NLP -Lab 03

Aimee Haas, Loic Sine, Rick Beaudet

February 11, 2023

1 Introduction

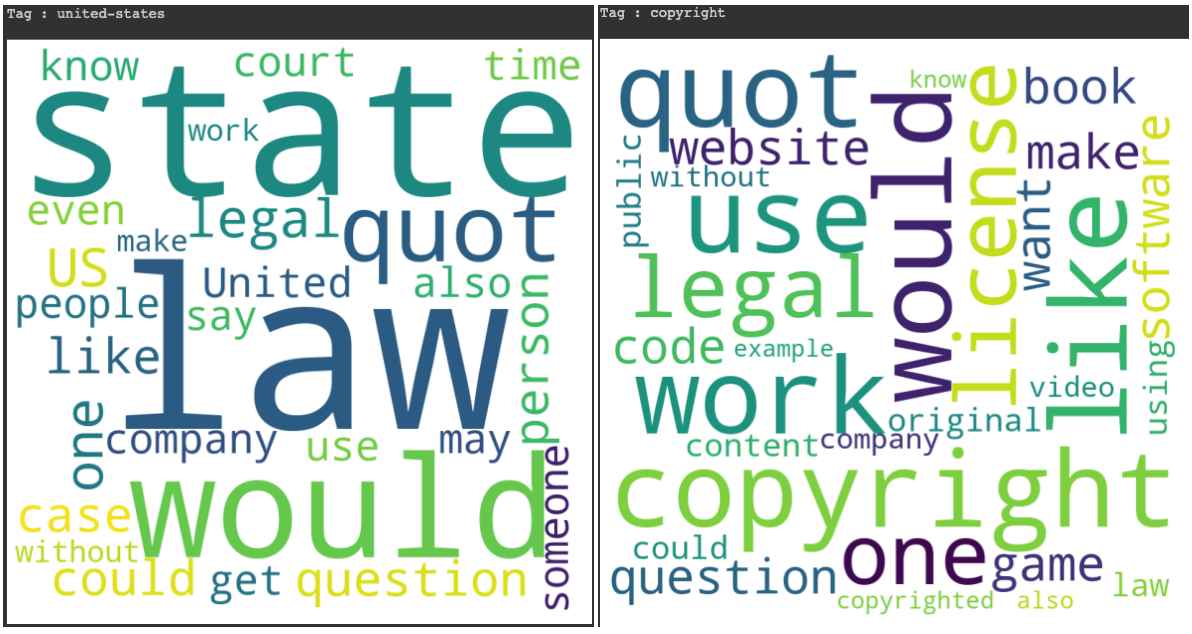
In this lab we worked with python NLTK library to parse posts from law.stackexchange.com. The notebook code for our group can be found at github repo. In order to run the code you will need to have the posts.xml file from the law.stackexchange.com zip file, as well as the post_parser_record.py file and the Post.py file. We noted recently that there may have been a change made to the post_parser_record.py file that exists in bright space. Re downloaded today and there is a encoding parameter set on line 14 that they python interpreter gets very upset about. After removing the parameter, it worked fine.

2 Step 2

2.1 Creating word clouds for top 30 common tokens after removing stopwords

Since we were tasked with gathering the top 5 most frequent tags amongst all of the posts, we interpreted that we needed to create a word cloud of top 30 tokens for each of the 5 tags so we could compare.

2.2 The Word Clouds for top 30 tokens after stopwords were removed



Our approach of creating 5 seperate word clouds allowed for us to see what 30 tokens were most frequently found for each of the top 5 tags. When examining each of the tags, the tokens displayed in the word cloud are easily associated with the tag. For 'united-states' we end up tokens such as: US, United, legal, court, etc. Between the tag being associated with law.stackexchange.com and 'united-states' these words make sense that they would frequently appear. For 'Copyright', we end with tokens like: legal, work, code, game, copyrighted, law, software, book, license, etc. For 'united-kingdom' we end up with tokens such as: UK, law, contract, work, pay, case, etc. For 'contract-law' we end up with tokens such as: pay, service, agreement, work, sign, signed, party, law, person, clause, etc. Finally for 'criminal-law' we end up with tokens such as: criminal, police, evidence, question, illegal, time, court, case, etc.

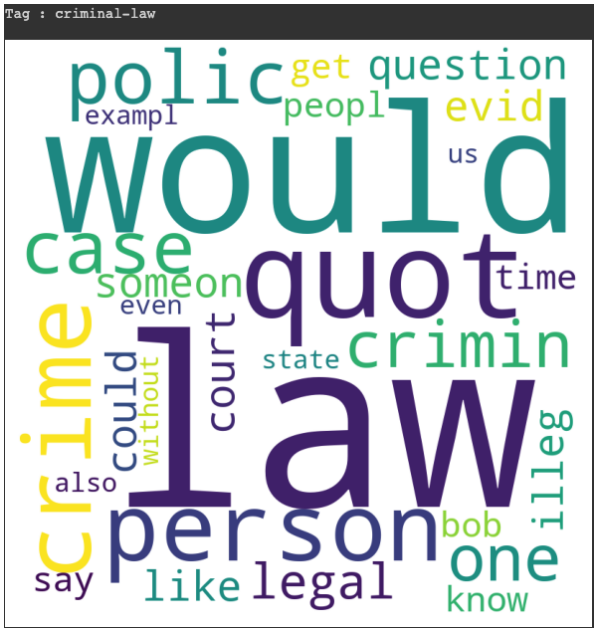
These tokens generated for each tag make sense based on what the tag is. We would say that the law specific tags such as 'contract-law', 'criminal-law' and 'copyright' seem to be more tightly associated than the other two. But this makes sense because they are more topic focused tags, where as 'united-states' and 'united-kingdom' are more geographical focused tags.

3 Step 3

3.1 Generating our top 30 tokens but this time running them through the NLTK Porter Stemmer before

Here we repeated the same process that we went through before but when generating our top 30 most frequent tokens we ran the tokens throught the NLTK Porter Stemmer before determining our lists. Stemming is the process of producing morphological variants of a root/base word.

3.2 The Word Clouds for top 30 tokens after stop words were removed and Porter stemmer was applied



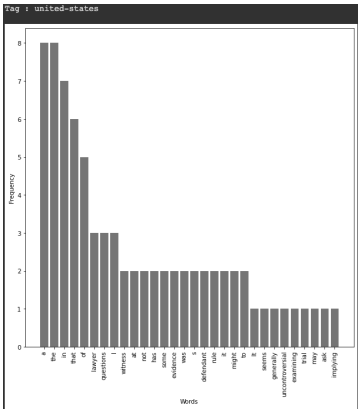
Based on what we know about how a Porter stemmer works, the results reflect what was expected to happen. Whe comparing the word clouds associated with a particular tag, they appear similar but not congruent. For instance the word cloud associated with the tag: 'criminal-law' has some of the same words such as: case, law, crime and legal. However there are some words that appear to be stemmed but not neccesarily in a relevant way, for instance police becomes polic, people becomes peopl and example becomes exampl. There are more interesting excutions of the stemmer however, illegal becomes illeg, and criminal becomes crimin. It is interesting to see how the stemmer can determine what does and doesnt get stemmed and makes it seem like stemming may have its place within NLP pre processing, but might not be our go to.

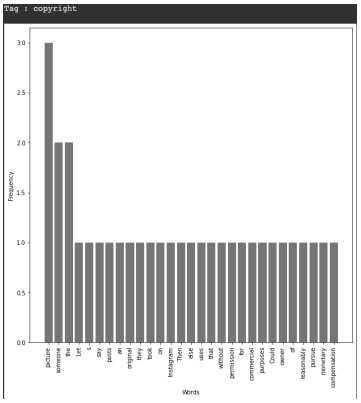
4 Step 4

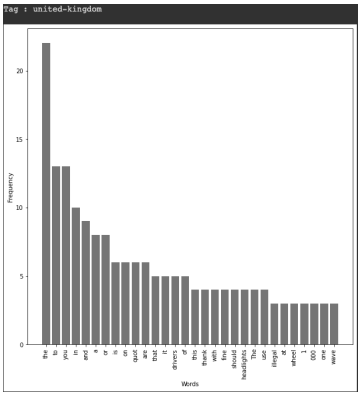
4.1 Generate a word distribution diagram

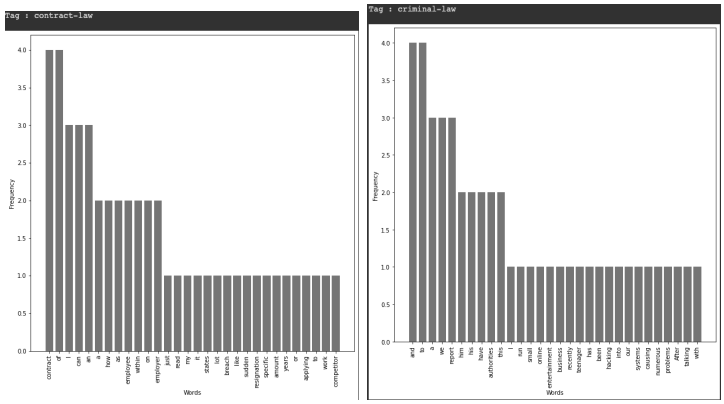
Here we repeated the same process of step 2, but this time we didn't remove the stop words before generating our top 30 most frequent tokens. We then needed to generate a diagram displaying (Rank Vs. Probability).

4.2 The word distribution diagrams generated from each of our 5 most frequent tags 30 most frequent tokens









We noticed based on the 5 graphs that were generated from our top 5 tags, 30 most frequent tokens amongst all posts in law.stackexchange.com, that they all follow a similar trend. This trend that they follow is very similar to that of perfect Zipf curve, and therefore we determined that the Rank Vs. Probability word distribution of each of the 30 tokens for each tag follows Zipf's Law.