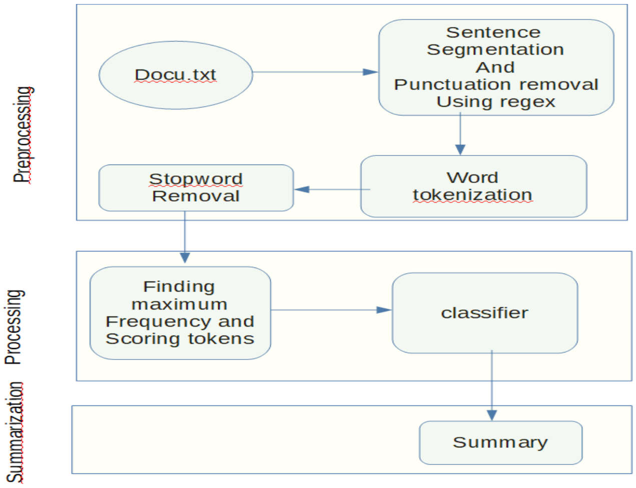
IR END-TERM PROJECT-1
INFORMATION RETRIEVAL (CSE4053)

Single Document Summary Extraction

This project includes various index construction methods. These methods are basically used to retrieve information from a documnet. The methods are implemented using Python3 and NLTK library. It takes an input of text file and produces the summary of that file.

Input: docu.txt

Model:


Preprocessing
Docu.txt
Sentence Segmentation And Punctuation removal Using regex
Stopword Removal
Word tokenization
Processing
Finding maximum Frequency and Scoring tokens
classifier
Summarization
Summary

