

Інтелектуальна система побудови контекстної реклами. Отримання, аналіз та апроксимація даних користувача

ПЕРЕПЕЛИЧНИЙ Я. І.

Науковий керівник: к.т.н., доц. Дідковська М.В.

2

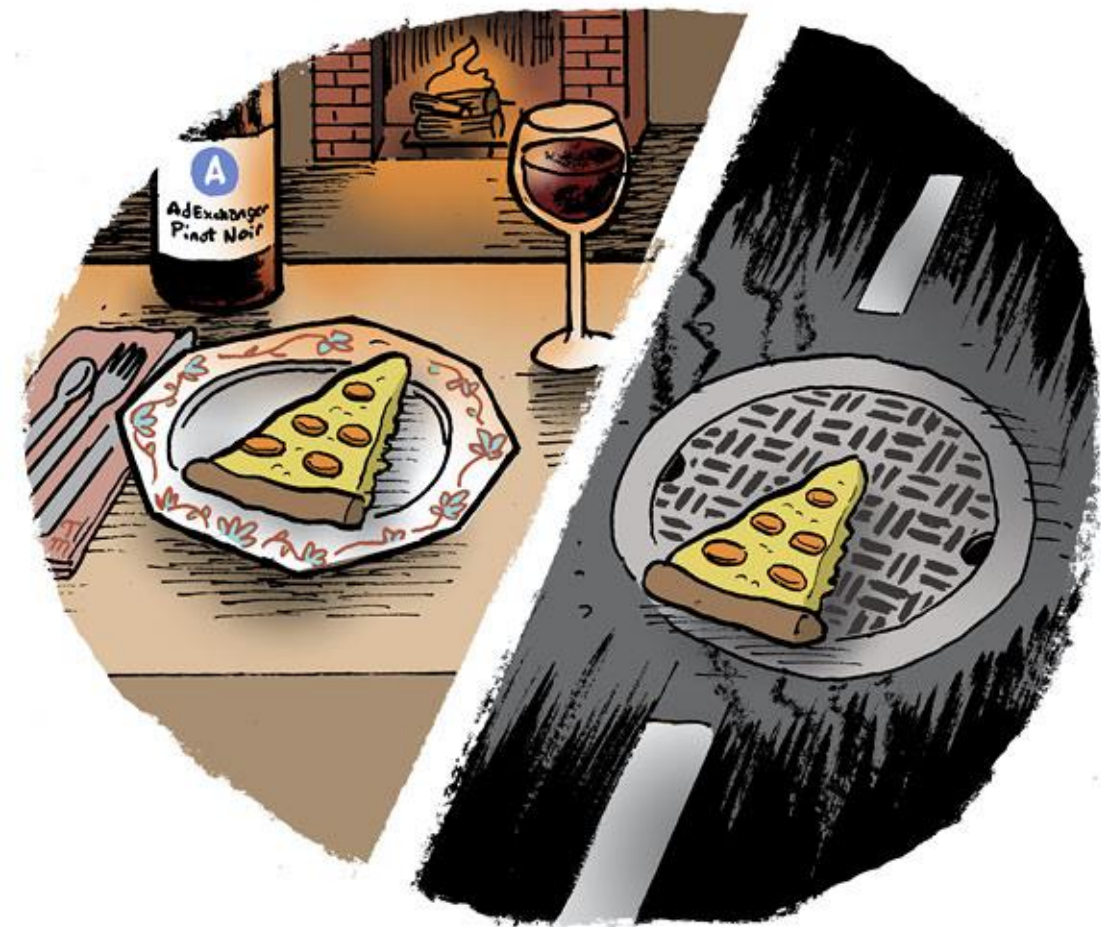




Вартість показу одного ролику на телебаченні буде перевищувати місячний бюджет контекстної рекламної кампанії.

Цільова аудиторія (контекст має значення)

4



Ваше рекламне оголошення буде показано людям, які цікавляться даною тематикою

Context Matters

Постановка задачі

5

Головною цілю даної роботи є розробити інтелектуальну систему побудови контекстної реклами.

