

Розробка рекомендаційної навчальної системи на основі онтологій

...

Виконав студент 3 р.н. ІПЗ

Ніверовський М. М.

Науковий керівник

К-т фіз. - мат. наук Жежерун О.П.

Вступ

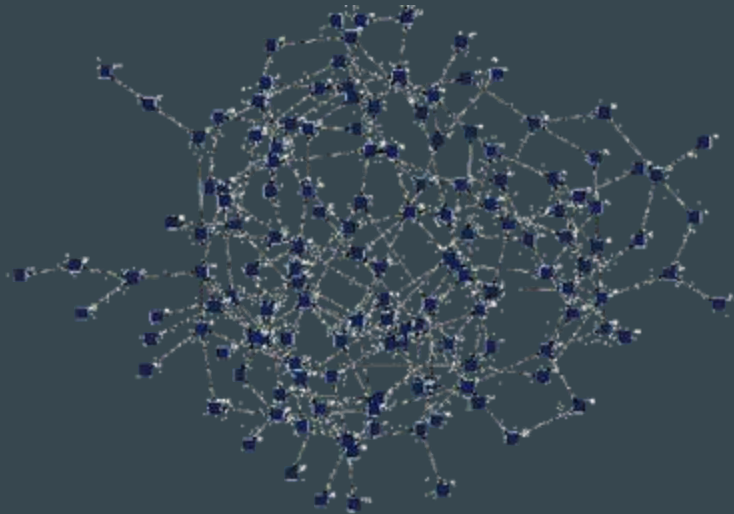
Дослідження *How much information*, яке провели у 2009 році показало, що кількість інформації яку споживає людина з 1986 року зросла у 5 разів.

Кожен день велика кількість людей читає велику кількість статей, новин, документів, книг, тощо. Для того щоб якось шукати це в інтернеті, класифікувати, оцінювати, використовується засоби природної мови.

Одним із найпопулярніших засобів розпізнавання людської мови є ЛСА – латентно-семантичний аналіз.

Постановка задачі

Дослідити техніку природного розпізнавання мови Латентно-семантичний аналіз та реалізувати один із засобів на мові програмування Python.



Застосування ЛСА

- Близькі по значенню
- Порівняння у матриці
- Порівняння один-до-багатьох
- Багатомовний пошук
- Оцінка ессе

Принцип роботи

1. Підготувати дані
2. Побудувати матрицю вживаності
3. Розкласти матрицю
4. Обробити результат

Розклад матриці

Найбільш поширений варіант розкладу матриці - це SVD (сингулярний розклад).

Посилаючись до теореми о сингулярному розкладанні:

$$A = U\Sigma V^t$$

де Σ – це діагональна матриця $m \times n$ з додатних елементів, які є сингулярними числами;

U – це матриця $m \times m$, яка складається з лівих сингулярних векторів;

V – це матриця $n \times n$, яка складається з правих сингулярних векторів;

