Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

Отчет по лабораторной работе №3
по курсу
”Проектирование Рекомендательных Систем”

Тема TF-IDF и LDA

Студент Якуба Д. В.

Группа ИУ7-33М

Оценка (баллы) _____

Преподаватели Быстрицкая А.Ю.

Москва — 2023 г.

Оглавление

ВВЕДЕНИЕ

Цель работы – изучить TF-IDF и LDA.

Для достижения поставленной цели потребуется:

- привести описание алгоритмов;
- привести описание используемых для исследования данных;
- привести зависимости скорости работы алгоритмов от заданных параметров.

1. Аналитический раздел

1.1 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) – это статистическая мера, используемая в информационном поиске и анализе текста для оценки важности слова в документе относительно всей коллекции документов. Эта мера может быть полезной и в рекомендательных системах для оценки сходства между элементам и пользователями. [1]

TF – частота слова, отношение числа вхождений некоторого слова к общему числу слов документа, так оценивается важность слова t_i в пределах отдельного документа:

$$tf(t,d) = \frac{n_t}{\sum_k n_k}, \tag{1}$$

где n_t — число вхождений слова t в документ;

$\sum_k n_k$ — общее количество слов в данном документе.

IDF – обратная частота документа, инверсия частоты, с которой некоторое слово встречается в документах коллекции.

$$IDF(t,D) = log\frac{|D|}{|\{d_i \in D | t \in d_i\}|}, \tag{2}$$

где $|D|$ — число документов в коллекции;

$|\{d_i \in D | t \in d_i\}|$ — число документов из коллекции D, в которых встречается t (когда $n_t \neq 0$).

Данная мера может быть использована в рекомендательных системах для:

- Представления контента, такого как текстовые описания товаров, фильмов или музыкальных треков; каждый элемент (например, товар) будет представлен его описанием-вектором, в котором каждое слово представлено его TF-IDF весом, что позволит понимать, какие слова играют важную роль в этом описании – выделать “тэги”;
- Определения сходства элементов и пользователе через косинусное сходство между векторами; элементы, чьи векторы более похожи на вектор пользователя, могут быть ему рекомендованы;
- Улучшения рекомендаций путем подсчета весовых коэффициентов для

слов или фраз в профилях пользователей; если пользователь часто взаимодействует с элементами, содержащими определенные ключевые слова, то можно увеличить вес для этих слов в профиле пользователя;

- Модификации; TF-IDF может быть использован вместе с другими методами рекомендации, например, с коллаборативной фильтрацией, для улучшения точности и разнообразия рекомендаций.

При этом TF-IDF имеет некоторые ограничения: он не учитывает контекст слов и не способен обрабатывать синонимы. [1]

1.2 LDA

LDA (Latent Dirichlet Allocation) – это статистическая модель, используемая в анализе текстовых данных для выявления скрытых тем в коллекции документов. Данная модель предполагает, что каждый документ в коллекции создается путем комбинирования нескольких тем, и каждая тема представляет собой распределение слов. [2]

Данная модель может быть использована в рекомендательных системах для [2]:

- Извлечения тематических профилей – путем применения LDA к текстовым данным, можно извлечь тематические профили для каждого элемента, которые представляют собой вероятностные распределения тем в каждом элементе;
- Рекомендаций на основе тем – при наличии профилей элементов и пользователя, можно измерить сходство между темами и рекомендовать элементы, которые имеют близкие тематические профили к профилю пользователя;
- Разнообразия рекомендаций – LDA может помочь в улучшении разнообразия, так как модель позволяет контролировать количество тем;
- Персонализации – модель может быть адаптирована к поведению конкретного пользователя, чтобы улучшить качество рекомендаций.

Преимущество LDA над TF-IDF заключается в том, что он учитывает более высокоуровневую структуру текстовых данных и может обнаруживать тематические зависимости между словами. Однако LDA имеет и ограничения: необходимость выбора количества тем (подбор значения данного параметра), и он может быть более сложным в реализации и настройке, чем TF-IDF.