Для досягнення цих цілей були поставлені наступні завдання:

1. Провести аналіз існуючих систем;
2. Розробити парсер, який збирає основну інформацію з HTML сторінки;
3. Проаналізувати існуючі алгоритми щодо класифікації тексту;
4. Розробити модуль для класифікації тексту;

Об'єкт та предмет дослідження

6

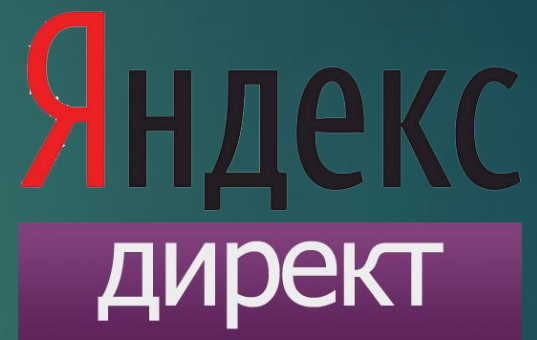
Об'єктом дослідження є Text Mining

Предметом дослідження є підходи до побудови контекстної реклами.



Існуючі підходи.

7



Реалізація парсеру

8

- ▶ Парсинг (від англ. Parse) – процес аналізу або розбору певного контенту на складові за допомогою роботів-парсерів (спеціальних програм або скриптів).
- ▶ Реалізований парсер в даній роботі збирає текст з спеціальних HTML тегів, таких як <meta>, з атрибутами keywords та description, <h1> та з атрибутом alt. Саме ці теги використовуються для пошукової оптимізації сайту, в них міститься основна інформація про дану HTML сторінку.



- ▶ Перш ніж класифікувати текст необхідно виключити стоп-слова або шумові слова. Це слова, які не несуть смислового навантаження, тому їх користь та роль для пошуку не суттєва. Наприклад для англійської мови це можуть бути a, an, and, for, from і так далі

Основні методи класифікації текстів:

- ▶ класифікація за допомогою дерев рішень;
- ▶ Байєсівська (наївна) класифікація;
- ▶ класифікація методом опорних векторів;
- ▶ класифікація за допомогою методу найближчого сусіда.

Байєсівська (наївна) класифікація

11

► Формула Байєса: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$;

► Основна задача: $\operatorname{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)}$ де c — клас документа, d — документ;

$$P(d|c) = \prod_{i=1}^n P(w_i|c), \text{ де } w_i \text{ — } i\text{-те слово документа;}$$

$$P(c) = \frac{D_c}{D}, \text{ де } D_c \text{ — кількість документів класу } C,$$

D — загальна кількість документів в навчальній вибірці.

