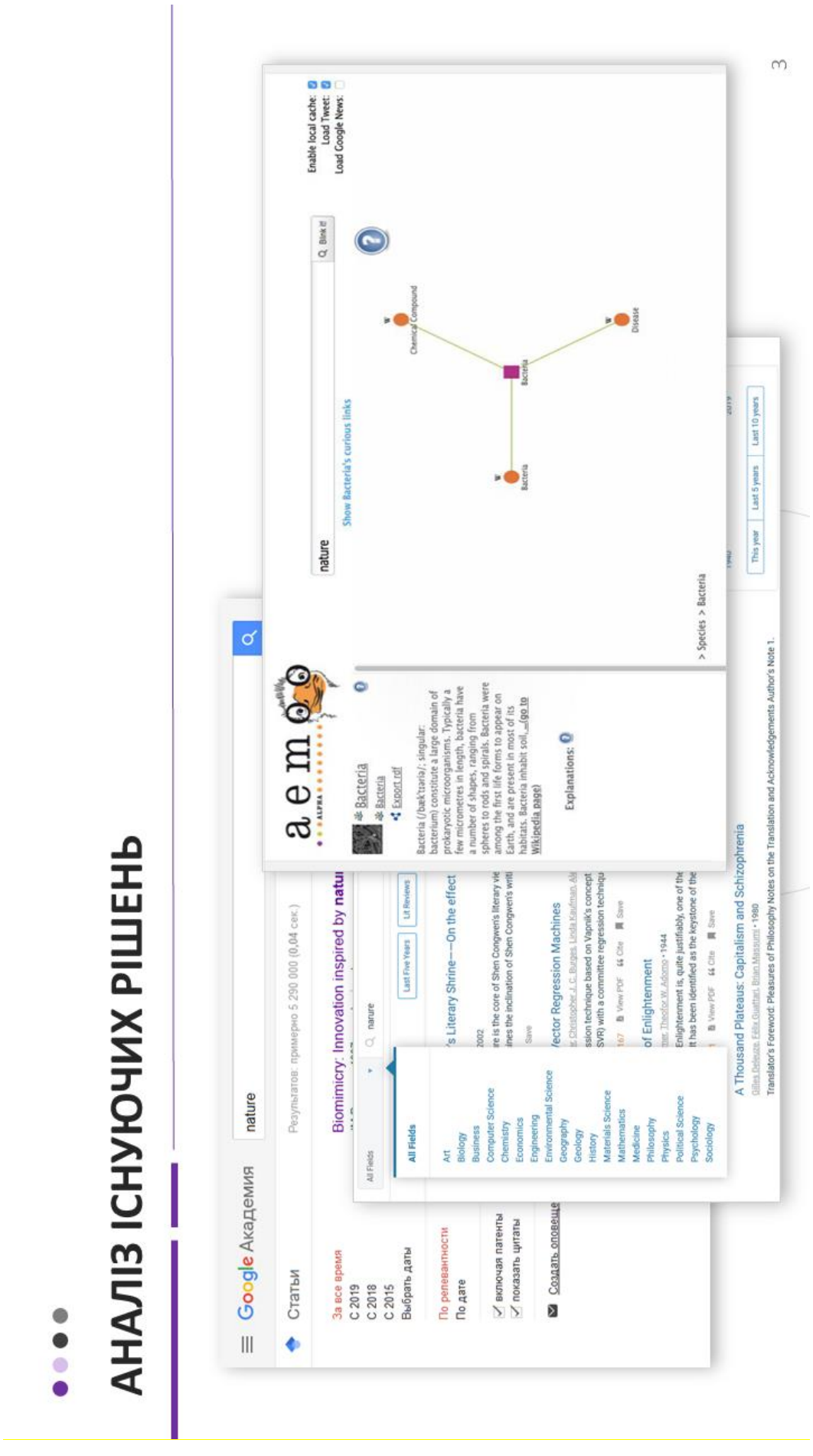


Рисунок А.3 – Слайд «Аналіз існуючих рішень»

ПОСТАНОВКА ЗАДАЧІ

- проаналізовані методоти тематичного моделювання та обрати найкращий;
- підібрати оптимальні параметри для отпимізації роботи системи та підвищення якості виділених тем;
- проаналізувати роботу бібліотек для автоматичного виправлення помилок у текстах;
- порівняти роботу методів оцінки складності текстів для сприйняття та обрано найкращий;
- розробити прототип системи розвідувального пошуку, яка при підборі результатів сортуватиме матеріали за складністю сприйняття.

