УДК 004.912

О. В. Бісікало, А. І. Лісовенко, Я. Е. Копецький

Розробка програмного забезпечення для визначення чисельних характеристик словосполучень текстового файлу

Вінницький національний технічний університет

Анотація

Запропоновано програмне забезпечення для визначення чисельних характеристик словосполучень текстового файлу на основі вільно доступних інструментів парсингу. Підхід використовує POS тегування, що дозволяє зменшити час обробки тестової інформації та форматувати отримані результати у вигляді таблиці MS Excel.

Ключові слова: текстовий файл, словосполучення, чисельні характеристики, парсинг, токенізація, програмне забезпечення.

Abstract

The software for determining numerical characteristics of phrases of a text file is offered on the basis of freely available parsing tools. The approach uses POS tags, which reduces the processing time of the test information and format the results in the form of a MS Excel spreadsheet.

Keywords: *text file, word combination, numerical characteristics, parsing, tokenization, software*.

.

Вступ

З розвитком інформаційних технологій зросла потреба у системах оброблення великих масивів даних, зокрема текстових, а також у зведенні результатів парсингу [1] таких даних до визначеної моделі для подальшої обробки. Тому навіть незначне скорочення часу обробки, у т.ч. внаслідок покращення відповідного методу оброблення даних дозволить збільшити обсяг отриманої бази знань предметної області.

Мета роботи полягає у розробленні методу визначення та узагальнення чисельних характеристик словосполучень текстових файлів.

Результати дослідження

Будемо вважати, що парсинг [2] – це синтаксичний аналіз певним чином форматованої інформації. Для текстової інформації важливим є POS тегування – етап автоматичної обробки тексту. Завдання останнього полягає у визначенні частини мови і граматичних характеристик слів в тексті з приписуванням їм відповідних міток (тегів). Тегування є одним з перших етапів комп'ютерного аналізу тексту. Алгоритми POS поділяться на дві групи: засновані на правилах і ймовірні. Список найбільш розповсюджених POS тегів наведено на рис.1.

Використовуючи фреймворк *Stanford Parser*, розроблене програмне забезпечення розділяє речення на слова, згодом визначає лексемні теги [3] кожного слова, зрештою рахує їх кількість на основі використання всього масиву отриманої інформації. На основі формалізованих понять [4] створюється бінарне дерево зв’язків, а далі, за допомогою ітераційного методу, проводиться обхід вершин дерева з подальшою градацією даних на слова та/або символи пунктуації.

Запропонований підхід дозволяє використовувати отримані результати у вигляді чисельних

характеристик словосполучень текстових файлів програмно, у т. ч. з можливістю експорту у таблиці MS Excel з метою подальшого застосування методів машинного навчання

Table 2
The Penn Treebank POS tagset.

1. CC	Coordinating conjunction	25. TO	*to*
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential *there*	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. $	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP$	Possessive pronoun	43.)	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. '	Left open single quote
22. RBS	Adverb, superlative	46. "	Left open double quote
23. RP	Particle	47. '	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. "	Right close double quote

Рис. 1. Список найбільш розповсюджених POS тегів

Висновки

У роботі проаналізовано поняття та структуру парсингу, зокрема запропоновано застосувати модель парсингу з використання POS тегів з метою визначення та узагальнення чисельних характеристик словосполучень текстових файлів.

Отже, механізм парсингу стає вагомою частиною обробки неструктурованої інформації у сучасному світі інформаційних технологій. Саме тому розглянуті методи та алгоритми застосування парсингу як потужного інструменту формального аналізу природно-мовних текстів знайдуть своє місце у майбутньому.

Список використаної літератури

1. Parsing Basics [Електронний ресурс]. — 2018. — Режим доступу до ресурсу: https://www.ipipe.ru/info/parsing.html. — Electronics and Computer Science of UK, 2002. — 221 p.

2. Psycholinguistics [Електронний ресурс]. — 2018. — Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Parsing. — A Probabilistic Model of Lexical and Syntactic Access and Disambiguation, 2004. — 12 p.

3. Computer language [Електронний ресурс]. — 2018. — Режим доступу до ресурсу: https://en.wikipedia.org/wiki/ Computer_language. — 2007. — 33 p.

4. Бісікало О.В. Формалізація понять мовного образу та образного сенсу природно-мовних конструкцій / О.В. Бісікало // Математичні машини і системи. — 2012. — № 2. — С. 70–73.

Бісікало Олег Володимирович — д-р техн. наук, професор, декан факультету КСА, Вінницький національний технічний університет, м. Вінниця, e-mail: obisikalo@gmail.com

Лісовенко Анна Ігорівна — к.т.н., асистент кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця

Копецький Ярослав Едуардович — студент групи І-15б, факультет комп'ютерних систем та автоматики, Вінницький національний технічний університет, Вінниця

Bisikalo Oleh V. — Dr.Sc. (Eng.), Professor, Dean of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia, email: obisikalo@gmail.com

Lisovenko Anna I. - Candidate of Technical Sciences, Assistant Professor, Department of Automation and Intelligent Information Technologies, Vinnytsia National Technical University, Vinnytsia.

Kopetsky Yaroslav E. — student, Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia.