NATIONAL UNIVERSITY OF SINGAPORE

BT5152: Decision Making Technology for Business | Assignment 4

Swetha Narayanan – A0074604J

Q1 Data Preprocessing of TED Talks transcript data

Steps Carried out for preprocessing:

1. **Encoding**
   - Converted transcript to UTF-8 encoded data
2. **Unique Data**
   - Removed 3 duplicate rows in transcripts
3. **Normalization**
   - Converted transcript to all lower case
   - Removed Emojis, special characters, numbers, punctuation
   - Stripped Unnecessary whitespace
4. **Stemming**
5. **Stopwords removal**

Q2 Determine the kind of emotions associated with the talks :

Approach:

- We obtained the following emotion words from the General Inquirer Dictionary : Happy , Sad, Feel ,Enlightenment, Excitement, Motivation, Virtue, Vice
- We use the pre-processed dataset to get a documentTermMatrix using term frequency
- We remove terms with very low frequencies
- Using tm_term_score, we compare the documentTermMatrix with the dictionary words to obtain the emotions associated with each talk

Interpretation from Heat map in HTML File

- In general, there are more positive emotions than negative emotions in the talks
- Emotions of virtue are the highest followed by motivation and enlightenment
- This correlates with the general theme of ted talks that usually promote feelings of confidence, motivation and positive vibe

Q3 Multi Class text classification and prediction of the ratings of the talks

Approach:

- We first process the ratings column to get the rating with highest count for each talk (Main_Rating). We get a documentTermMatrix using TF-IDF and remove columns with near zero variance
- Using Random Forest and hyper parameter tuning, we predict the rating for each talk
- We obtain the following Macro F1 and Micro F1 metrics on test data – F1 is a weighted measure of precision and recall

- Micro F1 : We sum up the individual TP,FP,FN,TN of the system for different sets and the apply them to get micro F1
- Macro F1: We take the average of the precision and recall on different sets
- In our case, Micro F1 is better than Macro F1, which suggest that the class with larger labels is better classified than the one with smaller labels

Performance Metrics

Macro Precision	0.664
Macro Recall	0.131
Macro F1	0.416
Micro F1	0.522

Commonly Misclassified Ratings

rating	**misclass_rate**
Inspiring	0.335
Informative	0.278
Fascinating	0.093

<u>Q4 Topic Modelling to find 10 related talks given a specific talk</u>

Approach:

- Using TF, we first converted the corpus to a documentTermMatrix
- We split the data set into training and test
- We used Gibbs Sampling and LDA topic modelling to perform topic modelling
- Quantitative : Using perplexity scoring on test data, number of topics of 10 to 200 were considered to find the optimal number of topics
- Qualitiative : Topic set of 200 gave a lower perplexity score, but keeping in mind the principle of parsimony , the fact that ted_talks generally fall into a topic set of <100, and the rate of decrease of perplexity decreases as number of topics increase, an optimal topic score of 100 was decided for this model
- Using the optimal topic model, pertopic_perterm_probability and pertalk_pertopic_probability was obtained
- pertalk_pertopic_probability was used to create a topicSimilarityMatrix
- This topicSimilarityMatrix is then used to find 10 related talks given a specific talk

Improvements:

- We have TF in this model. But TF_IDF is a better metric because we can improve the topics by reducing the weight of words which are present in all the documents.
- Improve LDA by using additional GibbsControl parameters like burnin rate

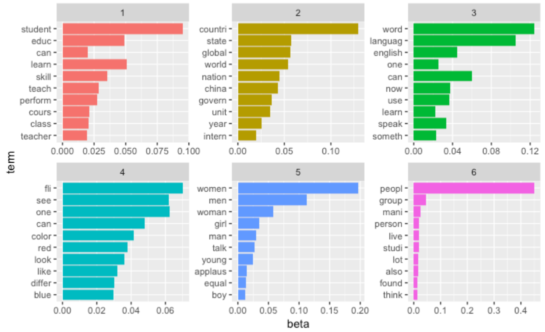
Inference

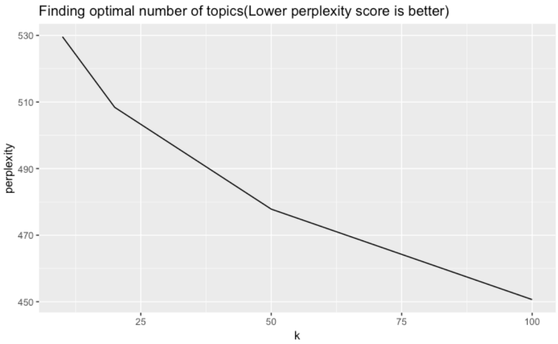
- We can infer topic 1 is about Education, topic 2 is about Government Policy etc
- Document 1 is 7.5% topic 94 (laughter) and 6.9% topic 18 (people)
- Top Related Topics on Training Set

Topic Name : Do schools kill creativity?	
I got 99 problems ... palsy is just one	0.76416727745029,
Every kid needs a champion	0.70169529694792,
Bring on the learning revolution!	0.681063041975946,
How to escape education's death valley	0.678175987468342,
Learning from dirty jobs	0.658497325495313,
The loves and lies of fireflies	0.655108518660787,
Art made of the air we breathe	0.643173752900327,
Three myths about corruption	0.636030610713695,

- Top Related Topics on Test Set

Topic Name : How art gives shape to cultural change	
Fun, fierce and fantastical African art	0.766375287937384,
Playing with space and light	0.693595099810349,
Why the live arts matter	0.671906689046494,
Gaming for understanding	0.568447869392222,
10 young Indian artists to watch	0.56223102433909,
Do the green thing	0.447780821785736,
Art of substance and absence	0.444602929840541,
Life science in prison	0.408797022054262,
A multimedia theatrical adventure	0.40369624884952,




Finding optimal number of topics(Lower perplexity score is better)