VISIT THE DOME

PreProcessing (Thang Nguyen)

Tasks

- Task 1: Collect two new dataset from original corpus
  - Dataset 1: rows having “main topic” is “Oral Questions”
  - Dataset 2: rows having “main tooic” is “Statements By Members”
- Task 2: Remove duplicate rows in the two datasets if any
- Task 3: Report 10 most occurring words, 10 most occurring two sequential words, 10 most occurring three sequential words in both dataset

Task 1

- Read through all row in all .csv file (13261 files)
- Check the “main topic” field (column 7/15)
- If the “main topic” is “Oral Questions”, add to Dataset 1
- If the “main topic” is “Statements By Members”, add to Dataset 2

Task 1

- Oral Question

basepk	hid	speechdate	pid	opid	speakeroldnam	speakerpos	maintopic	subtopic	subsubtopic	speechtext
3958728	ca.proc.d.1994-0...	25/02/1994	9ec9b873-a9...	116,0	Mr. Michel G...		Oral Questions	Francophon...		Mr. Speaker, "the cream of the crop of francophone officers in the armed forces will be assimilated if young
3958729	ca.proc.d.1994-0...	25/02/1994	99fe4f4b-b5b...	4464,0	Hon. David ...		Oral Questions	Francophon...		Mr. Speaker, in the interview I gave to a Canadian Press reporter two days ago, I did not say that I was sick of
3958730	ca.proc.d.1994-0...	25/02/1994	9ec9b873-a9...	116,0	Mr. Michel G...		Oral Questions	Francophon...		Mr. Speaker, I would like to inform the minister that if he is sick and tired now of Bloc Quebecois members, he is in for
3958731	ca.proc.d.1994-0...	25/02/1994	99fe4f4b-b5b...	4464,0	Hon. David ...		Oral Questions	Francophon...		Mr. Speaker, the hon. member is discounting the changes that have taken place in Canada outside Quebec since the
3958732	ca.proc.d.1994-0...	25/02/1994	9ec9b873-a9...	116,0	Mr. Michel G...		Oral Questions	Francophon...		Mr. Speaker, once again I feel I must reassure the minister. I did visit Kingston, Ontario, and the military college as well
3958733	ca.proc.d.1994-0...	25/02/1994	99fe4f4b-b5b...	4464,0	Hon. David ...		Oral Questions	Francophon...		Mr. Speaker, the hon. member is quoting one former officer. Perhaps after this one-week recess, I will have found 20 or
3958734	ca.proc.d.1994-0...	25/02/1994	ed3d4160-8a...	94,0	Mr. Gilles Du...		Oral Questions	Francophon...		Mr. Speaker, my question is for the Minister of National

- Statements By Members

basepk	hid	speechdate	pid	opid	speakeroldname	speakerpos	maintopic	subtopic	subsubtopic	speechtext
3952278	ca.proc.d....	19/01/19...	e62a3516-e718-...	7553,0	Mr. Leonard Hop...		Statements By Members	House Of C...		Mr. Speaker, I welcome you and all members of the House to a great job
3952279	ca.proc.d....	19/01/19...	a4756115-a90d-...	3849,0	Mr. Stéphane Ber...		Statements By Members	Los Angele...		Mr. Speaker, speaking on my own behalf and on behalf of the Official
3952280	ca.proc.d....	19/01/19...	2ee06d6b-f7f1-4...	5412,0	Mr. Ian McClellan...		Statements By Members	Los Angele...		Mr. Speaker, I rise before you today with a great deal of pride and
3952281	ca.proc.d....	19/01/19...	3c3a67cc-b4f6-4...	4255,0	Mr. Ronald J. Du...		Statements By Members	Internation...		Mr. Speaker, the United Nations has proclaimed 1994 the International
3952282	ca.proc.d....	19/01/19...	cae8968d-ec08-...	3482,0	Mr. Peter Adams ...		Statements By Members	Labotix Aut...		Mr. Speaker, let us start 1994 with an example of Canadians already
3952283	ca.proc.d....	19/01/19...	d8932c2a-9013-...	563,0	Mr. Roger Pomeri...		Statements By Members	Amateur Sp...		Mr. Speaker, as the press revealed last week, in Canadian amateur