Рисунок А.4 – Слайд «Постановка задачі»

ІМОВІРНІСНА ТЕМАТИЧНА МОДЕЛЬ

Розподілення термів у документі:

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td},$$

де $p(w|t,d)$ – імовірність появи терму w в документі d, що пов'язано з темою t;

$p(t|d)$ – імовірність належності документу d до теми t;

$p(w|t)$ – імовірність належності терму w до теми t.

Частотні оцінки імовірностей:

$$p(d,w) = \frac{n_{dw}}{n},\ p(d) = \frac{n_d}{n},\ p(w) = \frac{n_w}{n},\ p(w|d) = \frac{n_{dw}}{n_d},$$

де n_{dw} – число входжень терма w у документ d;
$n_d = \sum_w n_{dw}$ – довжина документа d в термах;
$n_w = \sum_d n_w$ – число входжень терма w у всі документи колекції;
$n = \sum_d \sum_w n_{dw}$ – довжина колекції у термах.

Рисунок А.5 – Слайд «Імовірніста тематична модель»

ОЦІНКА СКЛАДНОСТІ ТЕКСТІВ

Метод Фліша: $complexity = 206.835 - 1.015\left(\frac{totalWords}{totalSentences}\right) - 84.6\frac{totalSyllables}{totalWords}$

Метод Дейла-Челла: $complexity = 0.1579\left(\frac{complexWords}{totalWords} * 100\right) + 0.0496\frac{totalWords}{totalSentences}$

Метод Ганнінга: $complexity = 0.4\left(\frac{totalWords}{totalSentences} + 100\frac{complexWords}{totalWords}\right)$

Рисунок А.6 – Слайд «Оцінка складності текстів»

ІНСТРУМЕНТИ РЕАЛІЗАЦІЇ

ІНСТРУМЕНТИ ДОСЛІДЖЕННЯ

Мова програмування: Python

- бібліотека NLTK для попередньої обробки текстів природної мови;
- бібліотека textStat класифікатору визначення складності тексту;
- бібліотека Pandas для аналізу та візуалізації даних;
- бібліотека BigARTM для тематичного моделювання.

ІНСТРУМЕНТИ РЕАЛІЗАЦІЇ ПРОТОТИПУ СИСТЕМИ

Мова програмування Python

Bootstrap 4, JavaScript, HTML & CSS для розробки Front-End частини

Рисунок А.7 – Слайд «Інструменти реалізації»

РЕЗУЛЬТАТИ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ

Датасет:

463 000 статті із вікіпедії різних тематик із видаленими стоп-словами та приведені до формату «мішка слів»

Приклади тем (20 тем):

- party, political, law, president, court, election, minister, police, union, democratic;
- king, russian, roman, empire, republic, language, india, china, kingdom, emperor;
- army, air, force, military, aircraft, ship, battle, navy, forces, command;
- species, water, food, fish, plant, white, sea, black, areas, region;
- awards, magazine, award, business, show, radio, california, texas, million, america;

Показники розрідженості матриць без регуляризатору:

ϕ – 0.1817848

Θ – 0.00433088

Показники розрідженості матриць із регуляризатором з tau -8*1e5:

ϕ – 0.6884208

Θ – 0.0052103

Рисунок А.8 – Слайд «Результати тематичного моделювання»