2. Конструкторский раздел

2.1 Kaggle: Research Articles

В качестве источника данных был взят датасет, располагающийся в свободном доступе на веб-сайте Kaggle [3]. Датасет включает в себя статьи, которые описаны наименованием, аннотацией и тэгами. Предобработка включала в себя лемматизацию и удаление стоп-слов в аннотациях.

3. Технологический раздел

3.1 Средства реализации

В качестве используемого был выбран язык программирования Python [4].

Данный выбор обусловлен следующими факторами:

- Большое количество исчерпывающей документации;
- Широкий выбор доступных библиотек для разработки;
- Простота синтаксиса языка и высокая скорость разработки.

При написании программного продукта использовалась среда разработки Visual Studio Code. Данный выбор обусловлен тем, что данная среда распространяется по свободной лицензии, поставляется для конечного пользователя с открытым исходным кодом, а также имеет большое число расширений, ускоряющих разработку.

3.2 Библиотеки

При анализе и обработке датасета, а также для решения поставленных задач использовались библиотеки:

- pandas;
- numpy;
- matplotlib;
- sklearn [5].

Данные библиотеки позволили полностью покрыть спектр потребностей при выполнении работы.

4. Исследовательский раздел

4.1 Условия исследований

Исследование проводилось на персональном вычислительной машине со следующими характеристиками:

- процессор Apple M1 Pro,
- операционная система Ventura 13.5.2,
- 32 Гб оперативной памяти.

Временные затраты определялись с использованием библиотеки time.

Важно отметить, что время, затраченное на предобработку датасета (лемматизация и удаление соп-слов), в случае TF-IDF учитывается в общем времени исполнения.

4.2 Зависимость времени исполнения TF-IDF от значения параметра максимальной частоты встречаемости слова

На рисунке 4.1 представлен график зависимости времени исполнения TF-IDF от значения параметра максимальной частоты встречаемости слова.

