

Рекомендаційна система сайту новин

Студент групи ТВ-61м
Дрозд Дмитро
Степанович
Керівник: Стативка
Юрій Іванович.

BBC Nigeria

RELATED NEWS



22-year-old herdsman
arrested for allegedly killing

2019: We did not create any new
polling unit - INEC dismisses claim

Everybody knows I live a simple life
since I was a teacher - Ex-governor
denies knowledge of EFCC invitation

Delay in release of third list of looters,
prosecution of suspects worrisome -
Adeniran

Killer pest ravages tomato farms in
Kano, invites shortage of produce,
increase in price

Attahiru Jega bags another crucial
appointment in Kano

Here's the Chinese treatment a 47-
year-old civil servant used to cure all
his eye problems and restore his
eyesight

Актуальність

- Немає універсальних методів рекомендацій, кожне рішення унікальне і потребує адаптації під конкретний продукт.
- В наявних рішеннях від Google та інших виробників немає підтримки нігерійських діалектів (Hausa, Igbo, Yoruba, Urhobo, Ibibio, Edo, Fulfulde and Kanuri)
- Один запит до системи рекомендацій Google коштує 0.0005 дол, при кількості переглядів 16 мільйонів, сума становитиме 8000 доларів в місяць.

Мета

- **Метою** роботи є розробка ефективного програмного рішення для рекомендаційної системи статей на сайті новин.
- **Об'єктом** дослідження є рекомендаційна система сайту новин, яка надає персональні рекомендації користувачеві базуючись на його читацькій історії переглядів новин.
- **Предметом** дослідження є програмні рішення побудови рекомендаційних систем

Завдання диплому

- Провести огляд існуючих програмних рішень рекомендаційних систем;
- Проаналізувати існуючі методи та алгоритми побудови рекомендаційних систем;
- Розробити алгоритм рекомендацій, який би мав змогу не залежати від мови статті;
- Провести випробування рекомендаційної системи;

Наукова новизна одержаних результатів

Удосконалення алгоритму надання рекомендацій за рахунок використання:

- Історії переглядів користувача
- Кластеризації текстів новин

Як результат:

- Надання персональних рекомендацій користувачу
- Більш точне визначення тематики статті

