СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

КУРСОВ ПРОЕКТ
ПО СИСТЕМИ, ОСНОВАНИ НА ЗНАНИЯ

Тема:

Резюме на текст

Студент:
Николета Радостинова Далакчиева, група 1, ФН. 71867

София, януари 2021 г.

1. Формулировка на задачата

   Да се напише програма, която прави резюме на текст.

2. Използван алгоритъм
   - Оценяване тежестта на всяка фума Намираме думите, които не носят голямо значение, но се срещат изключително често (*the, and, u, че* и други), както числа, симполи и прочие. След като имаме списък с тези така наречени „stop words” , създаваме таблица, в която записваме колко пъти се среща дума, която не е от отзи списък.
   - Разбиваме текста на изреченя.
   - Оценяваме всички изречения спрямо тежестта на „stop words” и срещането.
   - Намираме средната тежест на всички изречения.
   - Взъмаме изреченията с най-голяма тежест.

3. Описание на програмната реализация

   - Програмата започва с main функцията, която очаква от потребителя да въведе текста. Програмата приема както файлове, така и линкове. Когато се подаде път до файл, просто зарежда информацията за файла. При подаване на линк, се взима текста от всички <p> тагове от html страницата.
   - След като текста е зареден се изплозва detect функцията от библиотеката langdetect, която разпознава езика на текста.

     ```
     # find language
     lang = detect(text_to_summer)
     ```

   - Спрямо разпознатия език се зареждат за съответния език “stop words”

     ```
     # get stop words
     stop_words = stopwords.stopwords(lang)
     ```

   - Разделяме думите от текста. Започваме да обхождаме всяка дума, като използваме PorterStemmer, за да я приведем в нейната основна форма. Ако думата е част от stop words, то не правим нищо, ако не, добавяме еденица към тази таблица с цел да намерим колко често се среща дадена дума.

Примерна таблица:

pythonidae	1
python	52
famili	5
…	…

- След като имаме готова таблицата, разделяме текста на изречения и почваме да ги обхождаме. Преброяваме stop words в това изречение. Пресмятаме и спрямо горната таблица чистата тежест на изречението. Примерно ако изречението съдържа думата python, увеличаваме тежестта му с 52. Пълната тежест изчисляваме като чистата тежест разделим на броя на stop words. Създаваме нова табица, в която ще записваме номера на изречението (започвайки от 1) и пълната тежест на изречението.

```
for sentence in sentences:
    stop_words = 0
    for word_weight in frequency_table:
        if word_weight in sentence.lower():
            stop_words += 1
            sentence_number += 1
            if sentence_number in sentence_weight:
                sentence_weight[sentence_number] += frequency_table[word_weight]
            else:
                sentence_weight[sentence_number] = frequency_table[word_weight]

    sentence_weight[sentence_number] = sentence_weight[sentence_number] / stop_words
```

- Знаейки тежестта на всяко изречение, намираме сбора от всички и ги разделяме на броя им (намираме средното аритметично).
- Имаме изреченията, имаме тежестта им, имаме средната тежест, остава да ги сглобим в текст. Това правим като вземем най-тежките изречения, т.е. тези които имат тежест по-голяма от средната.
- Вуаля! Вече няма нужда да четем целия текста, отсана само най-важното :)

4. Примери, илюстриращи работата на програмната система
   - [Резюме на текст за Чарлз Велики]

```
Enter path to the text:
test_file.txt
 Карл Велики. В продължение на векове името му е легенда. Карлос Магнус (" Чарлс Велики "), крал на франките и ломбардите, свещеният римски
 император, обект на многобройни епоси и романи - дори бил светец. Като фигура на историята той е по-голям от живота. Карл Велики беше женен пет
 пъти и имаше много наложници и деца. Той почти винаги държеше голямото си семейство, като понякога принуждаваше синовете му да водят кампании.
 Уважавал католическата църква достатъчно, за да натрупа богатство върху нея (акт на политическо предимство, колкото и духовно благоговение), но
 той никога не се подчинявал напълно на религиозния закон. Това причинява значително триене между братята, които майка им, Бертрада, се изгладила
  до смъртта на Карломан през 771 г.

Чарлс Завоевателя
Подобно на баща си и дядо си пред него, Карл Велики разширява и укрепва французите с насилствена сила. Той спонсорира манастири, където се
 запазват и копират древните книги. Докато изолацията им от римокатолическата църква изпрати известните ирландски манастири в упадък,
 европейските манастири бяха твърдо установени като пазители на знанието благодарение отчасти на франкския крал. С напътствия от доверения си
 съветник Алкуин, Карл Велики пренебрегва ограниченията на властта, наложени от Църквата, и продължава да се движи по собствен път като владетел
 на Франклин, който сега заема огромна част от Европа. Концепцията за император на Запад беше установена и щеше да придобие много по-голямо
 значение през идните векове. Наследството на Чарлз Велики
Докато Карл Велики се опитвал да възобнови интереса си към ученето и обединяването на различни групи в една нация, той никога не се занимаваше с
 технологичните и икономически трудности, пред които е изправена Европа, след като Рим вече не осигурява бюрократична хомогенност. Но това са
 само неуспехи, ако целта на Карл Велики е да възстанови Римската империя .
```

