Topic Modelling Approach

Preprocess

1. Remove Punctuation, Number

2. Word Segmentation

* Jieba
* Validity Check

3. remove single word

* Remove single characters

2. Stop Word Removal

* Filter Stop words

3. POS tag word removal (optional)

* Try different approaches
* Remove Noun or NE
* Remove ---

4. Word Frequency Check

* Plot Histogram
* Show Example Words
* Word Frequency Filtering
* Document Frequency Filtering

5. High/Low frequency word removal

* Try different threshold for high and low frequency words

6. Word Embedding

* Try Three embeddings
* BOW
* TF-IDF
* High dimensional word-embedding approach