Ati: Система за разпознаване на авторство

Симеон Христов

6MI3400191

SOFIA UNIVERSITY
St. KLIMENT OHRIDSKI

Курсов проект по
Извличане на информация

Факултет по математика и информатика
Софийски университет

Лектор: проф. Иван Койчев

Февруари 2023

Съдържание

1 Увод

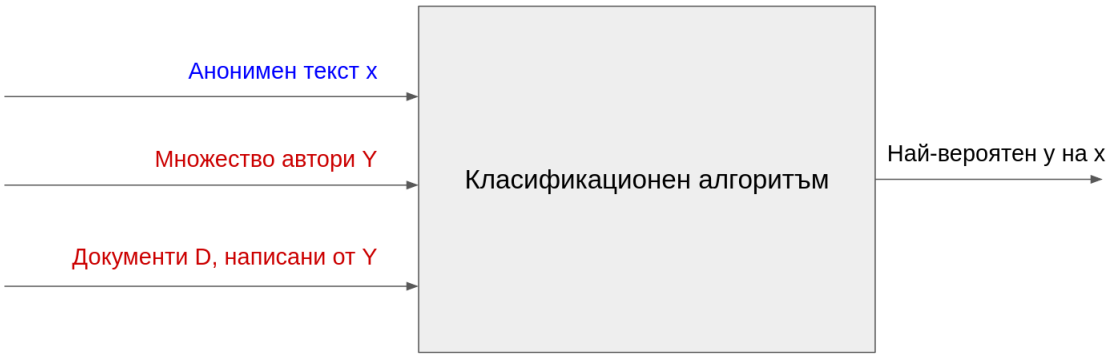
Целта на проекта е създаването на категоризационна система, която при даден корпус от документи - D, всеки от които е написан от един автор y, да идентифицира автора на анонимен текст x.

Разработената система може да бъде основата за разработване на приложение, което да:

1. **Проверява на авторство**: Дали даден текст наистина е написан от определен автор?
2. **Открива плагиатство**: Намиране на прилики между два или повече текста;
3. **Създава профил или характеризира на даден автора**: Извличане на информация за възрастта, образованието, пола и т.н. на автора на даден текст;
4. **Открива стилистични несъответствия** (както може да се случи при съвместно писане): Дали авторът настина е само един?
5. **Отговоря на въпроси**: В матурата по български език и литература има въпроси, които са фокусирани върху разпознаване на автора на даден отказ или разпознаване на автора, който пише за определен герой.

Задачите, които са реализирани от този проект са:

1. Създаване на паяк, копаещ документи (текстовете на съответните автори) от уеб страница.
2. Създаване на модел, който е трениран върху тренировъчно множество (80% от данните), валидиран върху валидационно множество (10% от оставащи данни) и оценен върху тестово множество (последните 10% оставащи данни).
3. Сравняване на поне 3 стилистични метрики за всички автори.
4. Сравняване на различни представяния на текст: *tf-tdf* и *transformer sentence embeddings*.



Фигура 1: Системата, представена "от птичи поглед".

5. Сравняване на класификатори: един и ансамбъл.
6. Създаване на потребителски интерфейс.

2 Преглед на областта по разпознаване на авторство

[ ] TODO: Подходи и методи за решаване; съществуващи решения; сравнителен анализ на решенията./ Подобни изследвания.

3 Проектиране

От гледна точка на машинното самообучение системата е представена схематично на Фигура 1.
Потребителският интерфейс е представен на .
[ ] TODO: Анализ на изискванията, Обща архитектура – напр. слоеве, модули, блокове, компоненти...; Модел на данните; Схема за представяне на знанията. Диаграми; Потребителски интерфейс (ако има); Ресурсни;. . .

4 Реализация, тестване/експерименти

4.1 Използвани технологии, платформи и библиотеки

[ ] TODO: Подходящи средства за реализация за проекта (технологии, платформи и библиотеки). Избор на средствата и начин за използването им;

4.2 Реализация/Провеждане на експерименти

[ ] TODO: Реализация (на модулите); За система/приложение: На кратко: планиране на тестването - тестови сценарии,...; Анализ на резултатите.

5 Заключение

[ ] TODO: Обобщение на направеното/резултатите. Идеи за по-нататъшно развитие, усъвършенстване или други експерименти.

6 Използвани технологии

[ ] TODO: (статии, книги, онлайн ресурси, форматирани съгласно MLA Style - http://www.library.mun.ca/guides/howto/mla.php) или друг подобен стандарт.

7 Използвана литература

Списък на фигурите