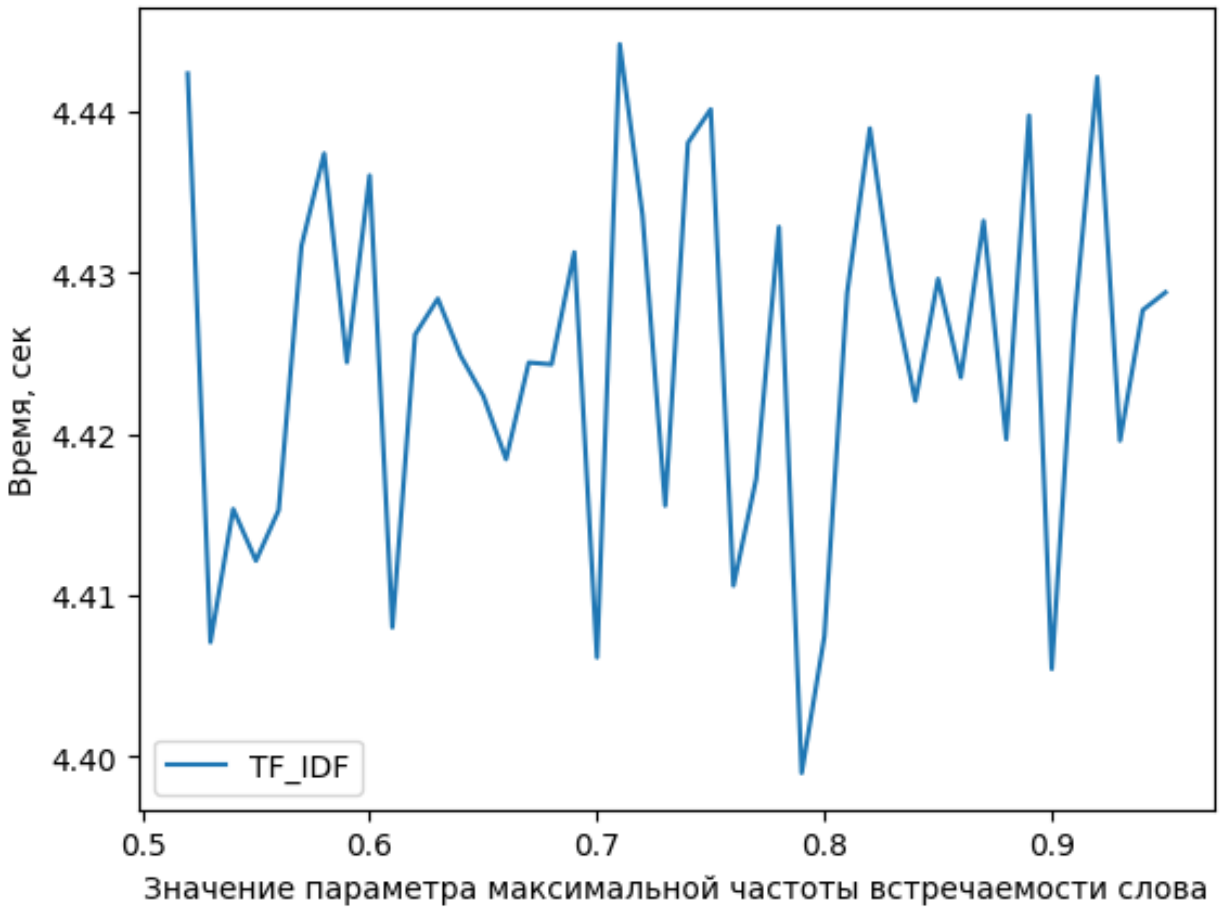

Рис. 4.1: График зависимости времени исполнения TF-IDF от значения параметра максимальной частоты встречаемости слова.

4.3 Зависимость времени исполнения TF-IDF от значения параметра минимальной встречаемости слова

На рисунке 4.2 представлен график зависимости времени исполнения TF-IDF от значения параметра минимальной встречаемости слова.

