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.

2. 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.

3. Описание на програмната реализация

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

```
function countWords(text) {
  var wordArr =
    text.replace(/[.,\/#!?$%\^&\*;:{}=\-_`~()]/g, '').
    split(' ')

  const res = {}

  wordArr.forEach(word => {
    word = word.toLowerCase()
    if (res[word]) res[word] += 1
    else res[word] = 1
  });

  return res
}
```

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

```
function calssify() {
  const alekoMap = countWords(aleko)
  const zahariMap = countWords(zahari)
  const inputMap = countWords(input)

  let res = {aleko: 0, zahari: 0}

  Object.keys(inputMap).forEach(curr => {
    if (!alekoMap[curr] && !zahariMap[curr]) {
      return
    }

    if (!alekoMap[curr]) {
      res.zahari += 2*inputMap[curr]
      return
    }

    if (!zahariMap[curr]) {
      res.aleko += 2*inputMap[curr]
      return
    }

    if (alekoMap[curr] < zahariMap[curr]) res.zahari +
= inputMap[curr]
    else if (alekoMap[curr] > zahariMap[curr]) res.ale
ko += inputMap[curr]
    else {
      res.aleko++
      res.zahari++
    }
  });

  let sum = (res.aleko + res.zahari) || 1
  output.aleko = Math.round( res.aleko / sum * 100 )
  output.zahari = Math.round( res.zahari / sum * 100 )
}
```

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.

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

```
function print() {
  if (output.aleko == output.zahari) {
    console.log("\nThe input can't be classified betwee
n Aleko Konstantinov and Zahari Stoyanov.\n")
    return
  }
  console.log(`\nThe input text is ${Math.max(output.al
eko, output.zahari)}% more likely to be authored by ${o
utput.aleko > output.zahari ? "Aleko Konstantinov then
 by Zahari Stoyanov" : "Zahari Stoyanov then by Aleko K
onstantinov"}.\n`)
}
```

4. <u>Примери, илюстриращи работата на програмната система</u>

Ако Алеко е написал „atanas atanas atanas”, а 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.”

5. Литература

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