ASSIGNMENT-3

Text Summarizer and extractor

Natural Language Processing

Submitted by:

101703320 Manav Sharma

9877392710 ,manavsharma136@gmail.com

BE Third Year- COE15

\

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala

PROJECT OVERVIEW

In this minor project one can enter the file with any long text or email and this Project will summarize the file and will extract phone numbers and email from that paragraph so it will save time and will summarize the whole story and will give out important data from that text.

The summarizer tokenizes input into sentences and further we will compute frequency map of words then this map will remove very high and low frequent words that words that don't contain much information and are very less frequent and in this way it will discard noisy words and will remove that words which donot contain much information . And finally, the sentences are ranked according to the frequency of the words they contain and the top sentences are selected for the final summary.

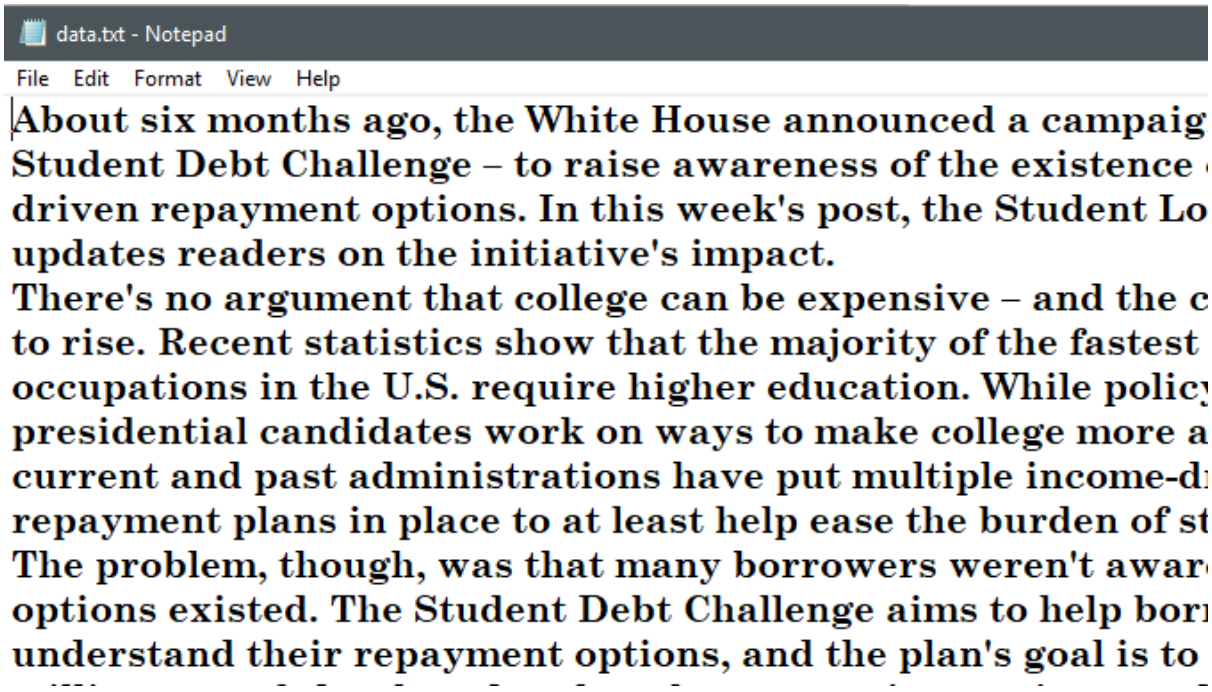
INPUT:

Any message or any email can be entered in the input file

Here is example of input text

About six months ago, the White House announced a campaign – the Student Debt Challenge – to raise awareness of the existence of income-driven repayment options. In this week's post, the Student Loan Ranger updates readers on the initiative's impact.
There's no argument that college can be expensive – and the costs continue to rise. Recent statistics show that the majority of the fastest growing occupations in the U.S. require higher education. While policymakers and presidential candidates work on ways to make college more affordable, the current and past administrations have put multiple income-driven repayment plans in place to at least help ease the burden of student debt.
The problem, though, was that many borrowers weren't aware that these options existed. The Student Debt Challenge aims to help borrowers better understand their repayment options, and the plan's goal is to enroll 2 million more federal student loan borrowers into an income-driven plan.
The idea behind the initiative is to get employers involved in helping spread the word about these options. For example, Fidelity Investments introduced a program called the Step Ahead Student Loan assistance program, which not only provides tenured employees with a student loan repayment benefit but also educates them on their repayment and, if eligible, forgiveness options. Rite Aid is also working with their 90,000 associates to ensure they are aware of the income-driven plans as well as reminding them of when to enroll and recertify their plans on an annual basis. And the National Housing Resource Center is training 500 housing counselors to work with clients to help them identify repayment plans that might help their overall financial circumstances.