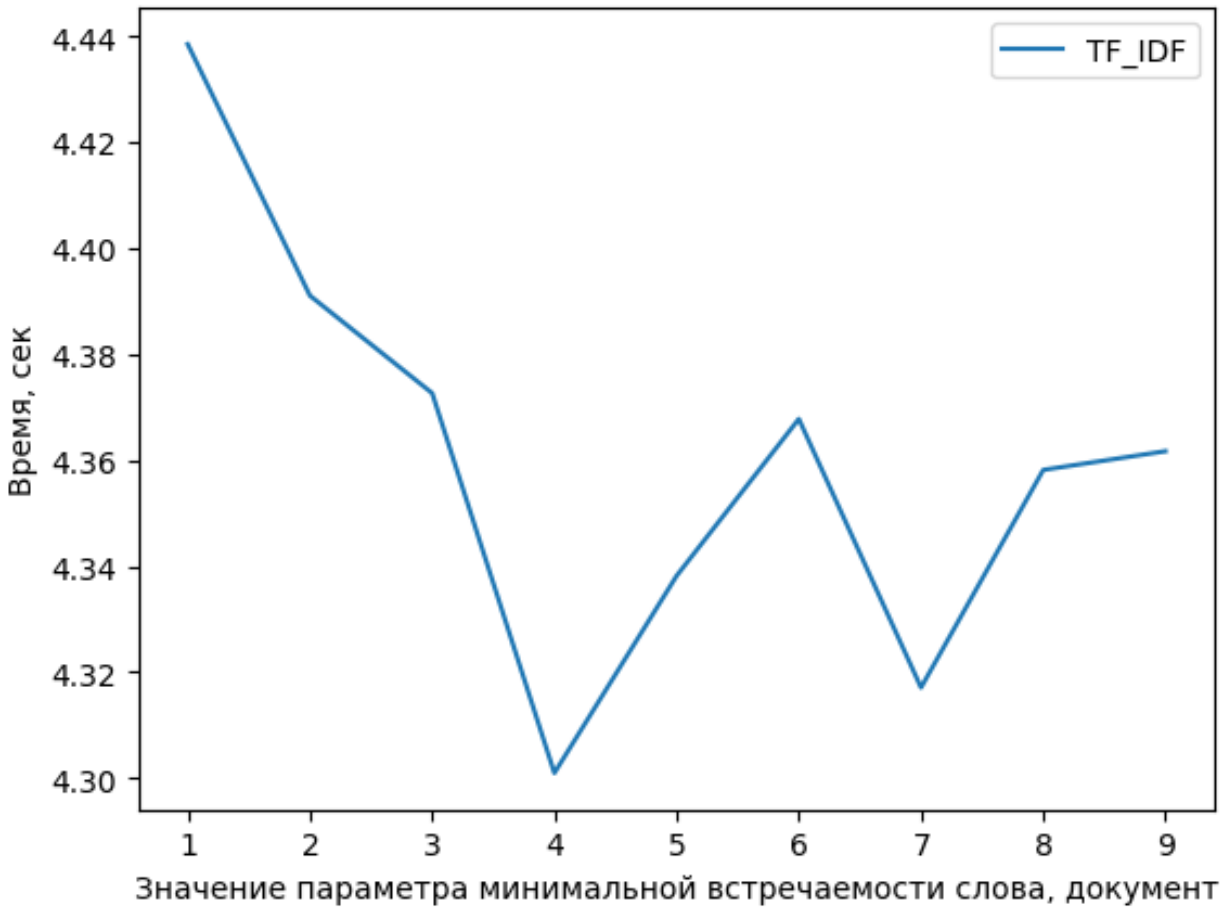

Рис. 4.2: График зависимости времени исполнения TF-IDF от значения параметра минимальной встречаемости слова.

4.4 Зависимость времени исполнения LDA от значения параметра количества тем

На рисунке 4.3 представлен график зависимости времени исполнения LDA от значения параметра количества тем с разделением по методу обучения.

