Data Preprocessing – Week 2

Visit the Dome

Ruta Bakhda, Thang Nguyen, Vamsi Vaddi

Overview

Goal :

Filter speeches that cover controversial topics

Steps :

Step1 : Remove clichés from the beginning and from the end of the sentences

Step 2 : Reporting ten most occurring subtopics in produced Oral questions and Statement by members

Step 3: Filter the speeches to controversial topics

Step 4 : Report most occurring controversial topics in Oral questions and statement by members

Step 1 : Removing clichés

How : Using tf-idf to find the weight of each token in all documents and removing less important (most common) words from the beginning and end of the speech

term	rank
minister	2347.351179
government	2077.974250
mr	2048.295076
oh	2031.947629
speaker	1975.840223
mr speaker	1966.255920
hon	1687.001108
member	1640.342418
canadians	1401.425980
canada	1322.287222
prime	1293.500025
prime minister	1178.516504
hon member	1151.919454
house	1100.329164
canadian	1028.480405
oh oh	1020.120927
hear	948.531846
order	924.100205
said	814.549490
question	805.352936

term	rank
canada	544.910868
government	531.197929
mr	472.210196
canadians	453.760010
member	452.104779
canadian	436.265668
speaker	424.786030
mr speaker	421.904975
hon	412.302531
hon member	393.282030
people	392.624096
minister	373.980126
today	331.308139
quebec	327.606120
women	322.773637
years	309.241096
community	306.154949
day	303.373948
house	290.746285
world	290.093486

Step 2 : Counting most occurring sub topics

How : Using pandas

Oral Questions	
Ethics	8753
The Environment	7197
National Defence	5092
Taxation	3903
Foreign Affairs	3900
Afghanistan	3493
Health	3181
Points of Order	3111
Public Safety	2996
The Economy	2964

Statement by members	
The Environment	533
Taxation	496
Agriculture	379
Justice	360
The Economy	351
The Budget	323
Health	284
Aboriginal Affairs	261
Human Rights	257
Violence Against Women	246

Step 3 : Finding controversial topics

How : Comparing our data with Wikipedia's controversial issues by counting frequencies of topic occurrence in speech text as well as comparing sub topic with list

Filters : Either controversial topics is sub topic or occurs more than twice in speech text

For Oral Questions,

133089 -> 30864

For Statement by Members,

44115 -> 25299

Step 4 : Reporting most occurring controversial

How : Pandas to find most occurring controversial topics

Oral Questions	
Taxation	2743
Ethics	1787
The Environment	1145
Status of Women	1068
Health	954
The Economy	875
Aboriginal Affairs	744
Foreign Affairs	736
Justice	731
Employment Insurance	635

Statement by Members	
Taxation	597
Violence Against Women	345
Agriculture	308
Status of Women	299
The Economy	260
The Budget	231
Justice	209
International Women's Day	207
Health	205
The Environment	172

Controversial topics by category

- 8760 – Statement by members
- 1066 – Oral questions

Suggestions to improve accuracy

- Following can be integrated to improve the accuracy of finding controversial topics
  - Sentence Similarity (Semantic Similarity)
    - Is the speech text related to the debate?
  - Sentiment Analysis
    - Checking the polarity of the speech text