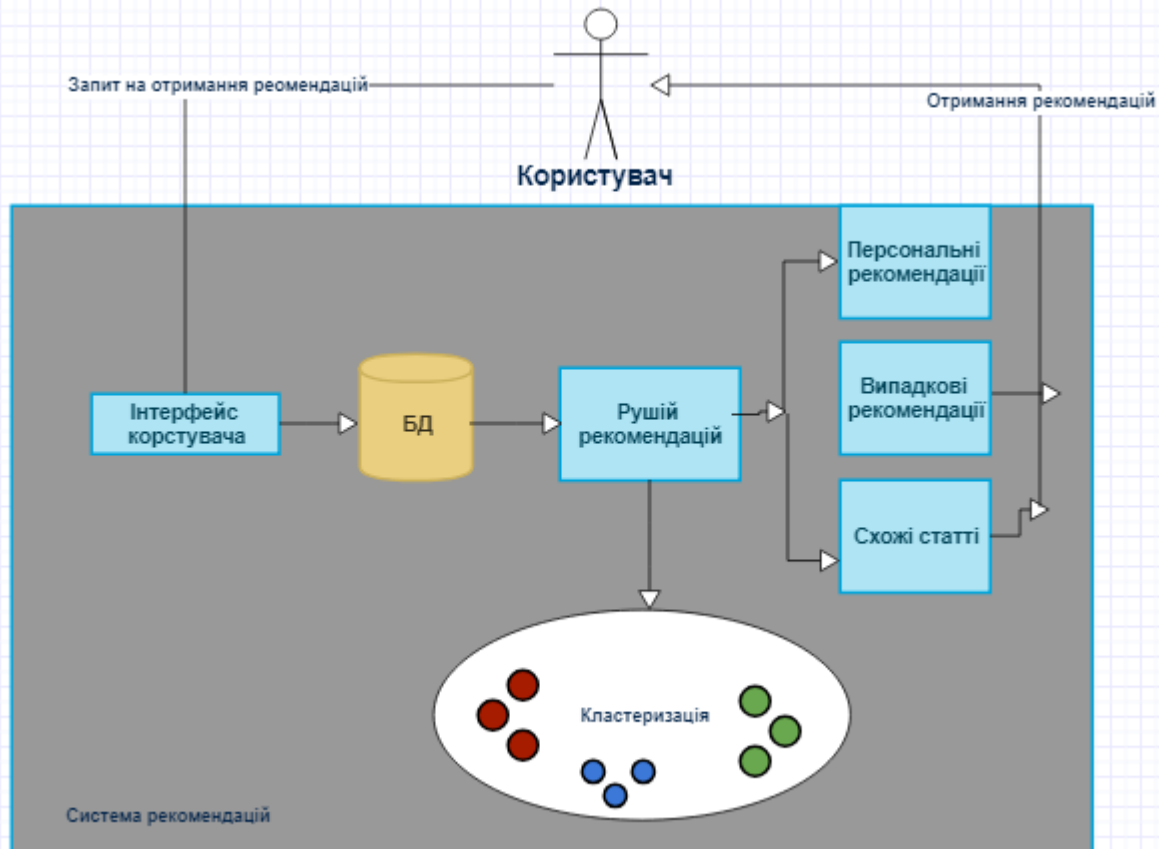
Article Data

- Text
- Number of View
- Category
- Title
- User viewed

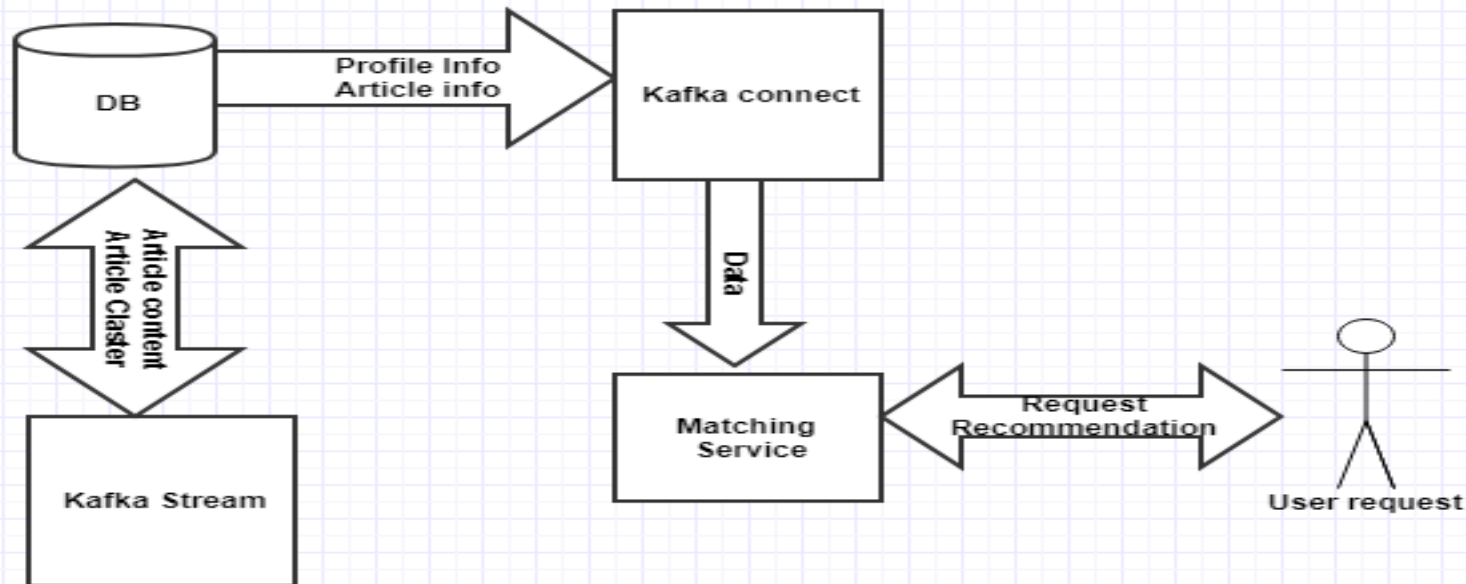
User Data

- E-mail
- Viewed Article
- Cluster

Схема роботи рекомендаційної



Програмні компоненти системи



Персональні рекомендації

1. По id користувача отримуємо історію переглядів;
2. Визначаємо топ-3 найбільш популярних категорій;
3. Виключаємо статті, які уже прочитав користувач;
4. Рекомендуємо статті з категорій, якими цікавиться користувач у відсотковому співвідношенні між категоріями;