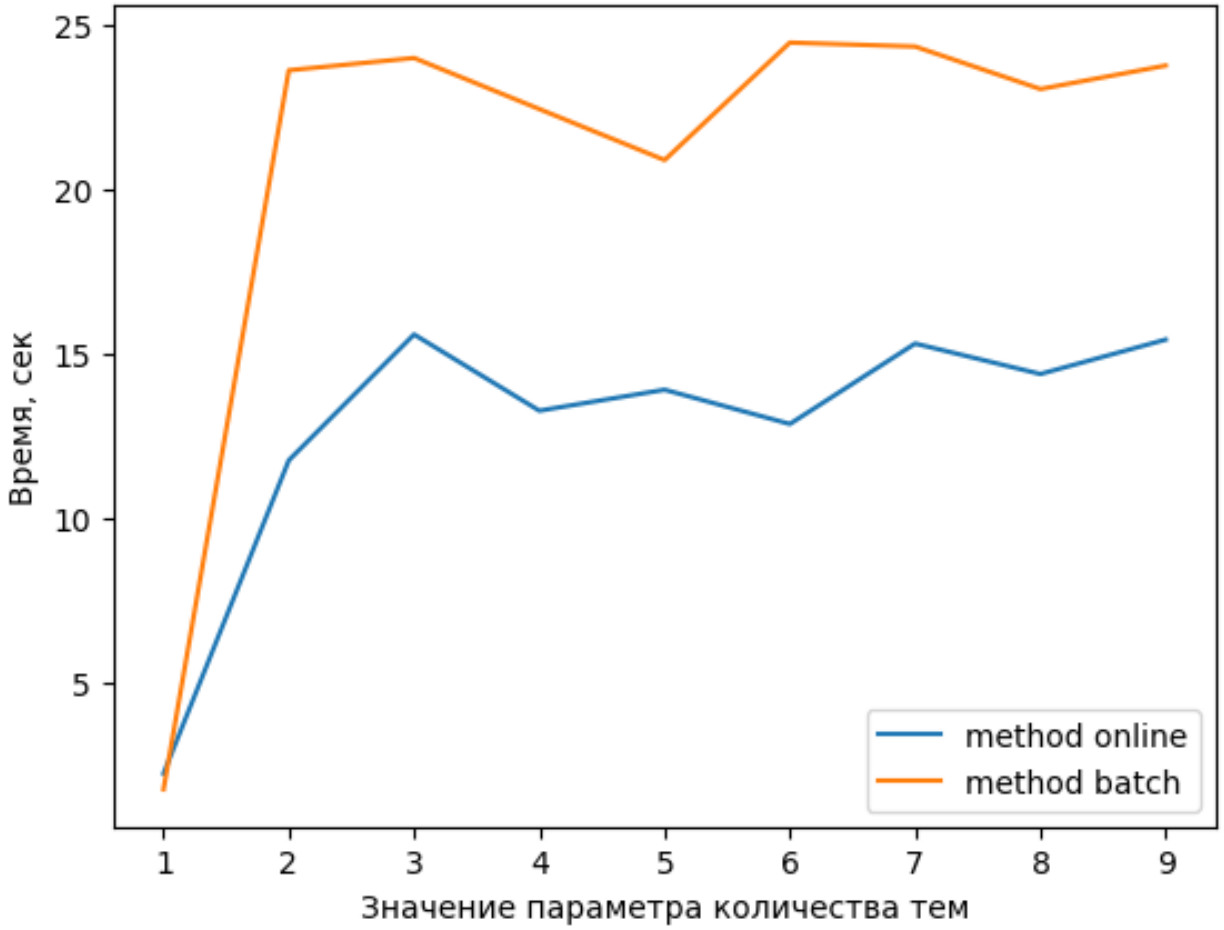

Рис. 4.3: График зависимости времени исполнения LDA от значения параметра количества тем.

4.5 Зависимость времени исполнения LDA от значения параметра количества эпох

На рисунке 4.4 представлен график зависимости времени исполнения LDA от значения параметра количества эпох с разделением по методу обучения.

