Lab 3 - Shea Durgin, Aleksandra Milinovic, Caleb Wentworth
Github - https://github.com/sheaDurgin/NLP23/tree/main/Lab3

Step 2:
The first commonality I found between the 5 wordclouds was how “would”, would always show up big and in the center. Since we are parsing a question and answer forum website, it makes sense that a word such as “would”, would show up so frequently. For example, a common way of asking a question is using would at the start of the sentence, “would x cause y to happen?”. Also the tokens in the common-tags are found usually at the upper echelon of the top-30 tokens as well.

Step 3:
The porter stemmer didn’t stop the reign of “would”. The biggest difference that the porter stemmer caused between these wordclouds and the first ones, is that it stemmed some of the common words! Words that would end in e, es, ed etc. sometimes appeared on the wordclouds that weren’t in the first versions. For example, “requir” showed up on the wordcloud for the tag “united-states”, none of it’s unstemmed counterparts were on the original wordcloud.

Step 4:
All 5 of the zipfs law word distribution graphs had a similar look to each other. What we are looking for to verify that they follow zipfs law is whether or not their slopes look to be around -1. While none of the slopes are a perfectly straight line, they pass the eye test for a slope around -1.