Step-1: Document Pre-processing

-> sentence segmentation

-> punctuation removal using regex

-> word tokenization

-> stop word removal

Step-2: Processing

-> counting max frequency and scoring tokens

-> classifier

-> scoring sentences using token scores

-> sort sentences acc. to the scores

-> take best 4 sentences

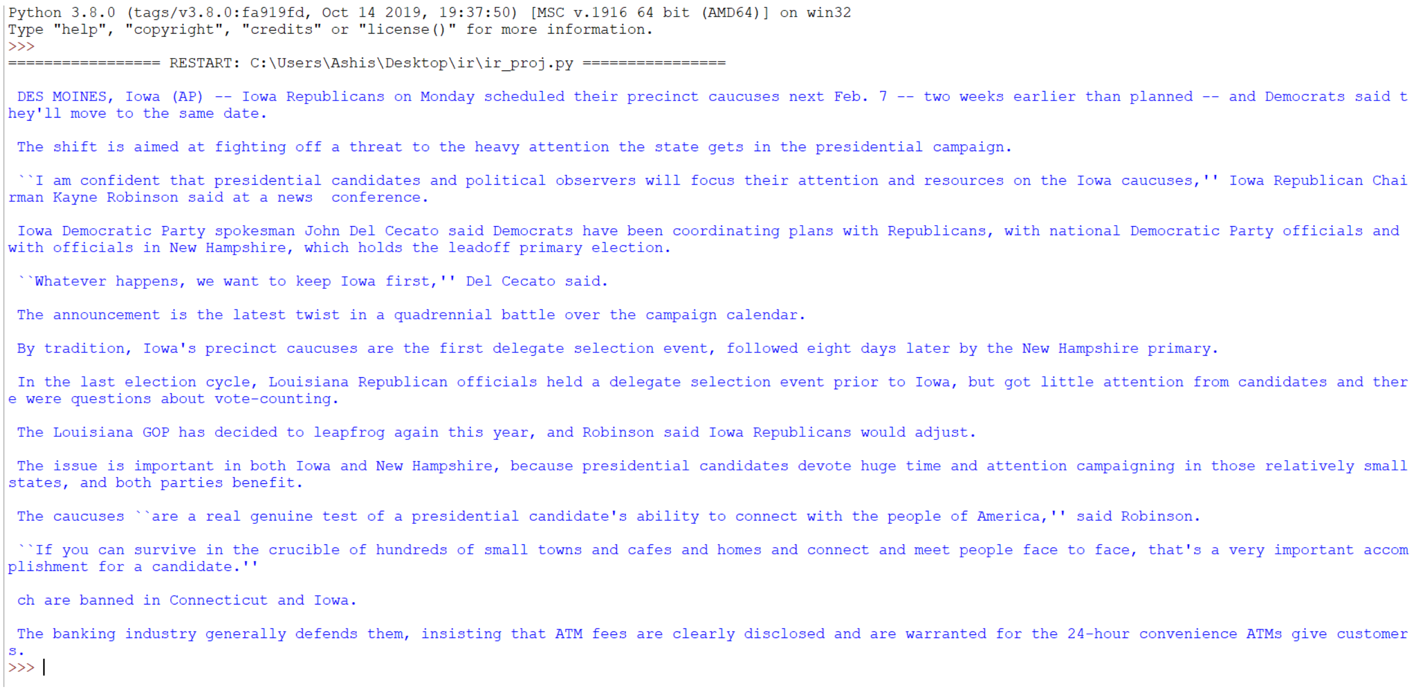
Step-3**:** Generate summary

Document Preprocessing

Sentence Segmentation:

input: docu.txt

output:

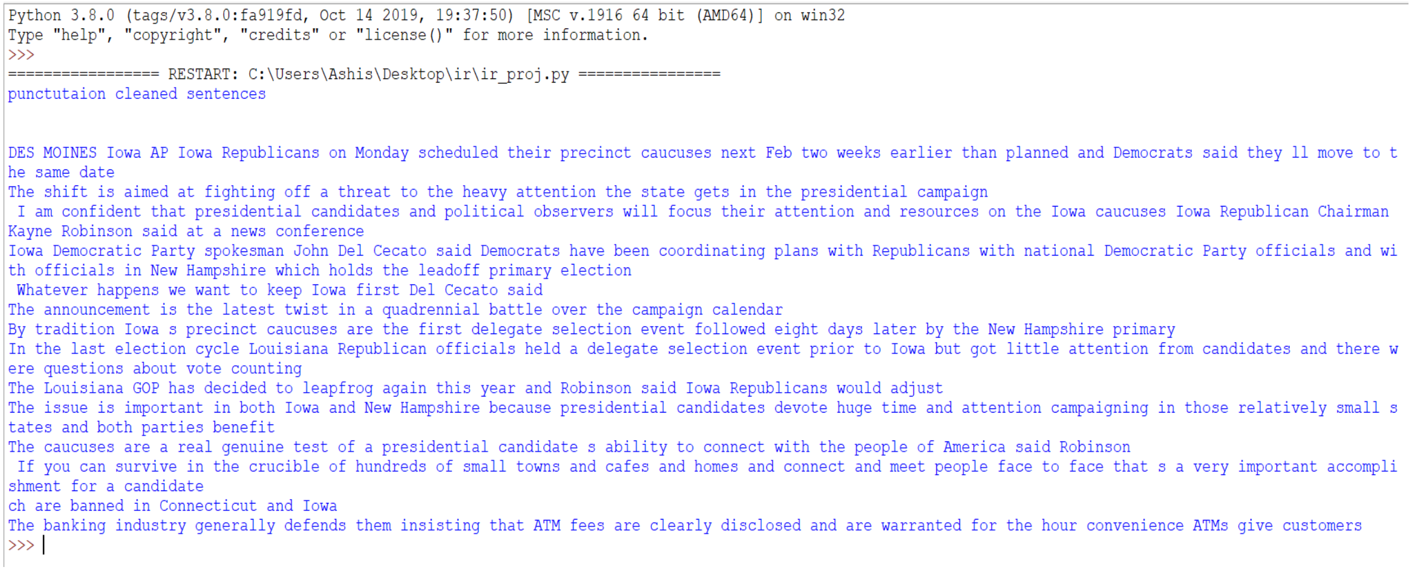


```
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> 
================= RESTART: C:\Users\Ashis\Desktop\ir\ir_proj.py ================

 DES MOINES, Iowa (AP) -- Iowa Republicans on Monday scheduled their precinct caucuses next Feb. 7 -- two weeks earlier than planned -- and Democrats said t
hey'll move to the same date.

 The shift is aimed at fighting off a threat to the heavy attention the state gets in the presidential campaign.

 ``I am confident that presidential candidates and political observers will focus their attention and resources on the Iowa caucuses,'' Iowa Republican Chai
rman Kayne Robinson said at a news  conference.

 Iowa Democratic Party spokesman John Del Cecato said Democrats have been coordinating plans with Republicans, with national Democratic Party officials and
with officials in New Hampshire, which holds the leadoff primary election.

 ``Whatever happens, we want to keep Iowa first,'' Del Cecato said.

 The announcement is the latest twist in a quadrennial battle over the campaign calendar.

 By tradition, Iowa's precinct caucuses are the first delegate selection event, followed eight days later by the New Hampshire primary.

 In the last election cycle, Louisiana Republican officials held a delegate selection event prior to Iowa, but got little attention from candidates and ther
e were questions about vote-counting.

 The Louisiana GOP has decided to leapfrog again this year, and Robinson said Iowa Republicans would adjust.

 The issue is important in both Iowa and New Hampshire, because presidential candidates devote huge time and attention campaigning in those relatively small
states, and both parties benefit.

 The caucuses ``are a real genuine test of a presidential candidate's ability to connect with the people of America,'' said Robinson.

 ``If you can survive in the crucible of hundreds of small towns and cafes and homes and connect and meet people face to face, that's a very important accom
plishment for a candidate.''

 ch are banned in Connecticut and Iowa.

 The banking industry generally defends them, insisting that ATM fees are clearly disclosed and are warranted for the 24-hour convenience ATMs give customer
s.
>>> |
```

