

ВИЗНАЧЕННЯ КЛЮЧОВИХ СЛІВ ТЕКСТУ ПРИРОДНОЇ МОВИ ЗА ДОПОМОГОЮ ПАКЕТУ NLTK.

Траченко С.С.

Науковий керівник – д.т.н., проф. Бісікало О. В.

Через швидкий розвиток інформаційних технологій та геометричні обсяги збільшення текстової інформації в мережі Інтернет набула значної актуальності проблема оцінки сенсу тексту. Для її розв'язання використовують різні методи визначення ключових слів у тексті, оскільки саме за ключовими словами оцінюється загальний зміст тексту та його позиція у списку результатів запиту до пошукових машин.

Для реалізації задачі дослідження пропонується застосування можливостей лінгвістичного пакету NLTK на мові Python. Компонентна технологія на основі NLTK з залученням інтерфейсу для WordNet передбачає використання класу SynSet (описує набір синонімів) та наступних функцій:

- sent_tokenize – виокремлення речень із тексту;
- word_tokenize – виокремлення ключів з тексту;
- `wsd.lesk` (`wsd` – клас, але в ньому лише 1 метод) – виокремлення SynSet із тексту, використовуючи алгоритм LESK;
- `wordnet.path_similarity` (аналогічно `lesk`) – знаходження того, наскільки два SynSet близькі між собою за значенням.

Загальна задача знаходження ключових слів зводиться до такого узагальненого алгоритму:

1) Отримання значень слів (SynSet) із кожного речення.

Виокремлення окремих речень з тексту та ключів із речень – тривіальна задача при використанні NLTK пакету. Але постає питання надлишковості артиклів та інших стоп-слів англійської мови, що перешкоджає адекватному визначенню ключових зв'язків орієнтованого графу. Для їх усунення використовується готовий список стоп-слів пакету NLTK.

2) Визначення ключовими таких слів, які за сумою зв'язків з іншими SynSet'ами мають найбільше значення.

Організація циклу між усіма виокремленими SynSet'ами та визначення ваги зв'язку між парами усіх SynSet'ів. Під час даного циклу зв'язки, що описують зв'язок «до» певного SynSet'у сумуються у відведеному для цього словнику, де ключем є SynSet, а значення – сума вхідних зв'язків.

Результуючий словник, що відсортований за сумами вхідних зв'язків, зшукає ключові слова тексту краще, ніж відповідний словник, відсортований за сумами вихідних чи загальних (вхідних та вихідних) зв'язків. Для реалізації запропонованого методу було розроблено власний скрипт, який отримує на вході текст, а на виході надає список SynSet, що відсортований за сумарним значенням вхідних зв'язків та візуалізує його у вигляді графу.