Task 2

- Rule of finding and removing duplicated rows
  - Using dataframe.drop_duplicates() in pandas to find out duplicated rows
  - Keep the first rows among all of duplicated rows
- Before removing duplicate rows
  - In “Oral Questions” dataset: 133084 rows
  - In “Statements By Members” dataset: 44114 rows
- After removing duplicate rows
  - In “Oral Questions” dataset: 126318 rows
  - In “Statements By Members” dataset: 43251 rows

Task 3

- Step 1: Build a dictionary for all unique words in each dataset
- Step 2: The key of each element is a word, the value is how many time this word appear in dataset
- Step 3: Sorting this dictionary
- Step 4: Pick up the first 10 elements

Task 3

- Oral Questions

#Line	Column1	Column2
1	mr.	123940
2	speaker	121393
3	minister	90704
4	government	89090
5	canada	49240
6	's	45383
7	canadians	41851
8	member	36431
9	would	33350
10	prime	32746

#Line	Column1	Column2
1	mr speaker	118299
2	prime minister	32500
3	hon member	12923
4	speaker government	6636
5	would like	5949
6	conservative government	5153
7	bill c	4994
8	liberal party	4567
9	minister finance	4042
10	across country	3905

Task 3

- Statements By Member

#Line	Column1	Column2
1	mr.	46603
2	speaker	42498
3	's	39603
4	canada	38918
5	government	29633
6	canadians	22001
7	canadian	21743
8	people	18798
9	today	16373
10	minister	16005

#Line	Column1	Column2
1	mr speaker rise	3249
2	mr speaker today	2608
3	speaker rise today	2277
4	mr speaker last	1608
5	mr speaker would	1354
6	speaker would like	1320
7	mr speaker yesterday	1144
8	mr speaker pleased	902
9	would like congratulate	879
10	would like thank	826

Task 3

- Oral Questions

```
10 most occurring words in Oral Questions
['mr.', 'speaker', 'minister', 'government',
'canada', "'s", 'canadians', 'member', 'would', 'prime']

10 most occurring two sequential words in Oral Questions
['mr speaker', 'prime minister', 'hon member', 'speaker government', 'would like',
'conservative government', 'bill c', 'liberal party', 'minister finance', 'across country']

10 most occurring three sequential words in Oral Questions
['mr speaker government', 'mr speaker minister', 'mr speaker would', 'speaker prime minister', 'mr speaker prime',
 'mr speaker member', 'mr speaker yesterday', 'minister national defence', 'mr speaker said', 'speaker would like']
```

- Statements By Members

```
10 most occurring words in Statements By Members
['mr.', 'speaker', "'s", 'canada', 'government', 'canadians',
'canadian', 'people', 'today', 'minister']

10 most occurring two sequential words in Statements By Members
['mr speaker', 'would like', 'prime minister', 'speaker rise', 'rise today',
'federal government', 'speaker today', 'let us', 'pay tribute', 'liberal government']

10 most occurring three sequential words in Statements By Members
['mr speaker rise', 'mr speaker today', 'speaker rise today', 'mr speaker last', 'mr speaker would',
'speaker would like', 'mr speaker yesterday', 'mr speaker pleased', 'would like congratulate', 'would like thank']
```

Issue

- Thang's result

#Line	Column1	Column2
1	mr.	46603
2	speaker	42498
3	's	39603
4	canada	38918
5	government	29633
6	canadians	22001
7	canadian	21743
8	people	18798
9	today	16373
10	minister	16005

- Ruta's result

	Word	Frequency
0	mr	46662
1	speaker	42512
2	canada	39839
3	government	29796
4	canadians	22074
5	canadian	22037
6	people	18858
7	today	16380
8	minister	16018
9	would	15952

Solution

- Stemming words
  - The process of reducing each word to its root or base
  - Ex: fishing, fished, fisher → fish
  - Reduce the size of vocabulary
  - Focus on the sense or sentiment of a document rather than deeper meaning
  - Using PorterStemmer() function of NLTK

Solution

Before

#Line	Column1	Column2
1	mr.	46603
2	speaker	42498
3	's	39603
4	canada	38918
5	government	29633
6	canadians	22001
7	canadian	21743
8	people	18798
9	today	16373
10	minister	16005

After

#Line	Column1	Column2
1	mr.	46603
2	canadian	43752
3	speaker	42591
4	's	39603
5	canada	38923
6	govern	31415
7	year	25229
8	peopl	19374
9	member	19153
10	commun	18892

The end