МОДУЛЬ ЗБОРУ ТА АГРЕГАЦІЇ НОВИН

Автор: студент 4го курсу

Група КА-35

Дума Віталій

Науковий керівник: к.т.н. доц. Дідковська Марина Віталіївна

МЕТА, ОБ’ЄКТ ТА ПРЕДМЕТ ДОСЛІДЖЕННЯ

Мета дослідження:

- Дослідити методи оцінювання тексту за подібністю;
- Розробити програмний продукт з автоматизованою агрегацією новин та застосувати його до конретної задачі.

Об’єкт дослідження – задача агрегації тексту новин.

Предмет дослідження – аналіз методів комп’ютерної обробки природньої мови.

АКТУАЛЬНІСТЬ ЗАДАЧІ

Актуальність даної задачі полягає у тому, що кількість інтернет-ЗМІ досить велика, а для того, щоб швидко отримати всю нову інформацію, необхідно існування одного ресурсу, який би міг збирати новини та агрегувати їх за змістом.

ПОСТАНОВКА ЗАДАЧІ

- Проаналізувати існуючі алгоритми обробки природньої мови, які дозволяють оцінювати подібність тексту;
- Проаналізувати можливості подолання недоліків відомих алгоритмів в контексті конкретної предметної області;
- Розробити програмний продукт – модуль збору та агрегації новин який дозволяє формувати якісно агреговані групи новин, на основі інформації зібраної з різних джерел, та показувати результати роботи у вигляді веб-додатку.

ВХІДНІ ДАНІ

Вхідними даними даної роботи є набір новин, кожна з яких містить заголовок, текст (зміст), назву веб-ресурсу, дату публікації та посилання.

В рамках системи збору та аналізу новин було реалізовано модуль веб-скрапінгу, який збирав вхідні дані з веб-сайтів: www.5.ua, www.pravda.com.ua, tsn.ua, ukr.segodnya.ua, tyzhden.ua

Нормалізація вхідних даних

Даний процес виконується за алгоритмом стемера Портера.

"Європарламент позбавив депутатської недоторканності. Про це повідомляє Так, французьке правосуддя вимагало зняти з Ле Пен недоторканні у справі про наклепна мера Ніцци Крістіана Естрозі, якого 2015 року вона звинуватила у фінансуванні ісламістських організацій. Зазначимо, самої Ле Пен під час ухвалення рішення не було, оскільки вона наразі бере участь у передвиборчій кампанії до парламентських виборів у Франції, на яких за результатами першого туру Також окремо французьке правосуддя розглядає справу щодо позбавлення ЛеПен недоторканності у справі про фіктивне працевлаштування помічників Національного фронту"

"європарламент позбав депутатськ недоторканност про повідомля так французьк правосудд вимагал знят ле пен недоторканн справ наклеп мер ніцц крістіан естроз як рок звинуватил фінансуванн ісламістськ організац зазначим сам ле пен час ухваленн рішенн був оскільк нараз бер участ передвиборч кампані парламентськ вибор франці як результат перш тур також окрем французьк правосудд розгляд справ щод позбавленн ле пен недоторканност справ фіктивн працевлаштуванн помічник національн фронт"

Метод W-shingles з алгоритмом MinHash

- Ділить текст на послідовності шинглів:

```
"shingles" : [
    "час ухваленн рішенн",
    "працевлаштуванн помічник національн",
    "крістіан естроз як",
    "щод позбавленн ле",
    "правосудд вимагал знят",
    "окрем французьк правосудд",
```

- MinHash хешує усі шингли, для зменшення обчислювальної складності;
- Коефіцієнт Джакарда для оцінювання подібності множин

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Латентно-семантичний аналіз + TF-IDF