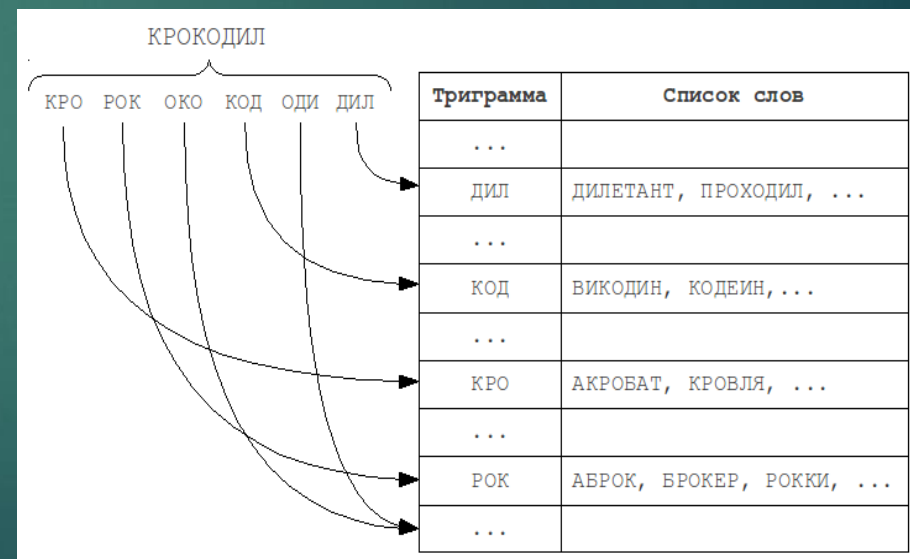
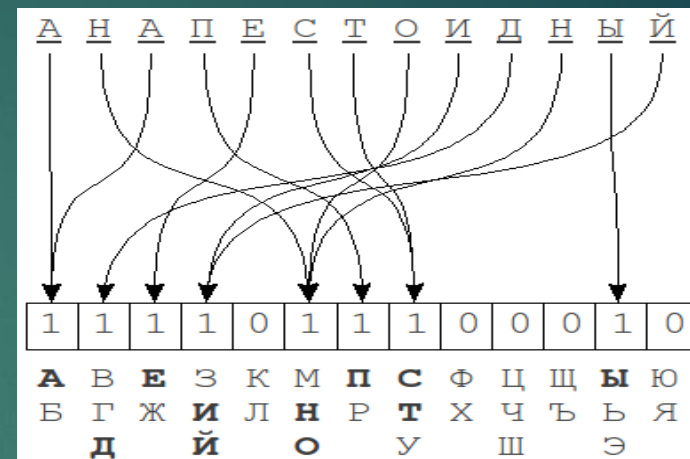
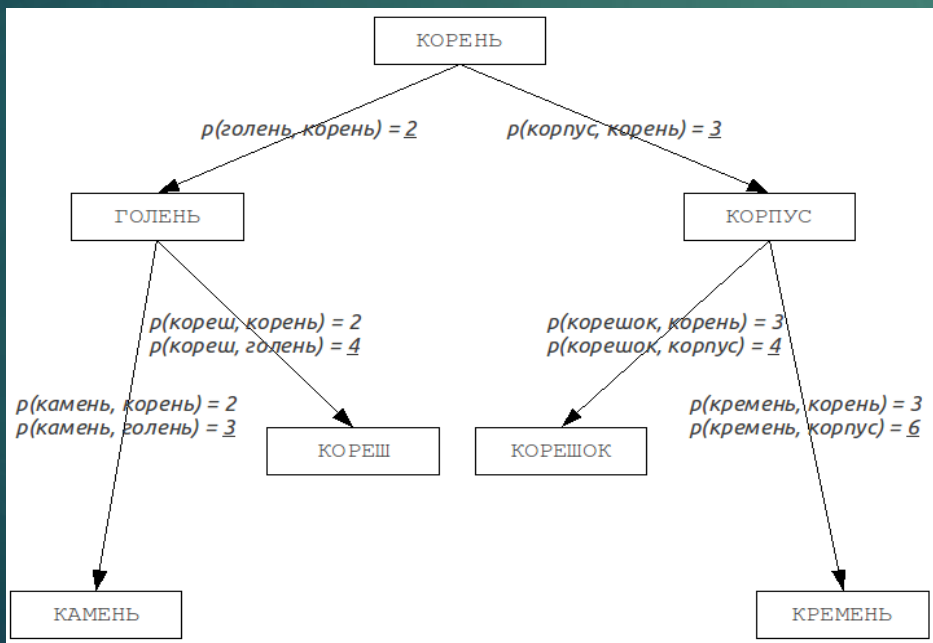
$$P(w_i|c) = \frac{W_{ic}}{\sum_{i' \in V} W_{i'c}} \text{ де } W_{ic} \text{ — кількість разів скільки } i\text{-те слово зустрічається в документах класу } c,$$

V — список всіх унікальних слів.

Алгоритми нечіткого пошуку

12

- ▶ Хешування за сигнатурою;
- ▶ ВК-дерева;
- ▶ Метод N-грам.



Практичне порівняння ВК-дерева та метода N-грам.