Removing punctuations using regex:

input: tokenized sentences (i.e output from the previous step)

output:

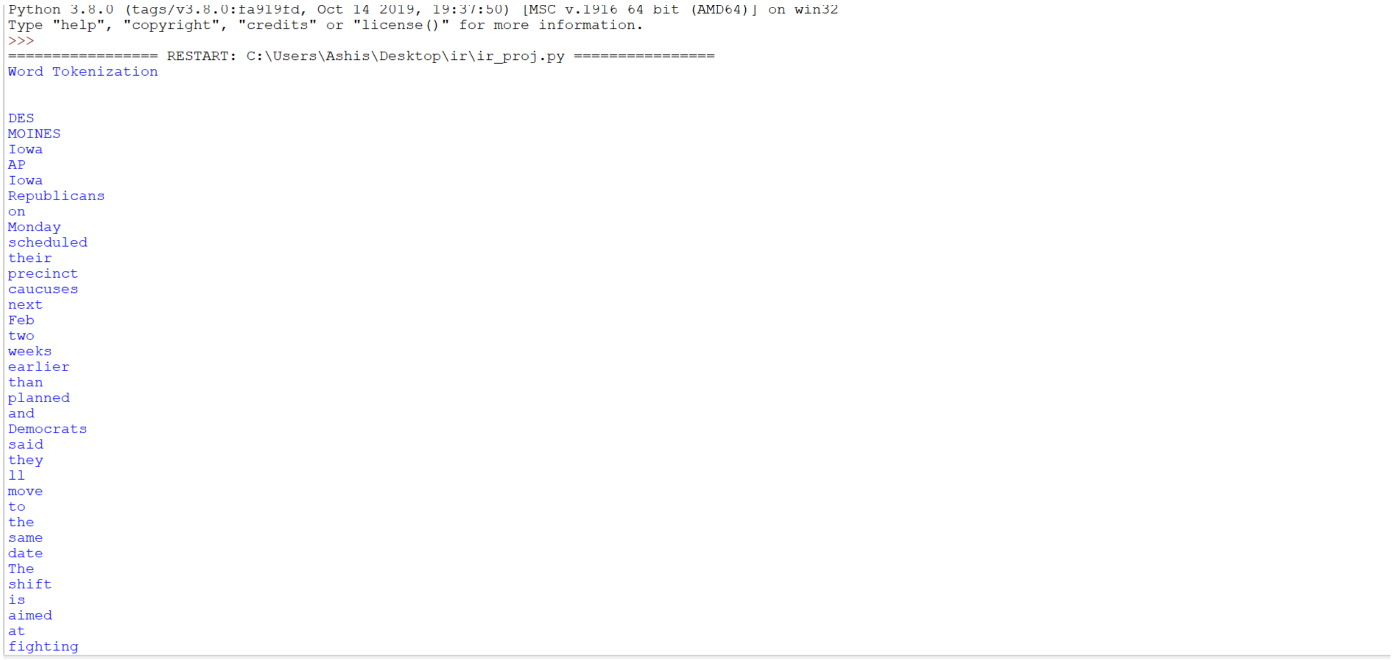


```
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> 
================= RESTART: C:\Users\Ashis\Desktop\ir\ir_proj.py ================
punctutaion cleaned sentences


DES MOINES Iowa AP Iowa Republicans on Monday scheduled their precinct caucuses next Feb two weeks earlier than planned and Democrats said they ll move to t
he same date
The shift is aimed at fighting off a threat to the heavy attention the state gets in the presidential campaign
 I am confident that presidential candidates and political observers will focus their attention and resources on the Iowa caucuses Iowa Republican Chairman
Kayne Robinson said at a news conference
Iowa Democratic Party spokesman John Del Cecato said Democrats have been coordinating plans with Republicans with national Democratic Party officials and wi
th officials in New Hampshire which holds the leadoff primary election
 Whatever happens we want to keep Iowa first Del Cecato said
The announcement is the latest twist in a quadrennial battle over the campaign calendar
By tradition Iowa s precinct caucuses are the first delegate selection event followed eight days later by the New Hampshire primary
In the last election cycle Louisiana Republican officials held a delegate selection event prior to Iowa but got little attention from candidates and there w
ere questions about vote counting
The Louisiana GOP has decided to leapfrog again this year and Robinson said Iowa Republicans would adjust
The issue is important in both Iowa and New Hampshire because presidential candidates devote huge time and attention campaigning in those relatively small s
tates and both parties benefit
The caucuses are a real genuine test of a presidential candidate s ability to connect with the people of America said Robinson
 If you can survive in the crucible of hundreds of small towns and cafes and homes and connect and meet people face to face that s a very important accompli
shment for a candidate
ch are banned in Connecticut and Iowa
The banking industry generally defends them insisting that ATM fees are clearly disclosed and are warranted for the hour convenience ATMs give customers
>>> |
```

