Ati: Система за разпознаване на авторство

Симеон Христов

6MI3400191

SOFIA UNIVERSITY
St. KLIMENT OHRIDSKI

Курсов проект по
Подходи за обработка на естествен език

Факултет по математика и информатика
Софийски университет

Лектор: проф. дмн Галя Ангелова

Февруари 2023

1 Въведение

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

2 Данни

3 Метод

Изискванията за разработване на системата 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*.
5. Сравняване на класификатори: един и ансамбъл.
6. Създаване на потребителски интерфейс.

4 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 Резултати. Дискусия

6 Заключение и бъдеща работа

Разработена е система, която може да класифицира даден текст или откъс от текст на базата на различни негови представяния.
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.