- Резюме на Уики страницата за питон.

```
Enter path to the text:
https://en.wikipedia.org/wiki/Pythonidae
 Among its members are some of the largest snakes in the world. Ten genera and 42 species are currently recognized. [4][5]
Pythons use their sharp, backward-curving teeth, four rows in the upper jaw, two in the lower, to grasp prey which is then killed by
 constriction; after an animal has been grasped to restrain it, the python quickly wraps a number of coils around it. Contrary to popular belief,
  even the larger species, such as the reticulated python, P. reticulatus, do not crush their prey to death; in fact, prey is not even noticeably
  deformed before it is swallowed. This sets them apart from the family Boidae (boas), most of which bear live young (ovoviviparous). [11][12]
Poaching of pythons is a lucrative business with the global python skin trade being an estimated US$1 billion as of 2012. [14] Pythons are
 poached for their meat, mostly consumed locally as bushmeat and their skin, which is sent to Europe and North America for manufacture of
 accessories like bags, belts and shoes. [15]  The poaching of the pythons is illegal in Cameroon under their wildlife law, but there is little
 to no enforcement. In Kenya, there has been an increase in snake farms to address the demand for snake skin internationally, but there are
 health concerns for the workers, and danger due to poachers coming to the farms to hunt the snakes. Pythons are disease vectors for multiple
 illness, including Salmonella, Chlamydia, Leptospirosis, Aeromoniasis, Campylobacteriosis, and Zygomycosis. [23] It is very common for the body
 fat of pythons to be used to treat a large variation of issues such as joint pain, rheumatic pain, toothache and eye sight. [24] Additionally,
 python fat has been used to treat those suffering from mental illnesses like psychosis. [25] Their calm nature is thought to be of use to treat
 combative patients. To improve mental illnesses, it is often rubbed on the temple. [24]
Python blood plays another important role in traditional medicine. [30]
```

- Резюме на статия за змиите.

```
Enter path to the text:
https://bg.wikipedia.org/wiki/%D0%97%D0%BC%D0%B8%D0%B8
 Змиите са удължени, студенокръвни безкраки влечуги от подразред Serpentes, близки родственици на гущерите, с които спадат към един и същи
 разред - Люспести. Подобно на останалите Люспести, змиите са амниотни гръбначни животни покрити с люспи. [3]
Вкаменените останки от змии са сравнително редки, поради това че скелетът на змията е изключително крехък и малък. [4] Въз основа на
 сравнителната анатомия днес има консенсус, че змиите произхождат от гущери. [4][6] Представителите на семейства Leptotyphlopidae и Typhlopidae и
  днес притежават следи от тазовия пояс. Предните крайници липсват при всички змии. В началото на еволюцията на змиите хомеозисните гени
 активират развитието на гръдните прешлени. Това стана успоредно с адаптивната радиация на бозайниците, след изчезването на динозаврите. Същите
 обаче са нямали връзка с прешлените. Сред тях са фосилизираните видове на изкопаемите родове Haasiophis, Pachyrhachis и Eupodophis. От своя
 страна те са произлезли от вараноподобни гущери. [5] Според тази хипотеза защитната мембрана на роговицата се е развила за предпазване на окото
 в морски условия. Въпреки това обаче точната им класификация в разреда остава спорна. [11]
Подразред Serpentes има два инфраразреда - Alethinophidia и Scolecophidia[11] Това разделение се базира на морфологичните характеристики и
 изследване на митохондриалната ДНК последователност. Признак за сменяне на кожата е помътняването на очите. Скелетът на змиите се състои само от
  череп, подезични кости, гръбначен стълб и ребра. Само змиите от надсемейство Henophidia имат следи от таза и задните крайници. Черепът на
 змията се състои от твърди и неподвижно свързани кости, предпазващи мозъка. Като прибавка към зрението, някои змии (някои отровници, питони и
 бои) имат инфрачервени сензори в ямките, които се намират между очите и ноздрите, което им позволява да виждат телесната топлина, излъчвана от
 жертвите или неприятелите им. Тъй като змиите нямат външни уши, слухът им е ограничен до усещане на вибрации, но това сетиво е много добре
 развито. [18] При всички отровни змии тези жлези се изливат през канали в набраздени или кухи зъби от горната челюст. [17][19] При апсидовите и
 отровниците тези зъби са разположени в предната част на устата и са кухи, така че да инжектират отровата по-ефективно, докато при змиите със
 задно разположени отровни зъби отровата преминава по открит прорез на задната страна на зъба. Обикновено отровата е смес от невротоксини (които
 въздействат на нервната система), хемотоксини (които въздействат на кръвоносната система), цитотоксини и много други токсини, които действат на
 тялото по различни начини. [17]
```

5. Литература

- Text Mining II, G. Neumann, M. Venkataramani, R. Altman, L. Hirschman, and D. Radev: http://web.stanford.edu/class/cs276b/handouts/lecture14.pdf
- Summarization Techniques, D. Radev: https://www.youtube.com/watch?v=N5N-HCUE3G4
  https://www.youtube.com/watch?v=cz8UImIopnQ
  https://www.youtube.com/watch?v=AgvfJddkzvE&t=527s
- https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/
- https://towardsdatascience.com/comparing-text-summarization-techniques-d1e2e465584e

- https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/