5. Додаємо декілька випадкових статей з категорій, якими не цікавиться користувач.

Алгоритм Кластеризації

1. Отримуємо текст статті;
2. Видаляємо стоп-столова з тексту;
3. Будуємо матрицю відповідності Tf-idf;
4. Будуємо вектори по матриці відповідності для статей;
5. За допомогою алгоритму К-середніх визначаємо кластер до якого відноситься вектор.

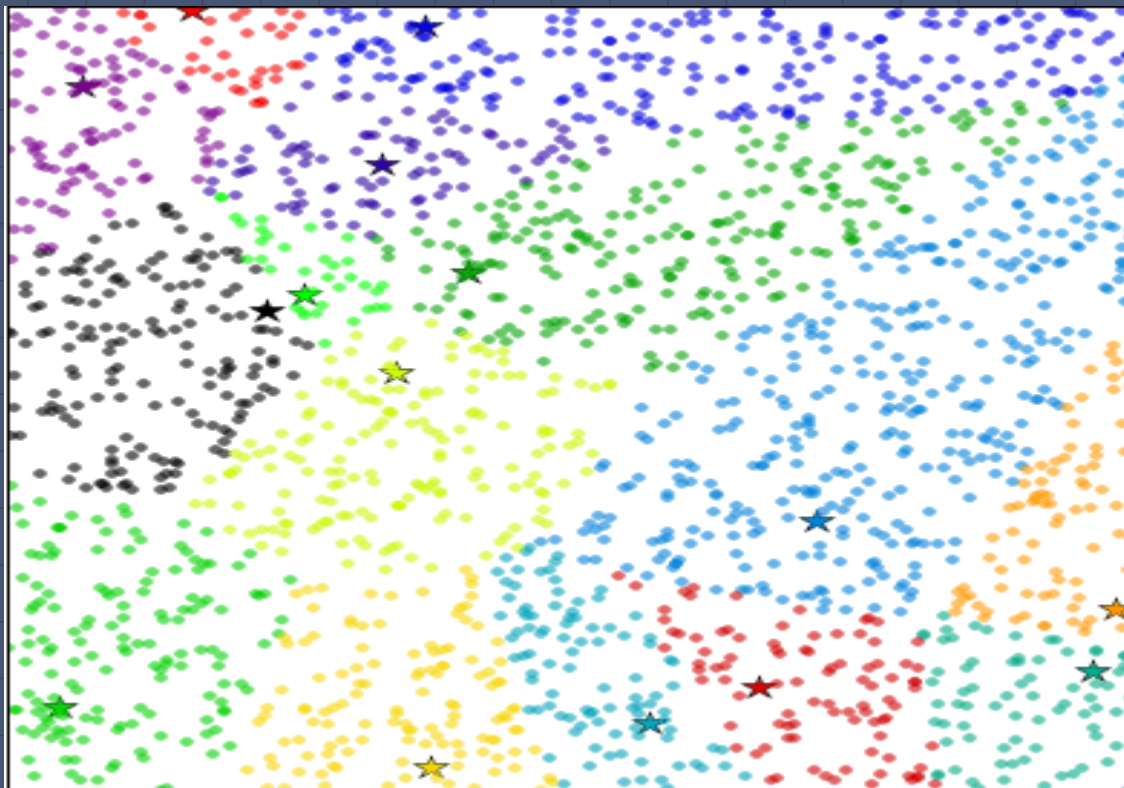
Tf-Idf

$$\mathbf{tf}(t, d) = \frac{f_d(t)}{\max_{w \in d} f_d(w)}$$

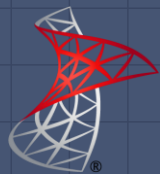
$$\mathbf{idf}(t, D) = \ln \left(\frac{|D|}{|\{d \in D : t \in d\}|} \right)$$