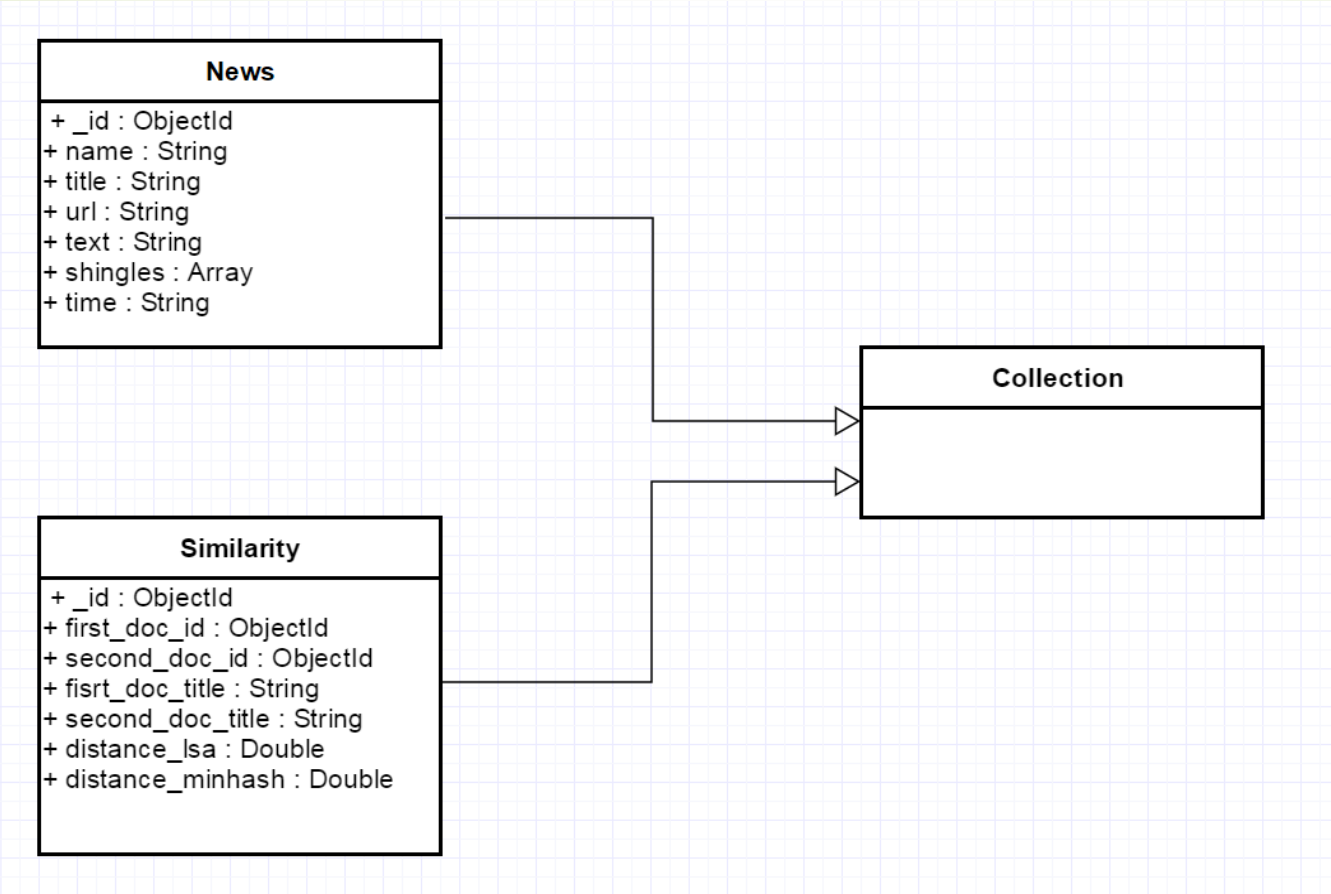
- Дозволяє виявляти приховані фактори та зв’язки між текстами;
- Прозорий та наочний алгоритм заснований на статистичній обробці текстів;
- Дозволяє ефективно розрізняти важливі фактори від шуму

Недолік:

- Висока обчислювальна складність

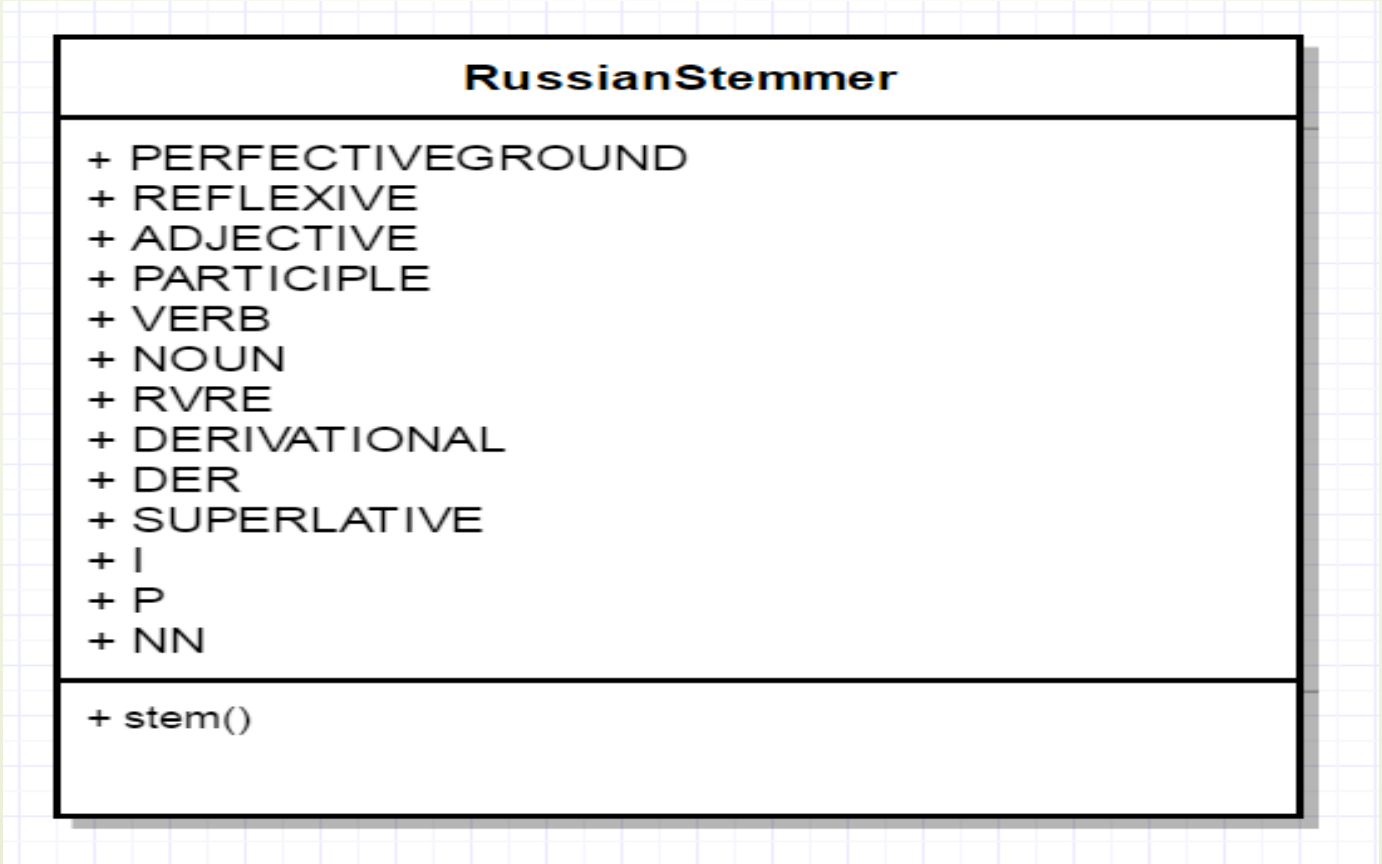
Архітектура програмного продукту

Модуль збору новин



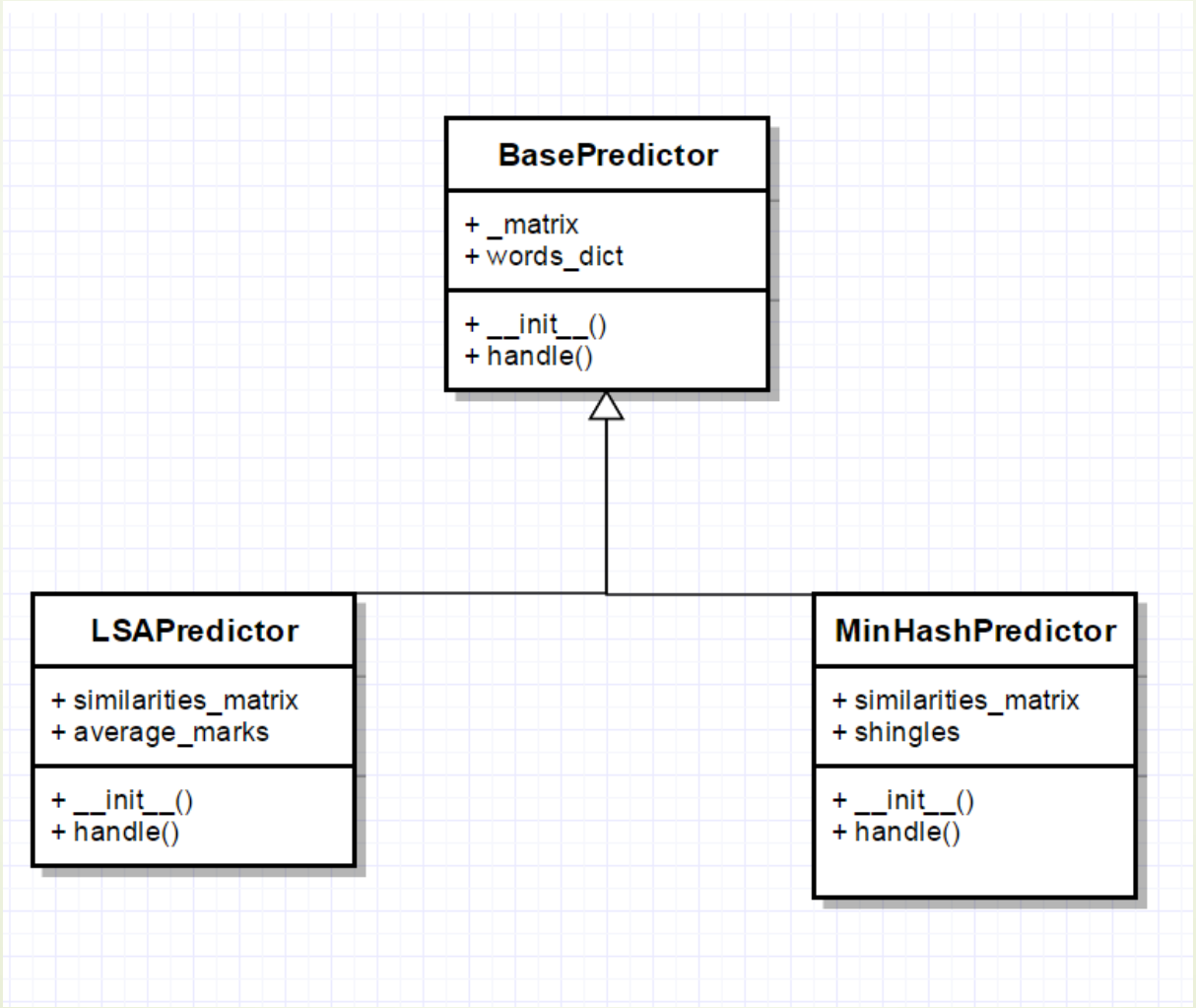
Архітектура програмного продукту

Модуль нормалізації тексту



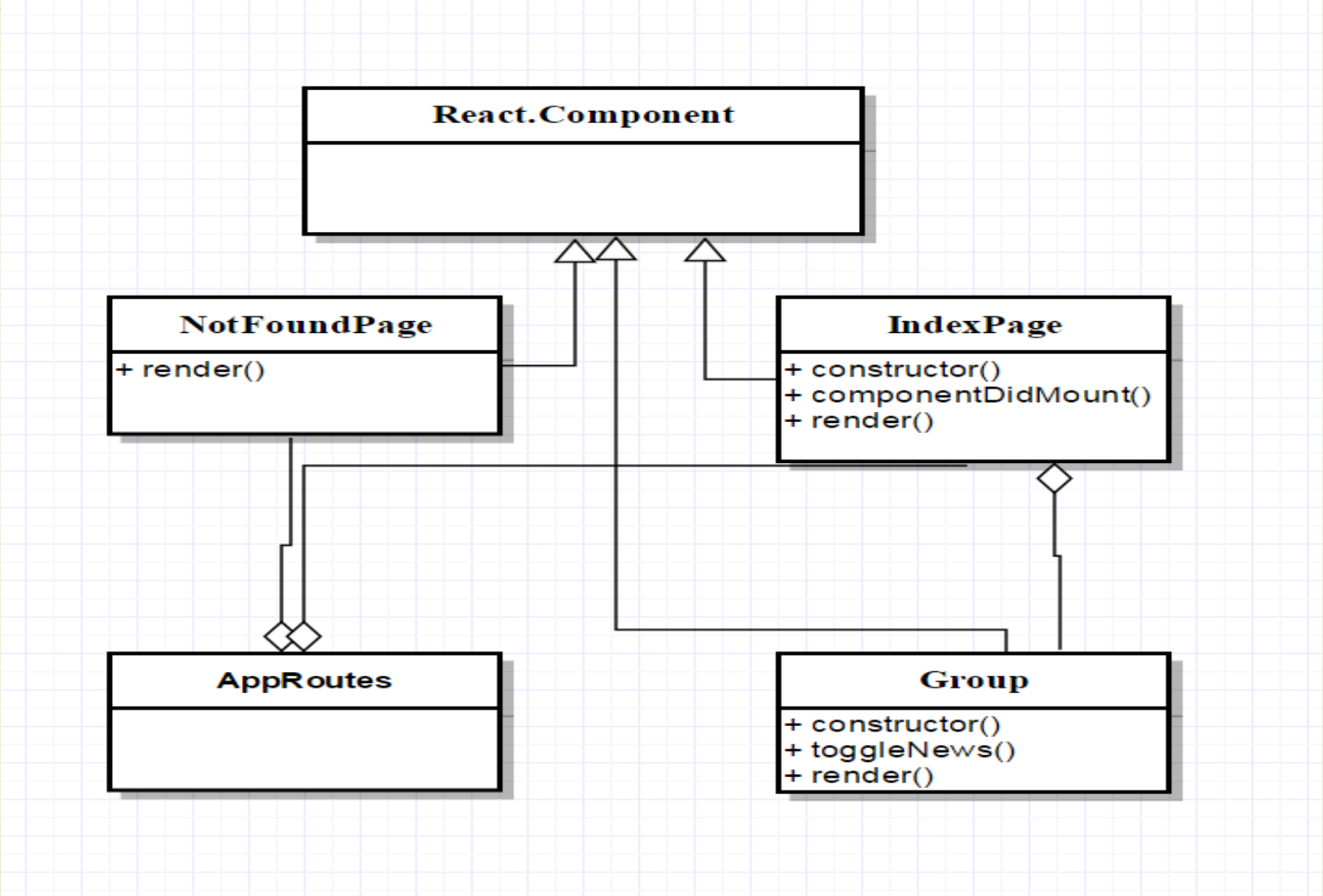
Архітектура програмного продукту

Модуль оцінювання подібності тексту



Архітектура програмного продукту

Модуль Front-End