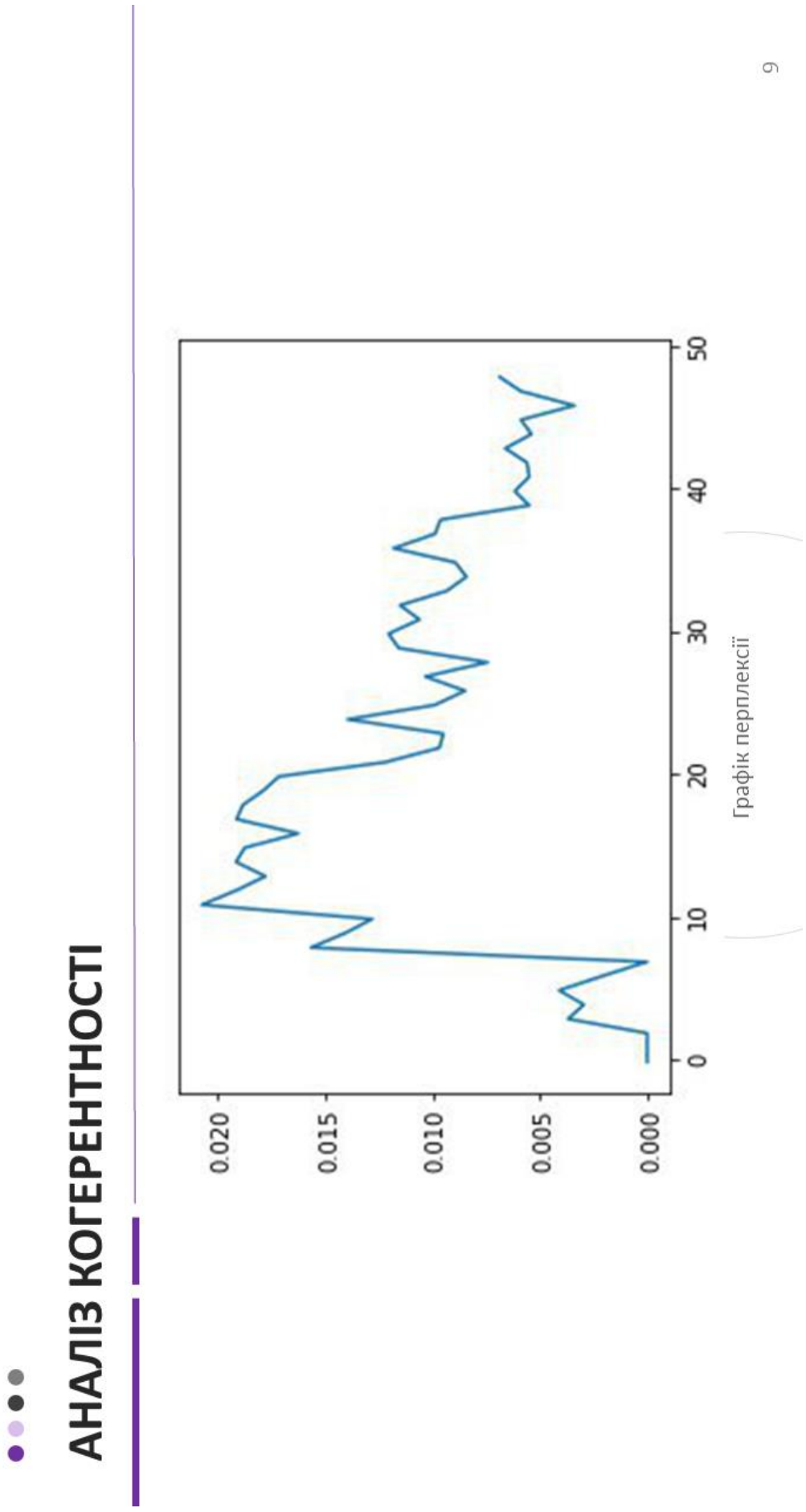


Рисунок А.9 – Сладй «Аналіз когерентності»

Рисунок А.10 – Слайд «Результати в залежності від кількості тем»

РЕЗУЛЬТАТИ В ЗАЛЕЖНОСТІ ВІД КІЛЬКОСТІ ТЕМ

10 ТЕМ:

Розрідженість ϕ : 0.5211899

Sparsity Θ: 0.0020706

party, law, president, political, court, election, minister, police, union, said,

station, game, army, railway, force, military, ship, battle, regiment, road,

film, album, song, band, show, you, episode, love, tv, television,

river, water, species, road, lake, km, jpg, population, often, areas,

air, space, system, aircraft, design, level, program, model, development, different,

district, linear, church, town, socorro, park, building, village, street, population,

25 ТЕМ:

Розрідженість ϕ : 0.7230567

Sparsity Θ: 0.0072765

army, air, force, military, aircraft, battle, navy, forces, command, ship,

district, municipality, polish, center, president, governor, fort, region, population, washington,

film, episode, tv, show, television, character, man, story, movie, role,

station, river, road, park, railway, street, opened, bridge, highway, airport,

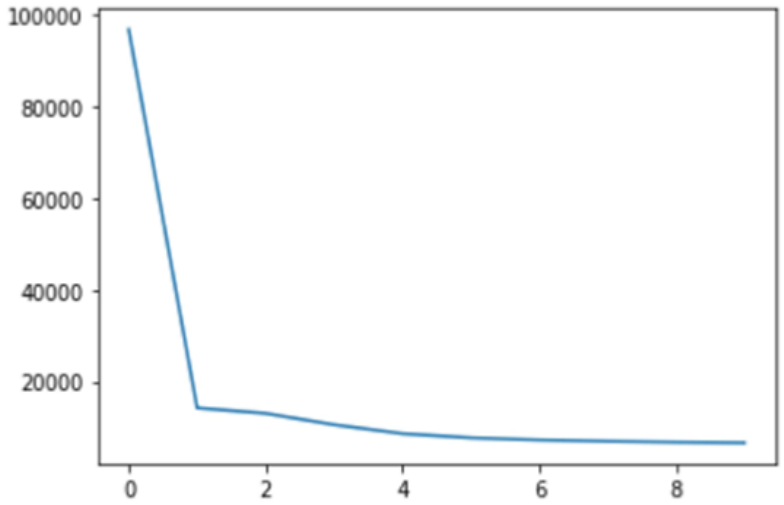
album, song, band, you, records, chart, songs, track, live, guitar,

system, data, systems, using, software, version, computer, standard, model, available,

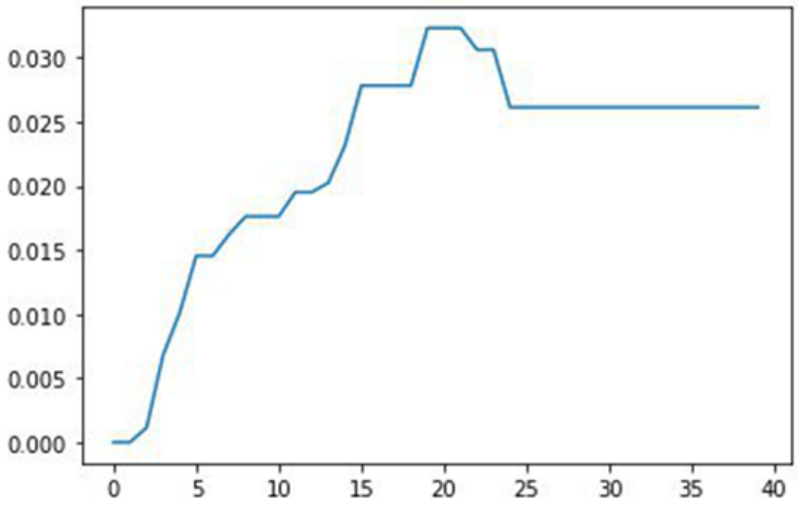
10

Рисунок А.11 – Слайд «Оптимальна кількість проходів по колекції»

ОПТИМАЛЬНА КІЛЬКІСТЬ ПРОХОДІВ ПО КОЛЕКЦІЇ



Графік залежності кількості проходів та перплексії



Графік залежності кількості проходів та когеретності

РЕЗУЛЬТАТИ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ - ВИСНОВКИ

У результаті дослідження було виявлено, що найбільш оптимальних показників вдається досягти із кластеризацією за 25 темами, застосуванню регуляризатору розрідження із значенням tau = -8*1e5, а також при 20 проходах по колекції.

При таких параметрах вдалося досягти збільшення показників розрідженості матриць ф та ѳ у кілька разів, покращенню інтерпетуємості тем та збільшенню їх кількості, а отже, і покращення підбору результатів пошуку.