Word Tokenization:

input: punctuation removed sentences

output:

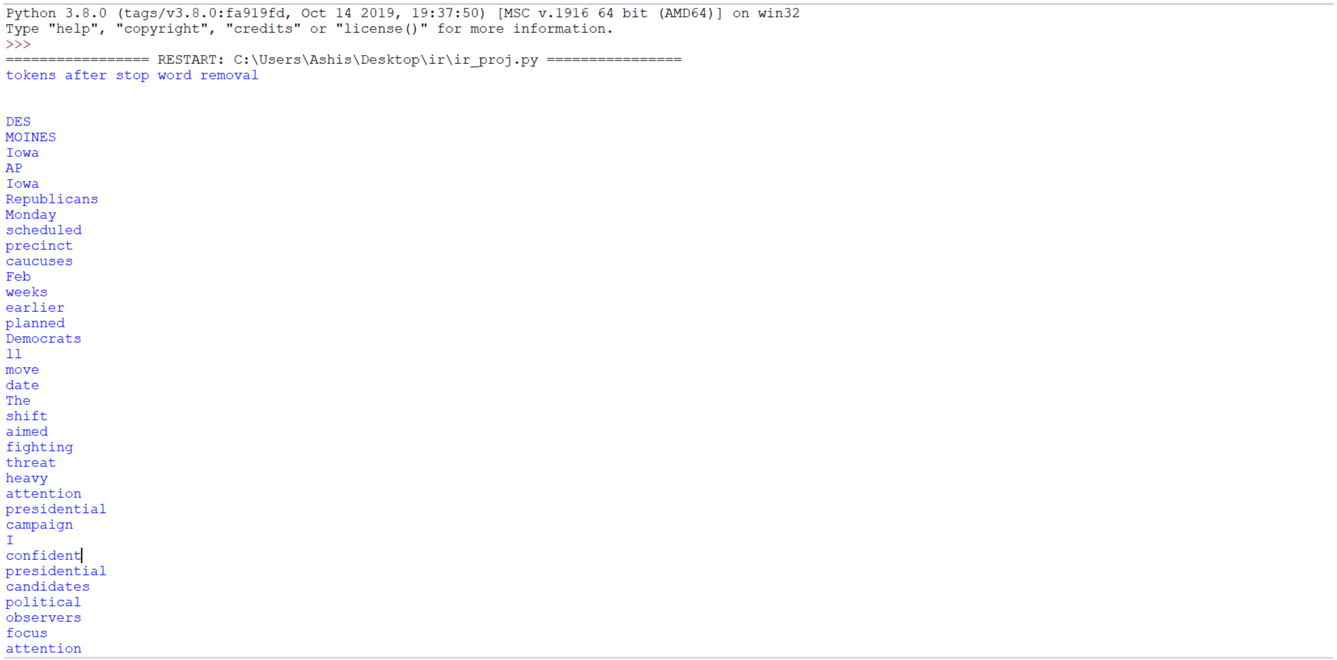


```
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
================= RESTART: C:\Users\Ashis\Desktop\ir\ir_proj.py ================
Word Tokenization


DES
MOINES
Iowa
AP
Iowa
Republicans
on
Monday
scheduled
their
precinct
caucuses
next
Feb
two
weeks
earlier
than
planned
and
Democrats
said
they
ll
move
to
the
same
date
The
shift
is
aimed
at
fighting
```

