УДК 004.91:004.82 (043.2)

Амонс В.В., Корчемний Р.Є.
Національний авіаційний університет

ОГЛЯД БІБЛІОТЕКИ GENSIM ДЛЯ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ

Найпопулярнішою мовою програмування для роботи з обробкою природної мови (NLP) є Python. Крім інтуїтивності розуміння, підтримки інтеграції з іншими мовами та інструментами, важливим аспектом вибору Python є те, що він надає розробникам широкий спектр бібліотек NLP. Однією з таких бібліотек є Gensim.

Gensim – це безкоштовна бібліотека обробки природної мови, що першочергово призначена для тематичного моделювання.

Тематичне моделювання (ТМ) – це метод виявлення тем, яким присвячений текст, що оброблюється. Кожен документ розглядається як комбінація тем, а кожна тема як комбінація пов'язаних слів. У бібліотеці Gensim ефективно реалізовані популярні алгоритми ТМ, такі як приховане семантичне індексування (LSI) та прихований розподіл Діріхле (LDA).

В обох випадках вказується кількість тем як вхідні дані. Тематична модель, у свою чергу, надає ключові слова для кожної теми та їх відсотковий зміст у кожному документі.

Об'єкт моделі підтримує індексацію. Тобто, якщо ви передаєте документ (список слів) в модель, то вона надає 3 типи тем: теми, що належать цьому документу разом із відсотками; теми, до яких належить кожне слово у цьому документі; теми для кожного слова в цьому документі та значення phi (ймовірність того, що слово відноситься до цієї конкретної теми).

Вхідний текст може мати наступні форми: речення, що зберігається в об'єкті списку Python; один текстовий файл; декілька текстових файлів. Усі алгоритми можуть обробляти великі вхідні дані, не завантажуючи весь файл в пам’ять.

Швидкість та ефективність пам’яті є головними особливостями Gensim. BLAS (Basic Linear Algebra Subroutines) - бібліотека базових підпрограм лінійної алгебри. Gensim використовує ці низькорівневі бібліотеки за допомогою своєї залежності від NumPy – пакету Python для наукових обчислень. Незважаючи на те, що код верхнього рівня Gensim написаний виключно на Python, він фактично виконує оптимізований Fortran/C.