Рисунок А.12 – Слайд «Результати тематичного моделювання - висновки»

Рисунок А.13 – Слайд «Порівняння результатів оцінки складності текстів»

ПОРІВНЯННЯ РЕЗУЛЬТАТІВ ОЦІНКИ СКЛАДНОСТІ ТЕКСТІВ

	0	1	dale_chall	gunning_fog	flesch_kincaid
4980	wikipedia-194230	Why Can't We Be Friends? "Why Can't We Be Fr...	6.99	6.29	4.2
8868	wikipedia-5657238	Maayavi Maayavi is a 2005 Tamil comedy-drama...	6.69	6.83	4.7
7079	wikipedia-3879432	Avacha Bay Avacha Bay () is a Pacific Ocean ...	6.17	8.28	4.9
7442	wikipedia-14978382	Touchdown Club of Columbus The Touchdown Clu...	5.87	6.00	5.0
9420	wikipedia-1222346	Battle Cry of Freedom The "Battle Cry of Fre...	5.82	6.67	5.1
...	...	...	...	...	...
9815	wikipedia-8707922	Anatomical neck of humerus The anatomical ne...	11.78	36.93	33.7
2753	wikipedia-6538673	Bronze Bauhinia Star The Bronze Bauhinia Sta...	11.76	38.02	36.0
2138	wikipedia-39256416	Index of Andhra Pradesh-related articles Thi...	13.10	39.01	40.8
2018	wikipedia-35045680	Meaningful learning Meaningful learning is o...	13.97	48.00	47.0
9732	wikipedia-8614989	Vestibule of the ear Definition. The vestibu...	12.84	52.84	50.9

Оцінка складнодні читання текстів за різними формулами

РЕЗУЛЬТАТИ АНАЛІЗУ ФОРМУЛ ОЦІНКИ СКЛАДНОСТІ ТЕКСТІВ

Загалом, формули показували дуже схожі оцінки складності, особливо для найскладніших текстів. Найпростіші тексти дещо різнилися, але їх складність сприйняття дійсно корелювала із оцінкою за формулами.

Швидкість роботи формул (10 000 текстів):

- Фліша – 7.18 с
- Дейла-челла – 8.7 с.
- Ганнінга – 6.46 с.

	dale_chall	gunning_fog	flesch_kincaid
dale_chall	1.000000	0.669183	0.542440
gunning_fog	0.669183	1.000000	0.906679
flesch_kincaid	0.542440	0.906679	1.000000

Графік кореляції формул оцінки складості текстів

Рисунок А.14 – Слайд «Результати аналізу формул оцінки складності текстів»

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ - ВИСНОВКИ

- Застосування регуляризатору розрідженості допомагає покращити тематичну модель та оптимізувати матриці ф та Θ.
- Когерентність корелює із інтерпретовністю тем. Для визначення оптимальної кількості тем можна спиратися на графік когерентності, також бажано приділяти увагу значенням, які лежать трохи за межами зони із високою когеретністю. Збільшення кількості проходів по документу покращень не дають.
- При виборі кількості проходів по колекції, краще спиратися не сходження моделі (графік перплексії), але і на відношення когерентності та кількості проходів по колекції.
- Варто із обережністю використовувати бібліотеки автоматичного виправлення помилок, оскільки вони погано працюють із іменованими сутностями.
- Застосування формул оцінки складності тексту для сприйняття може бути застосовано у якості параметру для сортування результатів розвідувального пошуку. Суттєвої різниці у якості оцінки формул Фліша, Дейла-Челла та Ганнінга не виявлено.

