Ted Talks Content Analysis

Which is more persuasive?

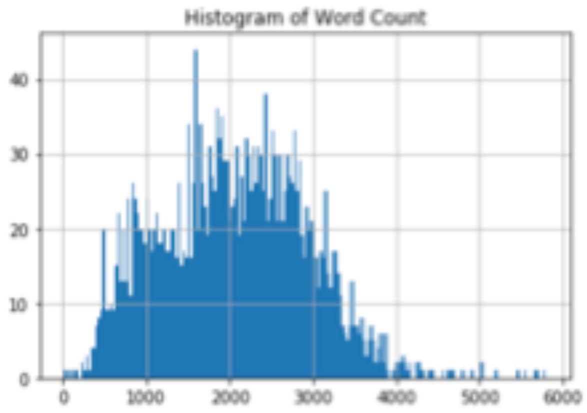
Data

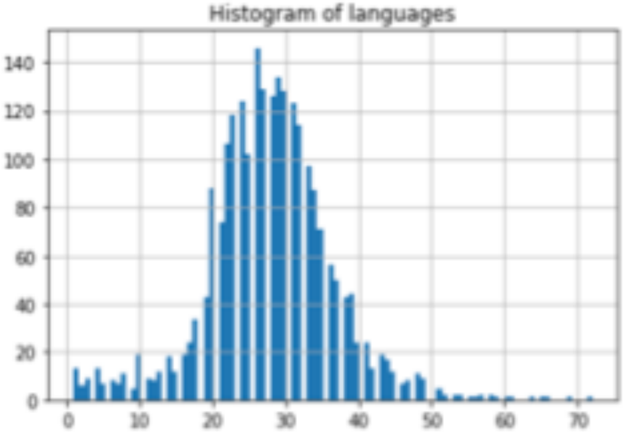
- Word count
- Language available
- Talk Transcript

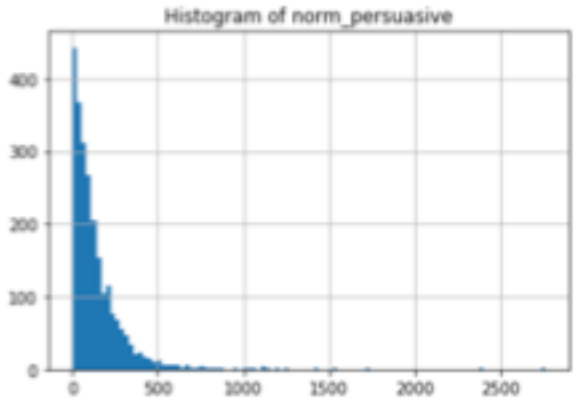
--Predict object:

- Persuasive score

No Missing Value

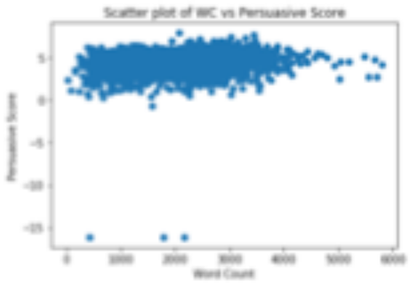


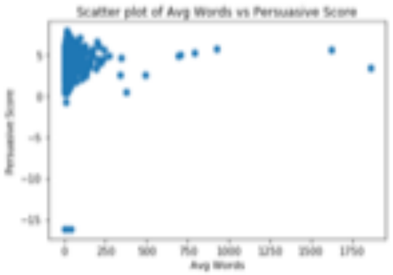




Feature Engineering

- Log transformation – persuasive score
- Sentence Tokenizing
  - Compute the number of sentences
  - Compute the average length of sentences



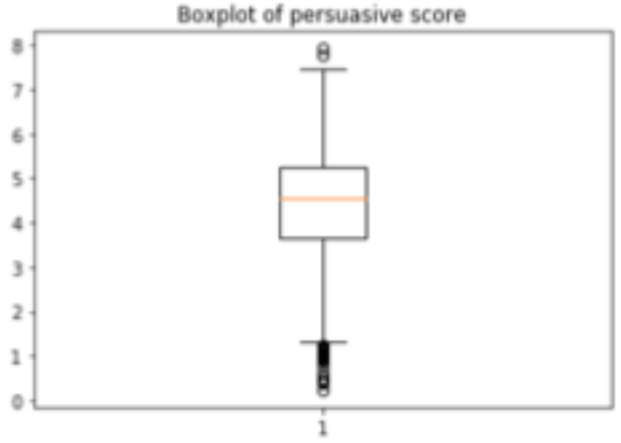


Tokenizing and Frequency counts

- Using kmeans to classify words into 80 categories;
- Count the frequency of each categories in every transcript of talk;
- Build new dataframe for modeling

