Огляд засобів латентно-семантичного аналізу

Дмитро Геннадійович Медведєв[*], Владислав Миколайович Пірогов[≠]
Кафедра інформатики та прикладної математики, Криворізький державний педагогічний університет, пр. Гагаріна, 54, м. Кривий Ріг, 50086, Україна
dimakrrog@kdpu.edu.ua[*], pirogov1818@gmail.com[≠]

Анотація. *Метою дослідження* є огляд бібліотек для латентно-семантичного аналізу. *Задачами дослідження* є аналіз вільно розповсюджуваних програмних продуктів, що реалізовують латентно-семантичний аналіз. *Об'єктом дослідження* є процес опрацювання даних. *Предметом дослідження* є опрацювання даних засобами Python. *Результатом дослідження* є вибір доцільних бібліотек Python, що реалізують алгоритми латентно-семантичного аналізу.

Ключові слова: латентно-семантичний аналіз; Python.

D. H. Medvediev[*], V. M. Pirohov[≠]. Review of tools for latent and semantic analysis

Abstract. The *objective* of the study is the review of libraries for latent and semantic analysis. *The tasks* of the study are the analysis of free software for latent and semantic analysis. The *object* is data processing. The *subject* is data processing in Python. The *research result* is choice of appropriate libraries which implements latent and semantic analysis in Python.

Keywords: latent and semantic analysis; Python.

Affiliation: Department of Computer Science and Applied Mathematics, Kryvyi Rih State Pedagogical University, 54, Gagarin avenue, Kryvyi Rih, 50086, Ukraine.
E-mail: dimakrrog@kdpu.edu.ua[*], pirogov1818@gmail.com[≠].

Алгоритм латентно-семантичного аналізу (індексування) – це метод опрацювання інформації, поданої природною мовою, що аналізує взаємозв'язок між документами та використаними в них термінами і зіставляє деякі чинники (тематики) всім документам і термінам. Принцип роботи полягає в використанні алгоритмів факторного аналізу.

У процесі аналізу проектів, що реалізують алгоритм латентно-семантичного аналізу, було виявлено значну кількість вільно розповсюджуваних програмних засобів. Серед яких можна виділити SenseClusters, S-Space, Semantic Vectors.

SenseClusters – це програмний пакет, розроблений на мові Perl. Основні функцій SenseClusters є:

– кластеризація схожих контекстів;
– розділення неоднозначності слів;
– класифікація документів різного роду;
– класифікація лексики;
– можливість роботи с біграмами.

Крім алгоритмів, що реалізовують латентно-семантичний аналіз, в SenseClusters є власні алгоритми, що виконують аналогічні функцій. Дана програма працює повільніше ніж її аналоги.

S-Space – це програмний пакет, розроблений засобами мови Java. Він є універсальним інструментом для побудови та опрацювання векторної моделі. S-Space оптимізована для швидкодії. Також програма містить у собі реалізацію багатьох алгоритмів для різних векторних моделей та їх подальшого опрацювання. Вхідні дані можуть прийматися у вигляді єдиного текстового файлу або рядками, що суттєво спрошує внесення даних. Вихідні дані можуть надаватися у бінарному чи текстовому форматі.

Semantic Vectors – це програмний пакет, розроблений мовою Java. Головною функцією пакета є створення векторно-семантичної моделі для визначення співпадінь між поняттями та запитом пошуку в інформаційній системі. Результат може бути подано у вигляді текстового або бінарного документа.

Для створення програмного забезпечення для латентно-семантичного аналізу доцільно використовувати вільно поширювану мову програмування Python [1; 2], до складу якої входять математичні бібліотеки NumPy і SciPy, що надають можливості швидкого виконання сингулярного розкладу матриці термів-на-документи. Для практичного використання рекомендуємо наступні бібліотеки:

1) models.lsimodel – Latent Semantic Indexing (http://radimrehurek.com/gensim/models/lsimodel.html, розробники – Радім Рехурек та Петр Сойка) реалізує швидкий інкрементальний сингулярний розклад, що надає можливість опрацьовувати надвеликі обсяги даних, які не можуть бути одночасно розміщені в оперативній пам'яті;

2) SemanticPy (https://github.com/josephwilk/semanticpy, розробник – Джозеф Уілк) відрізняться з-поміж інших гарною реалізацією алгоритму Портера для виокремлення коренів слів;

3) Document (Term) Similarity using Latent Semantic Indexing (https://github.com/neomoha/python-lsi-similarity, розробник – Мохамед Сордо) є простою бібліотекою для навчальних досліджень у галузі індексування та пошуку даних;

4) Scikit-learn (http://scikit-learn.org/stable/index.html, розробник – INRIA) – потужна промислова бібліотека для машинного навчання, що

надає різні алгоритми сингулярного розкладу матриці термів-на-документи;

5) LSA_Classification (https://github.com/chrisjmccormick/LSA_Classification, розробник – Кріс Маккормік) – надбудова над Scikit-learn для розв'язання задачі класифікації текстів;

6) semantic-analysis (https://github.com/Serafim-End/semantic-analysis, розробник – Нікіта Ендальцев) – реалізує семантичний аналіз текстів російською мовою.

Список використаних джерел

1. Sarkar D. Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data / Dipanjan Sarkar. – New York : Apress, 2016. – XXI, 385 p.

2. Поліщук О. П. Програмна підтримка комп'ютерного моделювання засобами мови Python / Поліщук О. П., Семеріков С. О., Теплицький І. О., Бойко А. С. // Інноваційні технології навчання в сучасній дидактиці вищої школи : матеріали Другої всеукраїнської науково-практичної конференції 13–16 березня 2007 р. – Полтава, 2007. – С. 71-72.

References (translated and transliterated)

1. Sarkar D. Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data / Dipanjan Sarkar. – New York : Apress, 2016. – XXI, 385 p.

2. Polishchuk O. P. Prohramna pidtrymka kompiuternoho modeliuvannia zasobamy movy Python [Software support of computer modeling by Python language tools] / Polishchuk O. P., Semerikov S. O., Teplytskyi I. O., Boiko A. S. // Innovatsiini tekhnolohii navchannia v suchasnii dydaktytsi vyshchoi shkoly : materialy Druhoi vseukrainskoi naukovo-praktychnoi konferentsii 13–16 bereznia 2007 r. – Poltava, 2007. – S. 71-72. (In Ukrainian)