NLP Project

**Text Preprocess**

I used python to load the dataset, given the characteristic of the files. I extracted the ‘name’ and ‘sentence’ from each file and combine every document together. Each row from dataframe is one xml file.

Then the pipeline I created for pre-processing text data include the following steps:

1. Lower case the word.

2. Remove punctuations.

3. Remove stop words.

4. Remove non-English words.

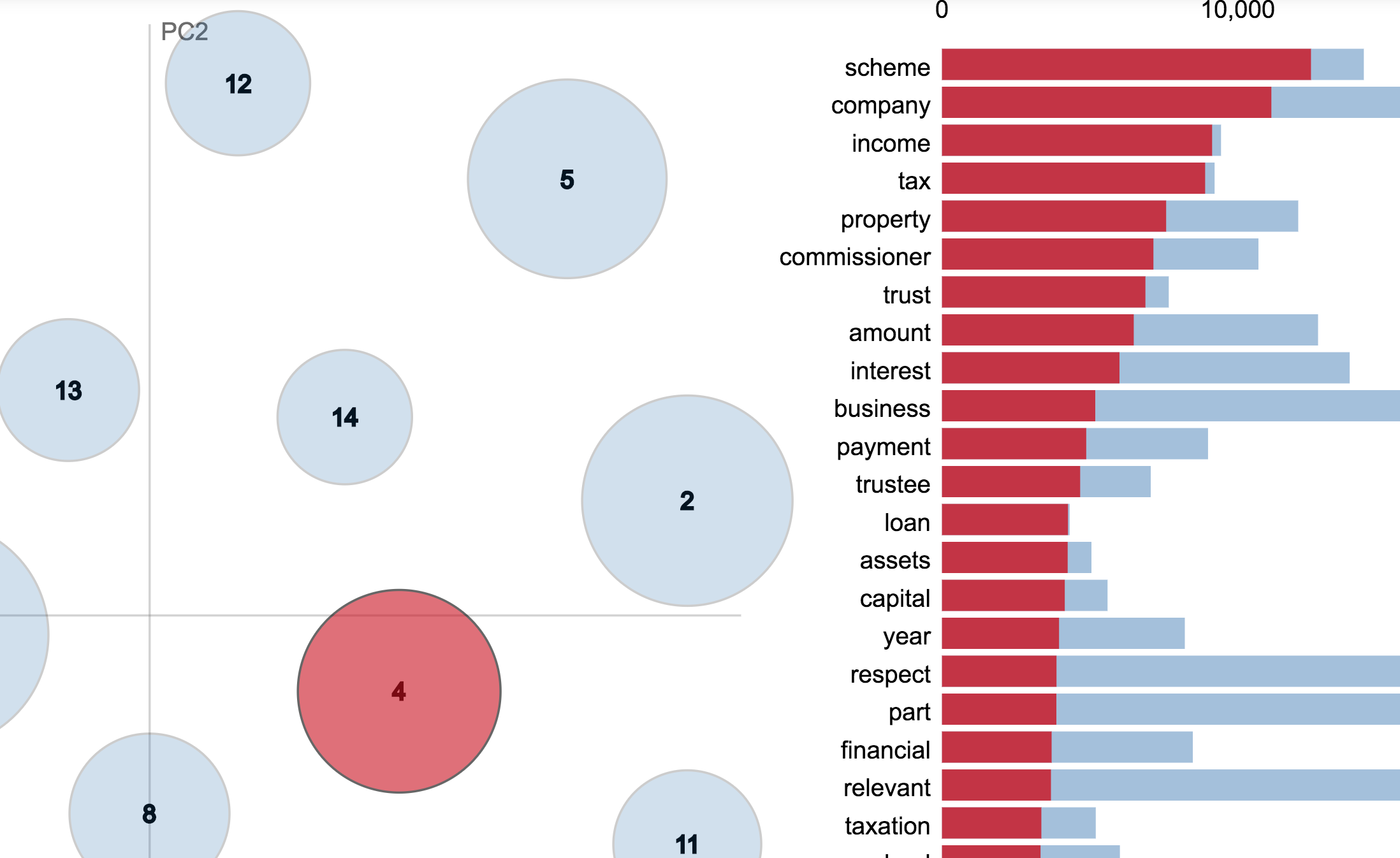
5. Remove the length of words which has low frequency. (length less than 2).

**NLP Process**

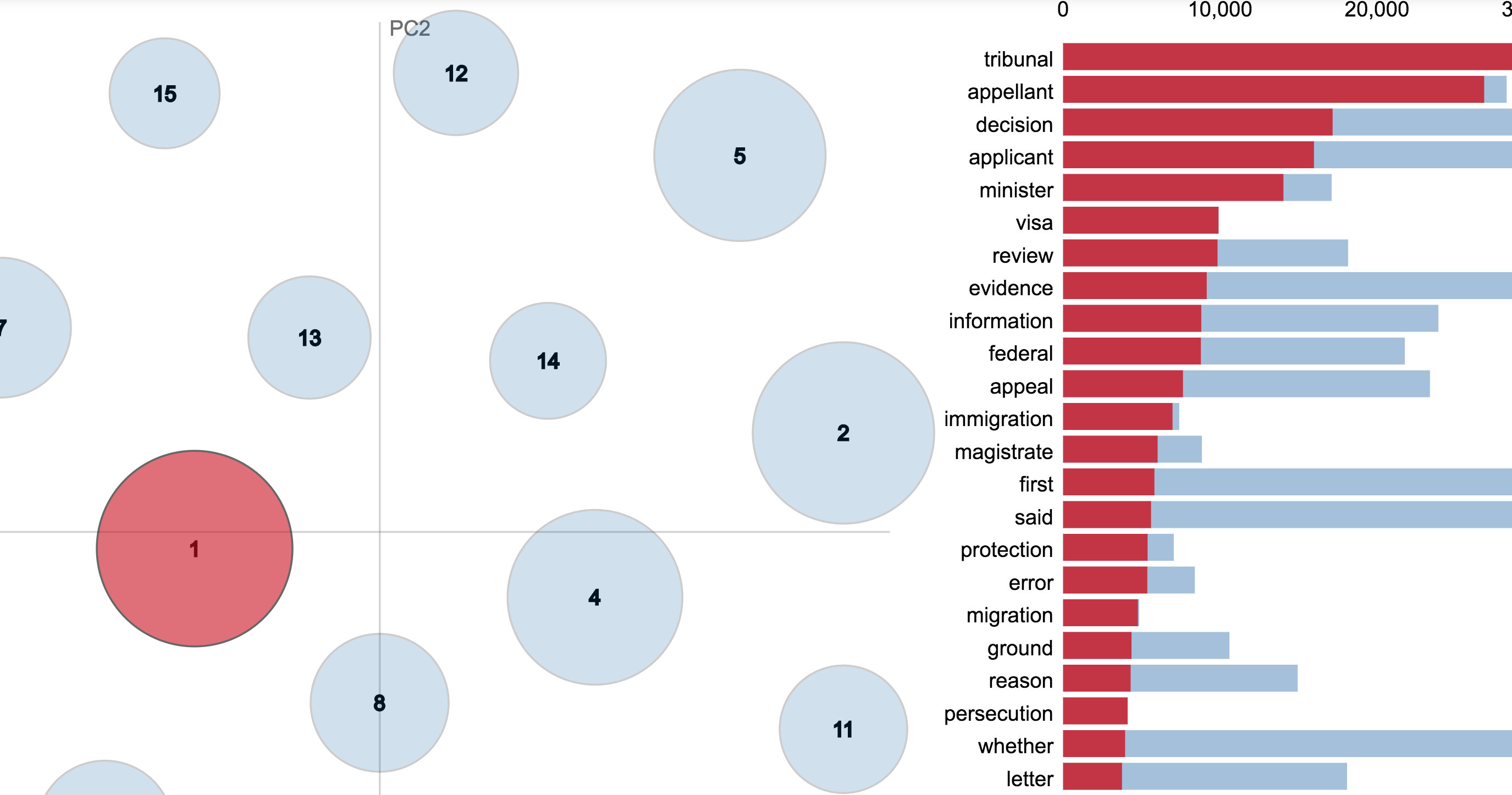
We firstly tokenized the preprocessed data in order to do the LDA model which needs the tokenized data first. Then I created dictionary on the tokenized data. I also used doc2bow to convert document into the bag-of-words (BoW) format for the preparation of the LDA model. Latent Dirichlet Allocation allows observations to be explained by unobserved groups that can explain why some of the data are similar, which can be used as topic modeling. Number of topics also has been set as 15.

A Bokeh plot was created for finding the word which is similar to each other.

A dashboard by using the pyLDAvis was also created for better seeing the result.



For example, here’s the 4th topic. The words have ‘company’, ‘tax’, ‘trust’, ‘assets’, ’capital’ and other words which seem like the topic related to financial domain.



For the topic 1, the words have ‘tribunal’, ‘visa’, ‘federal’, ‘protection’, ‘immigration’ and other words which seem like they are related to immigration cases.