Stopword Removal and word frequency:

input: word tokens

output:

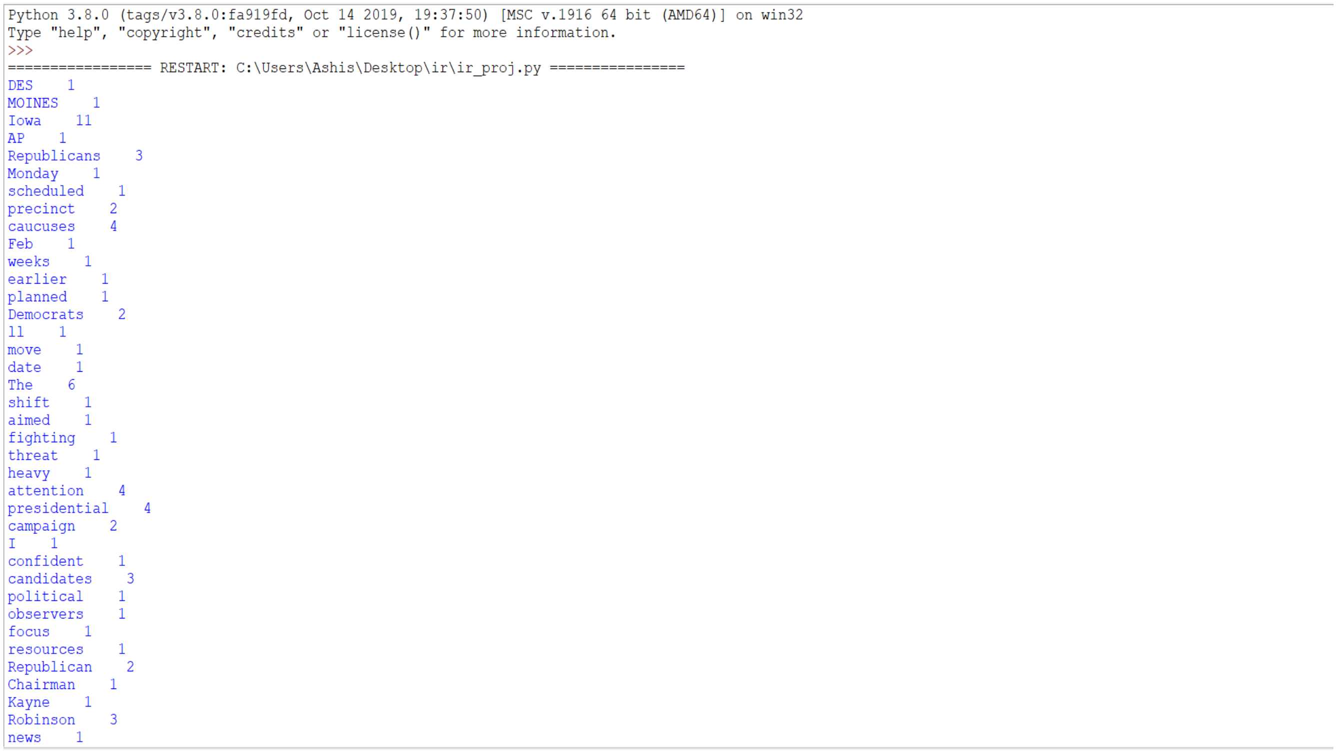


```
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
================= RESTART: C:\Users\Ashis\Desktop\ir\ir_proj.py ================
tokens after stop word removal


DES
MOINES
Iowa
AP
Iowa
Republicans
Monday
scheduled
precinct
caucuses
Feb
weeks
earlier
planned
Democrats
ll
move
date
The
shift
aimed
fighting
threat
heavy
attention
presidential
campaign
I
confident
presidential
candidates
political
observers
focus
attention
```