Labeling

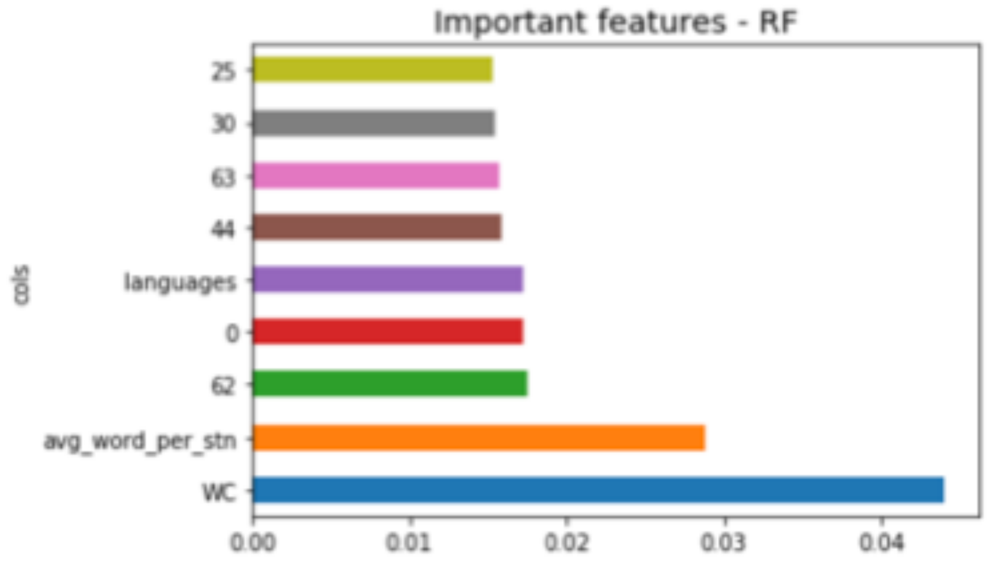
- Persuasive ： 1 if score > 3.5
- Not Persuasive : 0 if score < 3.5



Model comparing

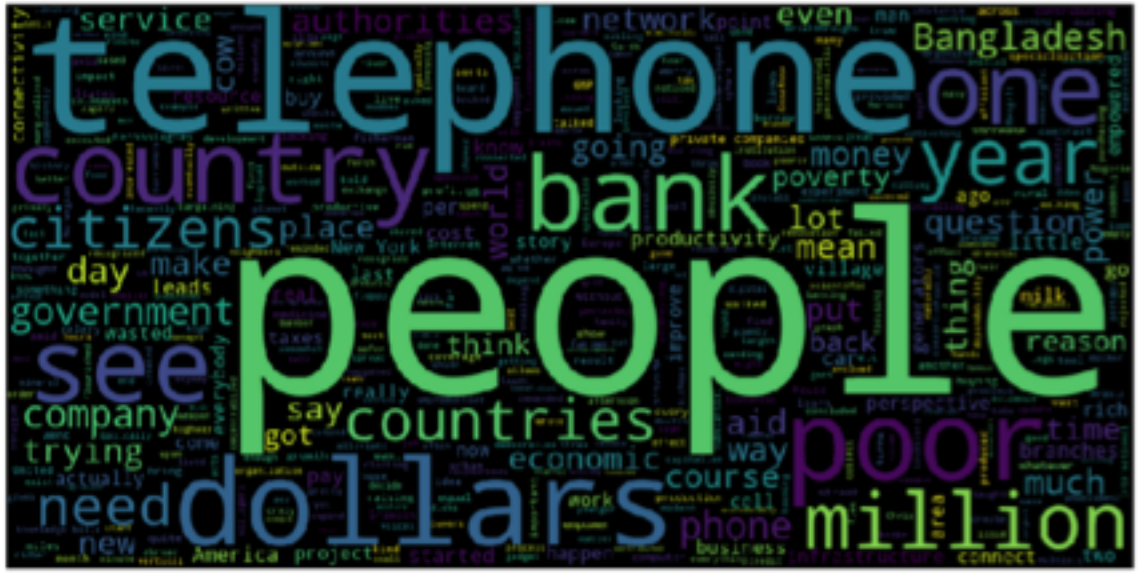
Accuracy	With stop words	Without stop words	With stop words & under- sampling	With stop words & over- sampling
Random Forest	0.784	0.782	0.782	0.778
Logistics Regression	0.762	0.762	0.775	0.684
SVC	0.784	0.784	0.784	0.773

Importance Features



- Length of the talk
- some groups of specific words

- An example of group of words


telephone
people
country
bank
dollars
poor
million
see
citizens
countries
government
company
year
one
Bangladesh
service
authorities
network
even
money
poverty
going
place
day
make
need
trying
question
mean
lot
reason
economic
course
phone
aid
way
much
time
think
say
got
new
America
project
business
world
productivity
little
power
milk
put
back
thing
village
story
cost
leads
wasted
really
actually
rich
work
cell
connect