Рисунок А.15 – Слайд «Результати досліджень – висновки»

ПРОТОТИП ПОГРАМНОЇ СИСТЕМИ

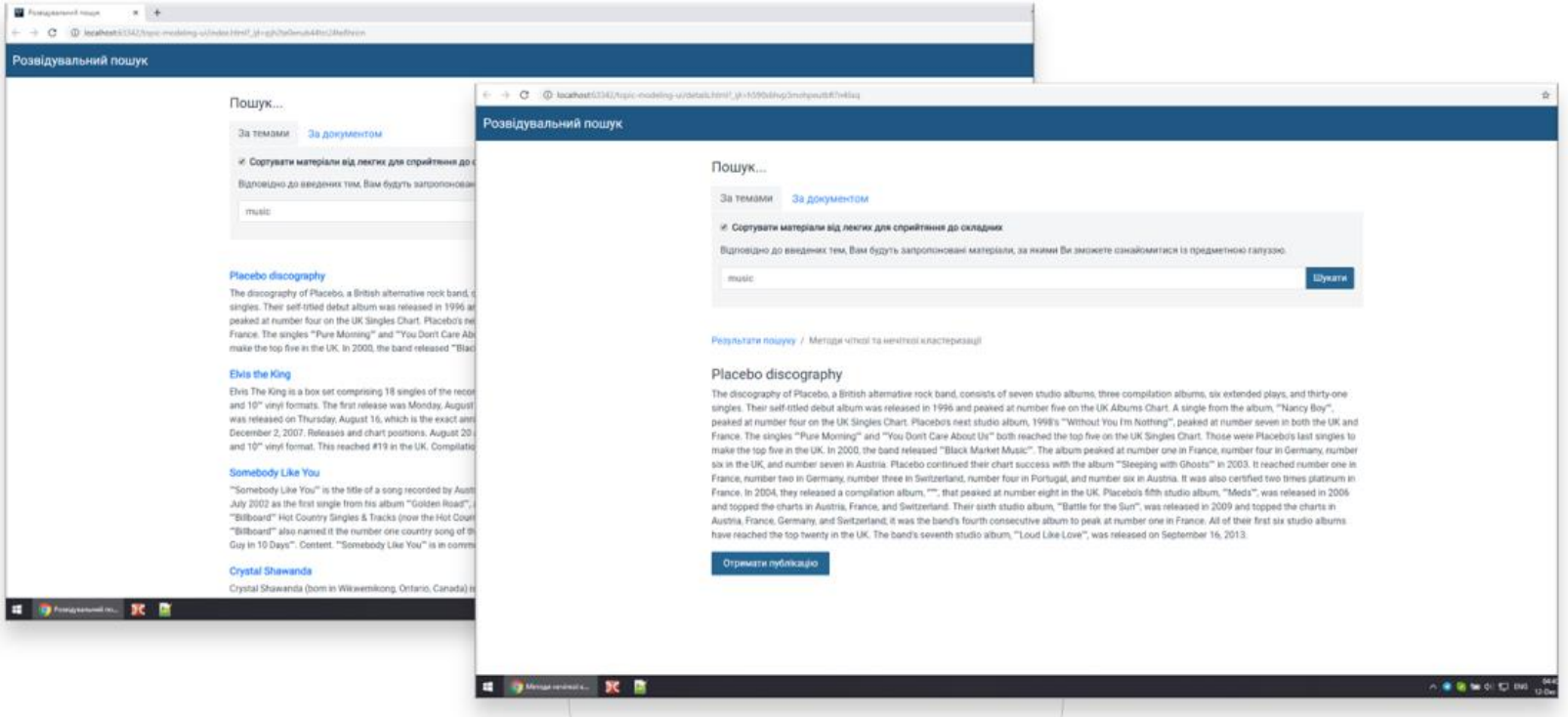


Рисунок А.16 – Слайд «Прототип програмної системи»

ВИСНОВКИ

У межах виконання атестаційного роботи були виконані наступні завдання:

- проведено аналіз предметної галузі;
- проаналізовані методи тематичного моделювання;
- підібрано оптимальні параметри, кількість тем та застосовані адитивні регуляризатори для оптимізації роботи системи для покращення тематичного моделювання;
- досліджено вплив кількості проходів по колекції на інтерпретуємість тематичних моделей;
- проаналізовано роботу бібліотек для автоматичного виправлення помилок у текстах;
- порівняно роботу методів оцінки складності текстів для сприйняття та обрано найкращий;
- розроблено прототип розвідувальної системи.

Рисунок А.17 – Слайд «Висновки»

ДОДАТОК Б

Відгук та рецензії