frequency count:

input: stop word cleaned tokens

output:



```
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
================= RESTART: C:\Users\Ashis\Desktop\ir\ir_proj.py ================
DES	1
MOINES	1
Iowa	11
AP	1
Republicans	3
Monday	1
scheduled	1
precinct	2
caucuses	4
Feb	1
weeks	1
earlier	1
planned	1
Democrats	2
11	1
move	1
date	1
The	6
shift	1
aimed	1
fighting	1
threat	1
heavy	1
attention	4
presidential	4
campaign	2
I	1
confident	1
candidates	3
political	1
observers	1
focus	1
resources	1
Republican	2
Chairman	1
Kayne	1
Robinson	3
news	1
```

Processing:

Finding maximum frequency:

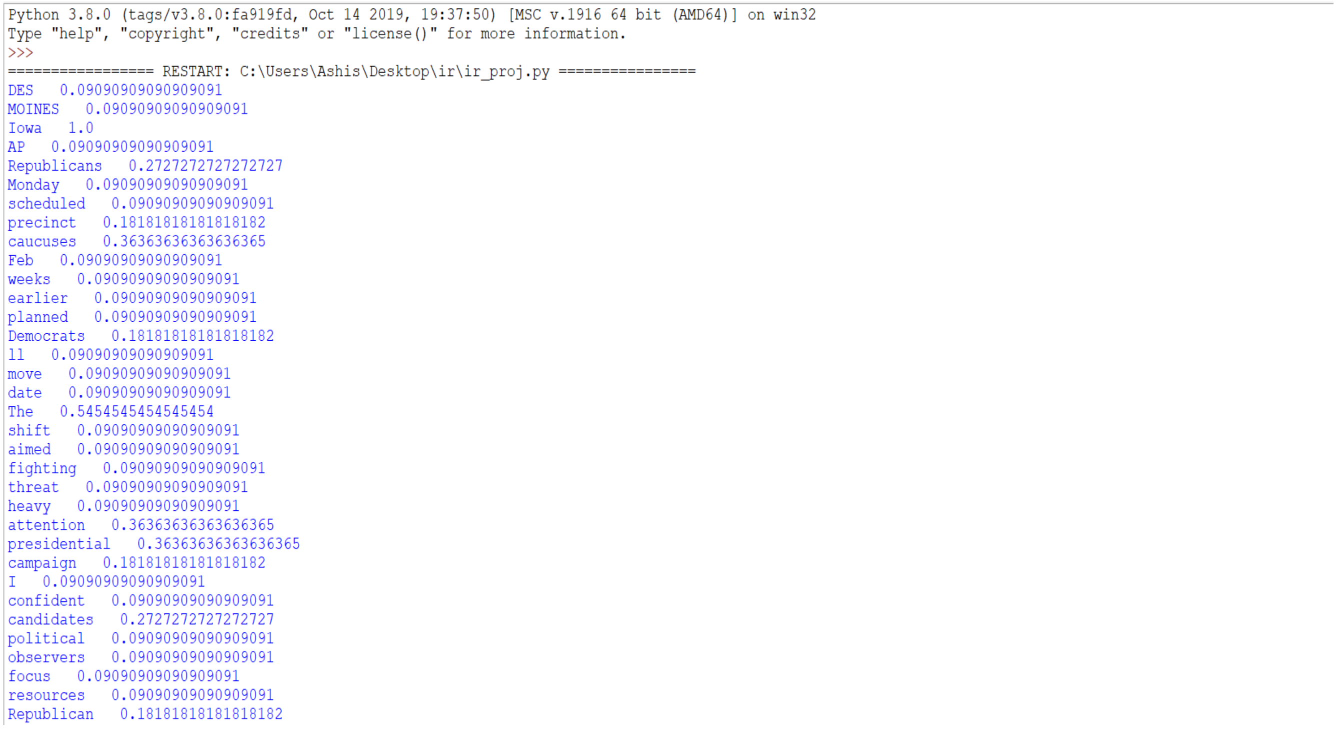
input: frequency of words apart from stopwords