V^t – це транспонована матриця до V

Засоби розробки



NumPy

NLTK 3.5



TensorFlow



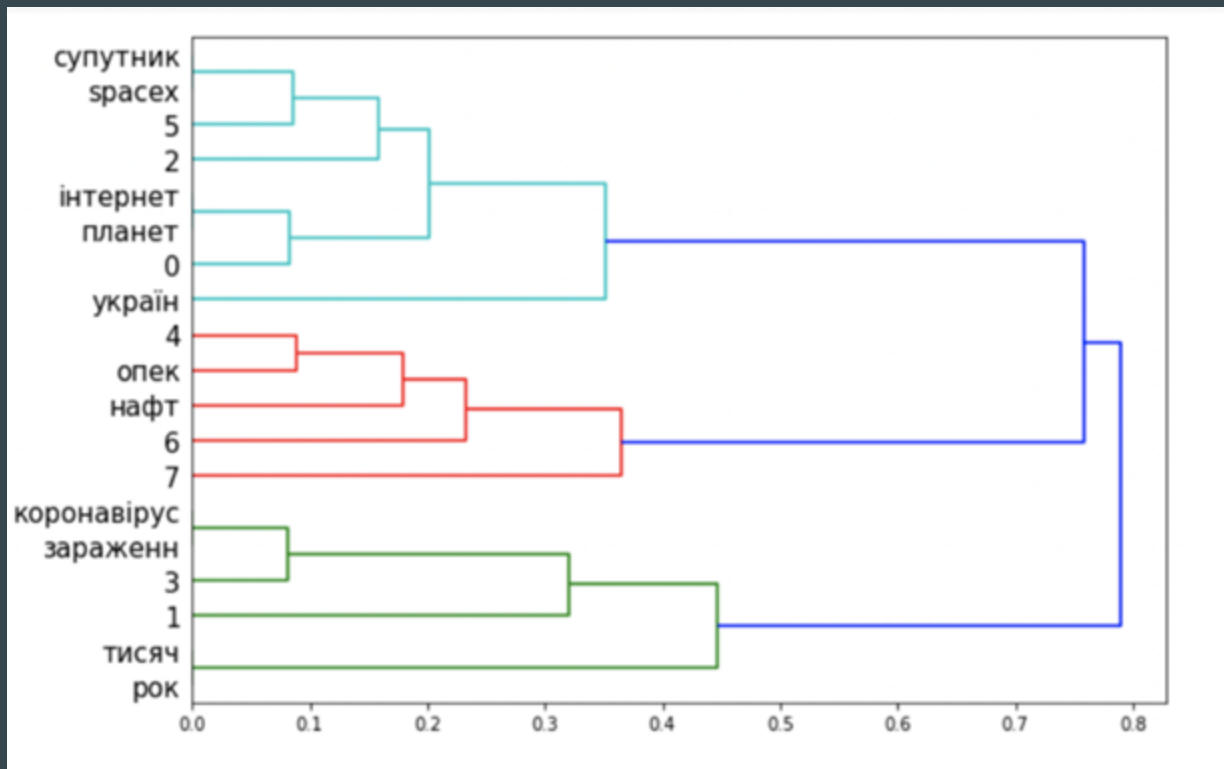
SciPy



Матриця вживаності

```
спасех [1. 0. 1. 0. 0. 1. 0. 0.]  
зараженн [0. 1. 0. 1. 0. 0. 0. 0.]  
коронавірус [0. 1. 0. 1. 0. 0. 0. 0.]  
нафт [0. 0. 0. 0. 1. 0. 1. 1.]  
опек [0. 0. 0. 0. 1. 0. 0. 1.]  
планет [0. 0. 1. 0. 0. 1. 0. 0.]  
рок [0. 1. 0. 0. 0. 0. 0. 1.]  
супутник [1. 0. 1. 0. 0. 1. 0. 0.]  
тисяч [0. 1. 0. 0. 0. 0. 0. 1.]  
україн [0. 0. 1. 0. 0. 0. 1. 0.]  
інтернет [0. 0. 1. 0. 0. 1. 0. 0.]
```

Результат класифікації статей



Переваги ЛСА

- Метод вважається одним з кращих для пошуку глибинних (латентних) залежностей серед великої кількості документів.
- ЛСА можна застосовувати як з навчанням так і без нього.

Недоліки ЛСА

- Цей метод працює значно повільніше при збільшенні кількості вхідних даних. Десь N^{2k} , де $N = N_{\text{doc}} + N_{\text{term}}$ – кількість документів та термів, k – розмір простору факторів.
- Стохастична модель не співпадає з реальністю.

Висновок

В ході роботи було детально досліджено латентно-семантичний аналіз. Можна зробити такі висновки: ЛСА є дуже функціональним засобом обробки природної мови, за допомогою нього можна шукати інформацію, знаходити залежності між текстами, фільтрувати їх, класифікувати, оцінювати тощо.

У ході практичної роботи було реалізовано засіб для класифікації статей та досліджена велика кількість програмних технологій, особливо технологій, які стосуються машинному навчанню. Були відібрані тільки кращі продукти, які користуються попитом та вже не один раз показали себе з кращого боку.