­­­­

СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

КУРСОВ ПРОЕКТ

ПО СИСТЕМИ, ОСНОВАНИ НА ЗНАНИЯ

Тема:

**Разпознаване на авторство (класификатор на автори). Класификатор, който различава текстове на Алеко Константинов от текстове на Захари Стоянов.**

Студенти:

Боян Иванов Дончев, група I, фн: 72029

Атанас Живков Атанасов, група I, фн: 72028

София, януари 2023 г.

1. Формулировка на задачата

Задачата реализира класифициране на текстове на Захари Стоянов и Алеко Константинов. Задачата приема текстов файл и извежда вероятността той да е написан от Захари Стоянов или Алеко Константинов. Реализацията на проекта е написана на javascript. За класификацията е необходима база от текстове на двамата автори.

1. Използвани алгоритми

Всяка дума от входния файл се сравнява с думите от базата текстове на двамата автори и се изпълнява следния алгоритъм:

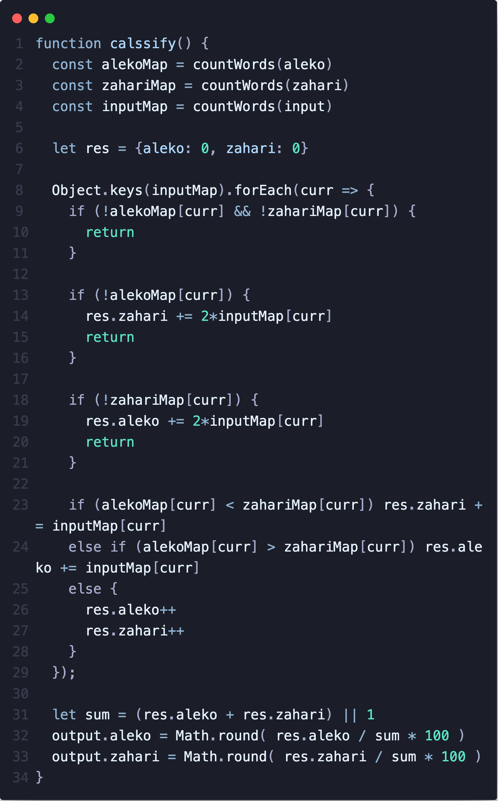
* Ако се среща в текстовете на единия автор повече от колкото на другия, увеличаваме процентната вероятност на съответния автор.
* Ако се среща само в текстовете на един от двамата автори, тогава увеличаваме вероятността двойно повече.
* Ако се среща в текстовете на двамата автори по равно, променяме вероятността пропорционално на дотук достигнатия резултат.
* Ако не се среща в нито един от текстовете, вероятността не се променя.

1. Описание на програмната реализация

Text

Description automatically generated

Функцията ***countWords()*** преобразува текст в честотна таблица на думите му. От текста се пропускат всички специални символи и се игнорира капитализацията на символите.



Функцията ***classify()*** създава честотни таблици от текстовете на двамата автори и входния текст и прилага алгоритъма описан в точка 2.

За пресмятане на процентната вероятност се използват два брояча, чиито стойности се увеличават спрямо алгоритъма.

Накрая се превръщат в проценти и се записват в глобална променлива.

Text

Description automatically generated

Функцията ***print()*** изписва форматиран текст с резултата от класифицирането, спрямо глобалната променлива.

1. Примери, илюстриращи работата на програмната система

Ако Алеко е написал „atanas atanas atanas”, a Захари е написал „boyan boyan boyan” и входният текст е:

* „atanas atanas boyan” 🡪 “The input text is 67% more likely to be authored by Aleko Konstantinov then by Zahari Stoyanov.”
* “boyan” 🡪 “The input text is 100% more likely to be authored by Zahari Stoyanov then by Aleko Konstantinov.”
* “nasko” 🡪 “The input can't be classified between Aleko Konstantinov and Zahari Stoyanov.”
* “boyan atanas” 🡪 “The input can't be classified between Aleko Konstantinov and Zahari Stoyanov.”

1. Литература

За база от текстове на двамата автори са използвани текстове от <https://chitanka.info/>.