УДК 004.91

ПОБУДОВА ОНТОЛОГІЇ ТЕКСТУ ПРИРОДНОЇ МОВИ ЗА ДОПОМОГОЮ ПАКЕТУ NLTK

Сергій Траченко – студент групи 1СІ-11б, Вінницький національний технічний університет (ВНТУ), Україна
Науковий керівник **– Олег Бісікало,** д-р техн. наук, професор, в.о. декана ФКСА, ВНТУ, Україна

Через швидкий розвиток інформаційних технологій та геометричні обсяги збільшення текстової інформації в мережі Інтернет набула значної актуальності проблема визначення сенсу тексту. Одним із перспективних інструментів для вирішення задачі вилучення знань з тексту вважають онтологокеровані інформаційні системи.

Для реалізації задачі дослідження пропонується застосування можливостей лінгвістичного пакету NLTK [1] на мові Python. NLTK є провідною платформою для побудови програм мовою Python з метою обробки даних природної мови. Пакет забезпечує зручні у використанні інтерфейси для більш ніж 50 корпусів і лексичних ресурсів, таких як WordNet, поряд з набором бібліотек обробки тексту для класифікації, токенізації, стемінізації та інших інструментів для розв’язання задач комп'ютерної лінгвістики.

Компонентна технологія передбачає використання таких класів:

SynSet – клас, що 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 [2];

wordnet.path_similarity (аналогічно lesk) – знаходження того, наскільки два SynSet близькі між собою за значенням.

Загальна задача знаходження онтології тексту 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 скінченого графу. Для їх усунення використовується готовий список стоп-слів пакету NLTK.

2) Визначення ключовими словами онтології таких, які за сумою вхідних 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 опи-

сують зв’язок «до» певного SynSet’у сумуються у відведеному для цього словнику, де ключем є SynSet, а значенням – сума вхідних зв’язків.

Отже, результуючий словник, що відсортований по сумах вхідних зв’язків, демонструє онтологічний набір SynSet’ів заданого тексту.

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.

Література

1. Steven Bird, Ewan Klein, and Edward Loper. – Gravenstein Highway North (Sebastopol): O’Reilly Media, Inc. – 2010.

2. M. Lesk**.** Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone // SIGDOC '86 Proceedings of the 5th annual international conference on Systems documentation. **–** University of Toronto (Canada, Toronto). **–** 1986. **–** 24-26**.**