$$\mathbf{tfidf}(t, d, D) = \mathbf{tf}(t, d) \cdot \mathbf{idf}(t, D)$$

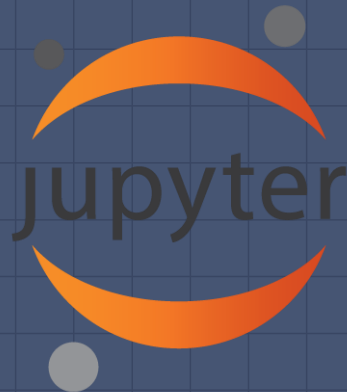
Алгоритм кластеризація (к-



Технології



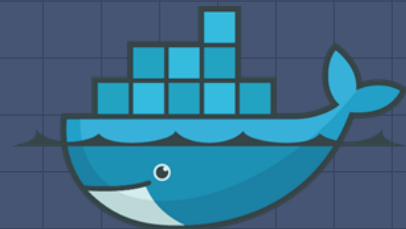
Microsoft®
SQL Server®



kafka



+ a b | e a u®
S O F T W A R E



Результат (Статистика)

	Випадкові рекомендації	Наша система	Різниця
Кількість переглядів	12 мільйонів	14 мільйонів	~17%
Середня кількість переглядів на користувача	2,3	3,1	~35%
Кількість переглядів з блоків рекомендацій	3 мільйони	4 мільйони	~33%

Результат (Сигнатура кластерів)

Top terms per cluster:

Cluster 0: ya ta na najeriya sun ba su

Cluster 1: news latest naij com read headlines page

Cluster 2: apc nigerians election fuel scarcity national congress

Cluster 3: pdp party apc jam iyyar democratic peoples

Cluster 4: president jonathan muhammadu obasanjo former goodluck senate

Cluster 5: governor fayose wike former rivers rufai ekiti

Cluster 6: follow comment twitter facebook share click us

Cluster 7: old year woman girl boy man photos

Cluster 8: man fetur wife photos young girlfriend lady

Cluster 9: manchester chelsea united arsenal league madrid city

Cluster 10: photos police government lagos actress new nollywood

Cluster 11: haram boko army borno terrorists insurgents attack

Результат (Виокремлені Кластери статей)

```
In [85]: # load cluster names
with open('cluster_names.json', 'r') as f:
    cluster_names = f.readlines()
    cluster_names = json.loads(cluster_names[0])
print(cluster_names)
```

```
{'0': 'Government', '6': 'Local news', '2': 'Short news', '1': 'Paganism (religion)', '8': 'Agro', '5': 'Criminal news', '4': 'President', '10': 'Wedding', '7': 'Football', '9': 'Politics', '3': 'Terrorism', '11': 'Sepparatism'}
```

Define cluster for article

```
In [89]: article = my_vectorizer.transform([parsed_article['title'] + parsed_article['desc']])
predicted_cluster = my_model.predict(article)[0]
cluster_names[str(predicted_cluster)]
```

```
Out[89]: 'Government'
```

Висновки

- Проаналізовано існуючі рекомендаційні системи, методи та алгоритми визначення тематики статей, а також алгоритми кластеризації статей
- Удосконалено алгоритм надання персональних рекомендацій користувачу за рахунок, використання алгоритмів кластеризації та врахування історії переглядів користувача
- Розроблено програмне рішення для реалізації рекомендаційної системи
- Результати випробувано та впроваджено в компанії “СОФТ Констракт, Київ”