Lab 3 - Shea Durgin, Aleksandra Milinovic, Caleb Wentworth

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:
A zipfs law diagram usually appears to have a line that starts at the top left and has a fast descent into what quickly becomes a plateau. This happens because you plot the tokens based on their frequency (y-axis, how many occurrences) and their rank (x-axis, most frequent is 1st rank). All 5 common-tags tokens followed zipf’s law even though their tokens and frequencies were unique.