output: 11

Scoring other tokens based on maximum frequency:

input: frequency of words with maximum frequency

output:

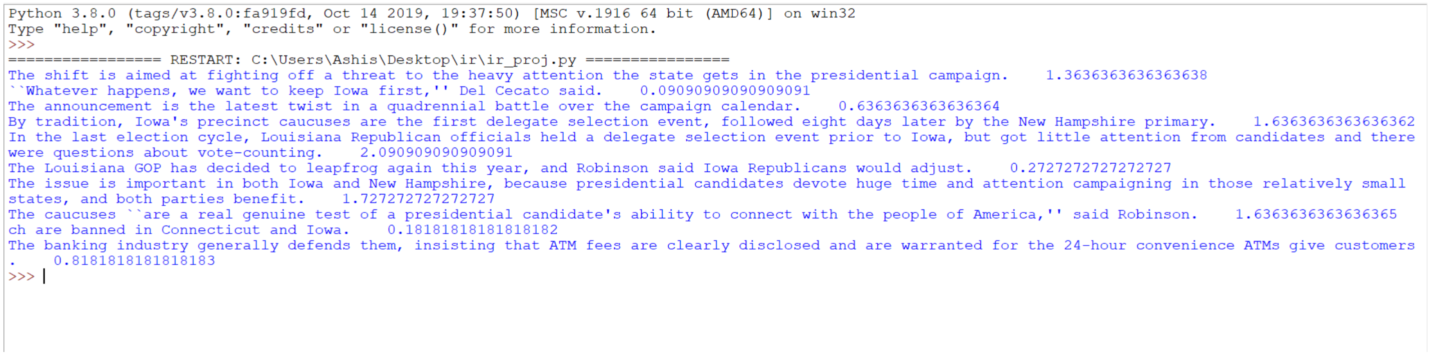


```
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
================ RESTART: C:\Users\Ashis\Desktop\ir\ir_proj.py ================
DES   0.09090909090909091
MOINES   0.09090909090909091
Iowa   1.0
AP   0.09090909090909091
Republicans   0.2727272727272727
Monday   0.09090909090909091
scheduled   0.09090909090909091
precinct   0.18181818181818182
caucuses   0.36363636363636365
Feb   0.09090909090909091
weeks   0.09090909090909091
earlier   0.09090909090909091
planned   0.09090909090909091
Democrats   0.18181818181818182
11   0.09090909090909091
move   0.09090909090909091
date   0.09090909090909091
The   0.5454545454545454
shift   0.09090909090909091
aimed   0.09090909090909091
fighting   0.09090909090909091
threat   0.09090909090909091
heavy   0.09090909090909091
attention   0.36363636363636365
presidential   0.36363636363636365
campaign   0.18181818181818182
I   0.09090909090909091
confident   0.09090909090909091
candidates   0.2727272727272727
political   0.09090909090909091
observers   0.09090909090909091
focus   0.09090909090909091
resources   0.09090909090909091
Republican   0.18181818181818182
```