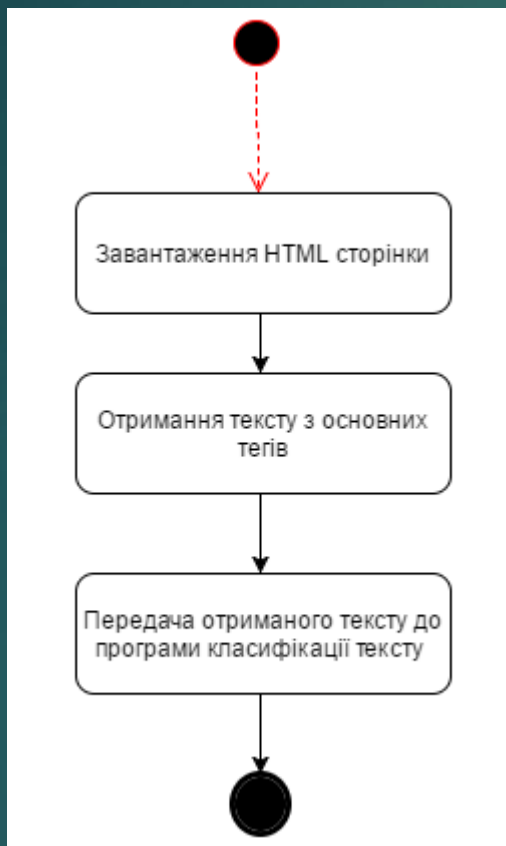
13

Пошук 100 слів з помилкою 2

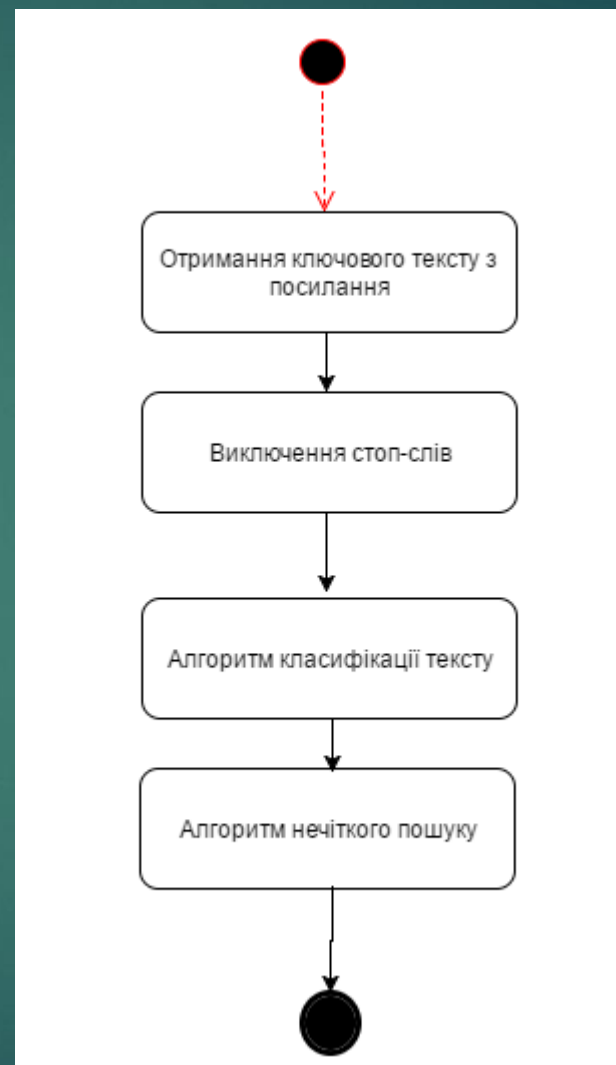
	ВК-дерева	метод N-грам
Час пошуку	60 мс	39 мс
Результат	63%	63%

Архітектура

14



Архітектура парсеру



Процес класифікації тексту

Порівняння

15

Аналіз 50 документів

	ABBYY FineReader Engine	Реалізований класифікатор тексту
Час роботи	10 с	13 с
Результат	85%	86%

Parser

Please input url for identification topics

ANALYZE

Topics:

Fly Fishing, Sports, Hobbies & Interests, Birdwatching,

Main keywords and sentences:

Berkley® Fishing, Trilene®, Fireline®, Powerbait®, Havoc®, Gulp!®, Gulp! Alive!® Berkley® Fishing,
Trilene®, Fireline®, Powerbait®, Havoc®, Gulp!®, Gulp! Alive!®, Vanish