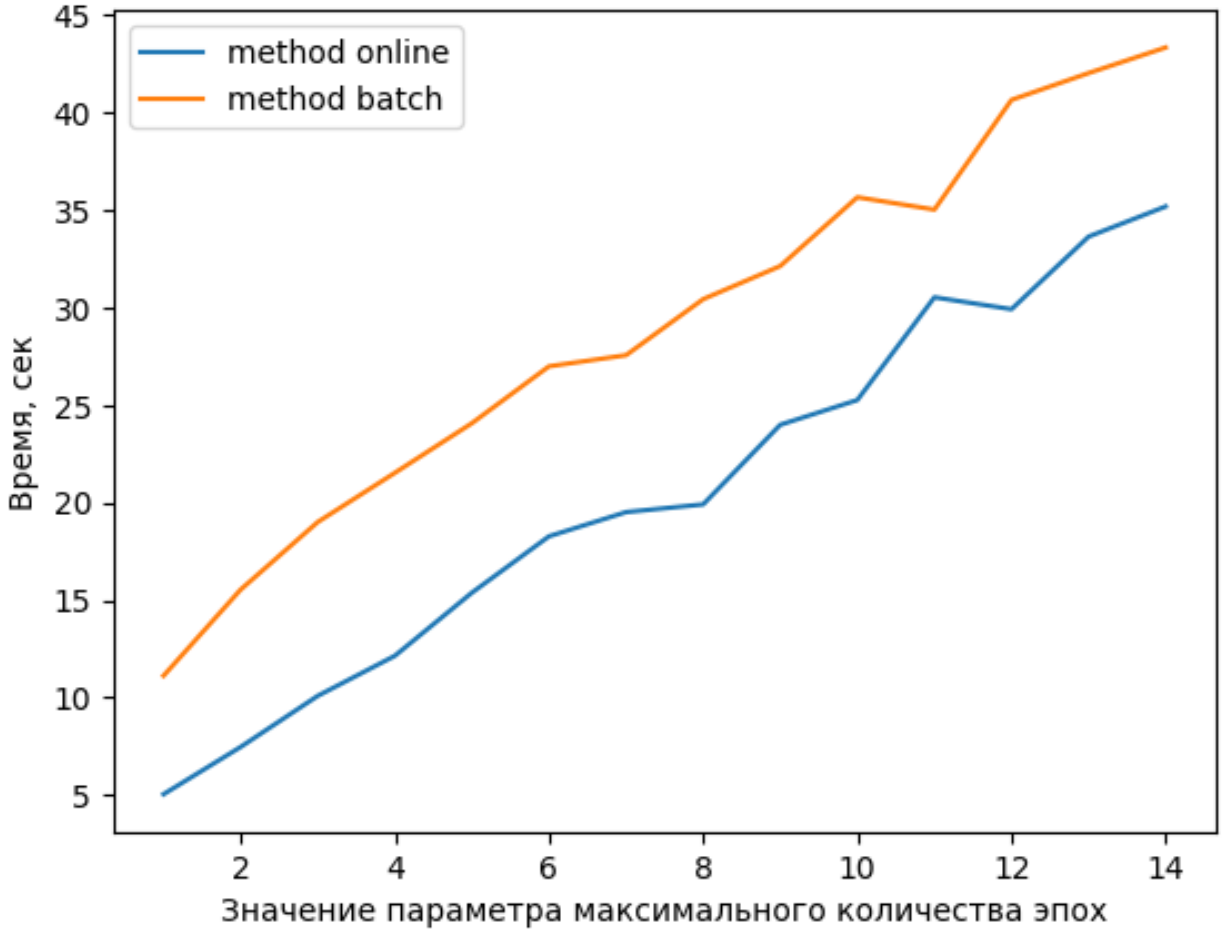

Рис. 4.4: График зависимости времени исполнения LDA от значения параметра количества эпох.

Заключение

В результате проведенных исследований легко заметить, что даже при факте включения во время исполнения предобработки данных, TF-IDF в среднем на данном датасете показал себя в ≈ 5 раз быстрее LDA.

Также можно увидеть, что в LDA лучше всего себя показал метод обучения “online”. При этом, что ожидаемо, при увеличении количества тем или эпох время исполнения алгоритма увеличивается.

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы было проведено сравнение алгоритмов коллаборативной фильтрации по пользователю и по объекту.

Были решены следующие задачи:

- приведено описание алгоритмов;
- приведено описание используемых для исследования данных;
- привести зависимости скорости работы алгоритмов от заданных параметров.

Проведенные исследования показали, что в среднем TF-IDF работает быстрее LDA в ≈ 5 раз, а в самом LDA параметр метода обучения “online” позволяет ускорить обучение модели.

Список литературы

1. Rajaraman A. Ullman J.D. Data Mining. 2011. С. 1–17.
2. Blei M. Ng Y. Jordan I. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. № 3.
3. MovieLens [Электронный ресурс]. Режим доступа: https://grouplens.org/datasets/movielens/ (дата обращения 16.09.2023).
4. Python official page [Электронный ресурс]. Режим доступа: https://www.python.org/ (дата обращения 10.05.2023).
5. Scikit-learn official page [Электронный ресурс]. Режим доступа: https://scikit-learn.org/stable/ (дата обращения 10.05.2023).
6. LibRecommender: official PyPl project page [Электронный ресурс]. Режим доступа: https://pypi.org/project/LibRecommender/ (дата обращения 16.09.2023).