Scoring sentences:

input: score of each token

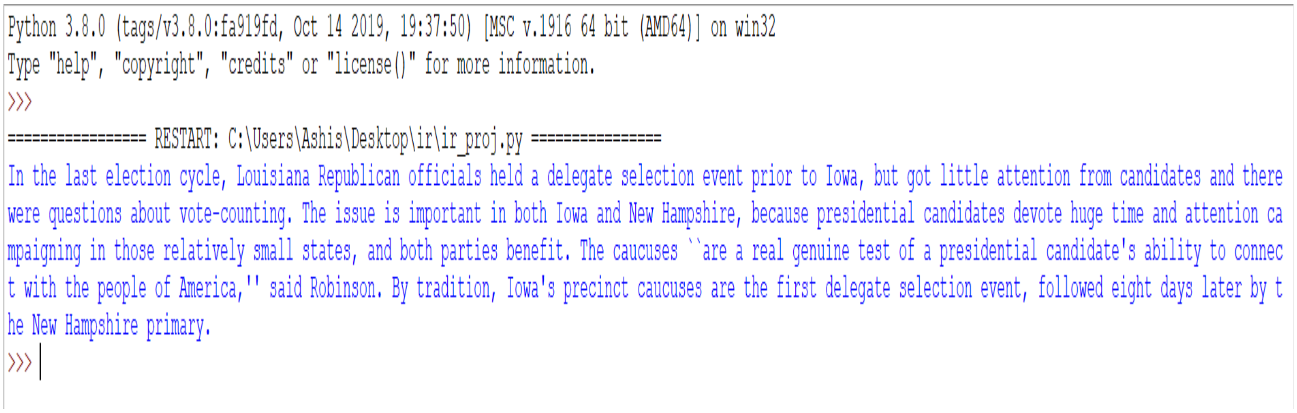
output:



```
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
================ RESTART: C:\Users\Ashis\Desktop\ir\ir_proj.py ================
The shift is aimed at fighting off a threat to the heavy attention the state gets in the presidential campaign.    1.3636363636363638
``Whatever happens, we want to keep Iowa first,'' Del Cecato said.    0.09090909090909091
The announcement is the latest twist in a quadrennial battle over the campaign calendar.    0.6363636363636364
By tradition, Iowa's precinct caucuses are the first delegate selection event, followed eight days later by the New Hampshire primary.    1.6363636363636362
In the last election cycle, Louisiana Republican officials held a delegate selection event prior to Iowa, but got little attention from candidates and there
were questions about vote-counting.    2.090909090909091
The Louisiana GOP has decided to leapfrog again this year, and Robinson said Iowa Republicans would adjust.    0.2727272727272727
The issue is important in both Iowa and New Hampshire, because presidential candidates devote huge time and attention campaigning in those relatively small
states, and both parties benefit.    1.727272727272727
The caucuses ``are a real genuine test of a presidential candidate's ability to connect with the people of America,'' said Robinson.    1.6363636363636365
ch are banned in Connecticut and Iowa.    0.18181818181818182
The banking industry generally defends them, insisting that ATM fees are clearly disclosed and are warranted for the 24-hour convenience ATMs give customers
.    0.8181818181818183
>>> |
```

Generate Summary:

input: sentences along with scores
output:



```
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> 
================ RESTART: C:\Users\Ashis\Desktop\ir\ir_proj.py ===============
In the last election cycle, Louisiana Republican officials held a delegate selection event prior to Iowa, but got little attention from candidates and there
were questions about vote-counting. The issue is important in both Iowa and New Hampshire, because presidential candidates devote huge time and attention ca
mpaigning in those relatively small states, and both parties benefit. The caucuses ``are a real genuine test of a presidential candidate's ability to connec
t with the people of America,'' said Robinson. By tradition, Iowa's precinct caucuses are the first delegate selection event, followed eight days later by t
he New Hampshire primary.
>>> 
```

Experimental setup

System configuration:

→ 8 GB, DDR-4 RAM
→ 1TB HDD
→ core i5-7th gen , 2.5 GHz processor
→ Python 3.8
→ Operating System: Windows 10 Home, 64 bit