These are just a few of the more than 40 organizations that have pledged to help spread the word and educate and advise borrowers on their repayment options. The Department of Education has also issued a free toolkit that employers can use to help educate their employees.
So far, the results are encouraging. After just three months, the campaign has increased participation to 23 percent of all borrowers.
That's a big jump from the 5 percent total enrollment four and a half years prior. The data also shows that these plans seem to be helping the exact populations it was intended to – those with higher debts and lower incomes.
What does this mean to prospective students and their families who are just starting to look into college financing? Although we always encourage families to minimize student debt to levels they know they will be affordable after graduation, choosing a field – such as finance and business – that tends to offer such employer assistance is another factor to consider. And now there is a growing trend of employers, both on their own and with the encouragement of the administration, who are assisting employees with managing their student debt.
While the benefit of student loan repayment isn't necessarily a new phenomenon, more recent college grads are looking for and more employers are now offering student loan repayment benefits in the form of payment reimbursement, help with student loan management or both. These benefits can help extend employee tenure, since young employees tend to switch jobs more often, as well as help relieve the stress of personal debt, which can increase employee productivity and decrease instances of internal fraud.
The bottom line is that when it comes to student debt, it takes a village. Employers helping their employees to take on the challenge of student debt is just another logical aspect step in this process.
you can call us at 9646860040 and manavsharma136@gmail.com



Data.txt file having paragraph which will be summarized

OUTPUT:

```
C:\Users\dell>cd desktop

C:\Users\dell\Desktop>cd ASD

C:\Users\dell\Desktop\ASD>python extract.py
THIS IS YOUR SUMMARIZED TEXT:

About six months ago, the White House announced a campaign - the Student Debt Challenge - to raise awareness of the existe
nce of income-driven repayment options.
While the benefit of student loan repayment isn't necessarily a new phenomenon, more recent college grads are looking for
and more employers are now offering student loan repayment benefits in the form of payment reimbursement, help with studen
t loan management or both.
The Student Debt Challenge aims to help borrowers better understand their repayment options, and the plan's goal is to enr
oll 2 million more federal student loan borrowers into an income-driven plan.
For example, Fidelity Investments introduced a program called the Step Ahead Student Loan assistance program, which not on
ly provides tenured employees with a student loan repayment benefit but also educates them on their repayment and, if elig
ible, forgiveness options.
While policymakers and presidential candidates work on ways to make college more affordable, the current and past administ
rations have put multiple income-driven repayment plans in place to at least help ease the burden of student debt.


Here is extracted emails
['manavsharma136@gmail.com']
Here is extracted numbers
['9646860040']

C:\Users\dell\Desktop\ASD>
```

CODE

```
"""
Created on Mon May 25 12:39:14 2020

@author: dell
"""

import re              #regex for extracting phone number and emails
import nltk
```

```
from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize, word_tokenize
from string import punctuation


import sys

stop_words = stopwords.words('english')




def phone(string):              #function that extract phone numbers based on
regular expressions
    r = re.compile(r'(\d{3}[-\.\s]??\d{3}[-\.\s]??\d{4}|\(\d{3}\)\s*\d{3}[-
\.\s]??\d{4}|\d{3}[-\.\s]??\d{4})')
    phone_numbers = r.findall(string)
    return [re.sub(r'\D', '', number) for number in phone_numbers]


def email(string):        #function that extract emailaddress based on
regular expressions
    r = re.compile(r'[\w\.-]+@[\w\.-]+')
    return r.findall(string)

def ie_preprocess(document):                                #preprocessing
function spliting and doing tokenization
    document = ' '.join([i for i in document.split() if i not in stop])
    sentences = nltk.sent_tokenize(document)
    sentences = [nltk.word_tokenize(sent) for sent in sentences]
    sentences = [nltk.pos_tag(sent) for sent in sentences]
    return sentences


def check(text):               #function for summarizing the text

        #Words that have a frequency term lower than mini or higher than maxi
will be ignored on basis of text rank algo.

        mini = 0.1      #setting frequency variable for taking words that are
more frequent these value 0.1 and 0.9 are arbitary can be different
        maxi = 0.9
        stopwords = set(stop_words+list(punctuation)+list("it's"))



   #we will Tokenize sentences and words

        sents = sent_tokenize(text)
        word_sent = [word_tokenize(s.lower()) for s in sents]


        #here will Compute the frequency of each word present in the text for
words in s.

        freq = dict()
```

```
        for s in word_sent:
            for word in s:
                if word not in stopwords:
                    if word not in freq:
                        freq[word] = 1
                    else:
                        freq[word] += 1
        # frequencies normalization and filerting   normalizing
        m = float(max(freq.values()))
        for w in list(freq):
            freq[w] = freq[w]/m
            if freq[w] >= maxi or freq[w] <= mini or w <= "a":
                del freq[w]

        ranking = dict()
        for i, sent in enumerate(word_sent):
            for w in sent:
                if w in freq:
                    if i not in ranking:
                        ranking[i] = freq[w]
                    else:
                        ranking[i] += freq[w]
        if 0 in ranking:
            del ranking[0]
        textLenReq = len(text.split())*0.25
        # sort sentences according to their values
        ranking = sorted(ranking,key=ranking.get,reverse=True)


        print('THIS IS YOUR SUMMARIZED TEXT:\n')

        print(sents[0])
        textLenReq -= len(sents[0].split())

        # print sentences in accordance with their values, as long as the
output length is less than 500 or 25% of total text length
        outputLen = 0
        i = 0
        while outputLen < textLenReq and outputLen < 500:
            #print('*')
            print(sents[ranking[i]])
            i += 1
            outputLen += len(sents[ranking[i]].split())


stop = stopwords.words('english')



if __name__ == '__main__':                #main function extracting data from
file and converting to string and than extracting phone number and emails
    string = open('data.txt', 'r').read()
    numbers = phone(string)
    emails = email(string)

    check(string)
    print('\n')
```

```
print("Here is extracted emails")
print(emails)
print("Here is extracted numbers")
print(numbers)
```