Аналіз практичних результатів

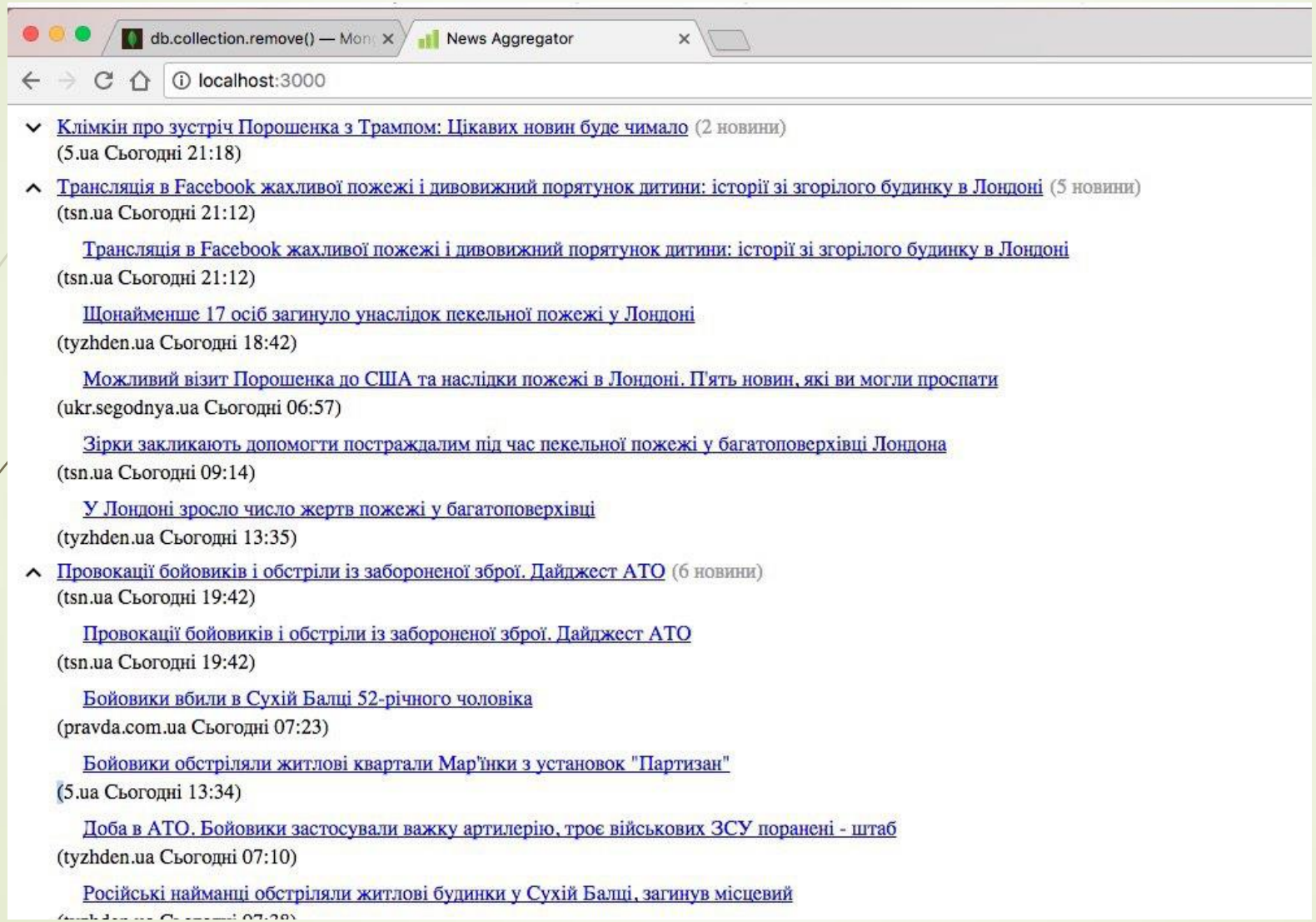
Вибірка із 100 новин

Метод	F1 measure	Precision	Recall
LSA	0.902	0.918	0.887
W-shingles	0.729	0.676	0.793
LSA + W-shingles	0.934	0.955	0.914

Вибірка із 300 новин

Метод	F1 measure	Precision	Recall
LSA	0.905	0.92	0.891
W-shingles	0.579	0.467	0.763
LSA + W-shingles	0.917	0.935	0.901

Веб-інтерфейс продукту



Використані технології

- Мова програмування Python 3.6

- Мова програмування JavaScript

- База даних MongoDB

ВИСНОВКИ

- Запропоновано **покращену** версію алгоритму W-shingles в композиції з MinHash;
- Запропоновано **модифікацію** Latent Semantic Analysis, яка використовує TF-IDF;
- Розроблено програмний продукт, що виконує збір новин, їх агрегацію та показ результатів. Система обробляє 600 новин за 12хв з точністю 90-92%.

ПОДАЛЬШІ ДОСЛІДЖЕННЯ

- Оптимізація запропонованого алгоритму для зменшення вимог до обчислювальних систем;
- Вивчення можливості побудови гібридного алгоритму, який здатний комбінувати розглянуті в роботі алгоритми, з метою покращення якості оцінювання подібності;
- Використання нейронних мереж для класифікації новин за тематикою та регіоном.

Дякую за увагу!