Intelligent system for building contextual advertising

PARSER

ANALYZE MULTIPLE LINKS

RESULTS

Multiple links

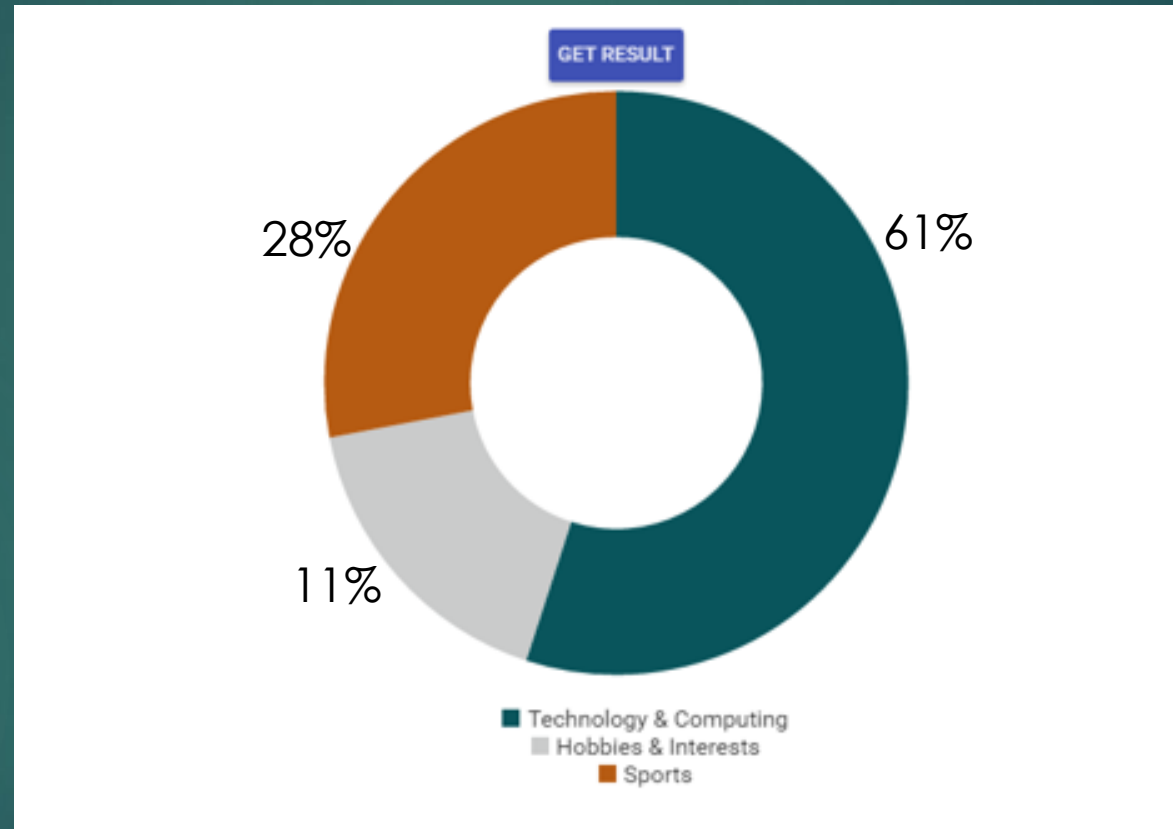
☒ Create new user

☐ Select an existing user

NewUserNameTest

Choose File urls.json

ANALYZE



Результати, якість роботи

19

Швидкість обробки запитів серверною частиною

Кількість посилань	Час обробки
5	3 с
10	5 с
15	6.74 с
20	8 с

Наприклад сервіс «Youtube» витрачає на передачу даних, 846 мс для окремо взятої сторінки

Застосування

20

- ▶ Соціальні дослідження
- ▶ Визначення впливу глобальних змін/подій на вподобання.
- ▶ Визначення тенденцій ринку

- ▶ Наукова новизна роботи: запропоновано для класифікації тексту використовувати метод байєсівського класифікатора та метод нечіткого пошуку
- ▶ Практична значимість: Реалізовано інтелектуальну систему побудови контекстної реклами, якість роботи модулю класифікації тексту становить 86%

Дякую за увагу!