CS378 assignment0

Qinxin Wang

January 23, 2020

1 Language Basics / Coding Warmup

Q6

a) The 10 most frequent words in the dataset using whitespace tokenization are : (’the’, 26423), (’of’, 13072), (’to’, 11264), (’a’, 10447), (’and’, 9533), (’in’, 8677), (’that’, 5778), (’for’, 4089), (’is’, 3664), (’Mr.’, 3349)

These words are mostly articles, prepositions, conjunctions and so on. They do not really have a meaning, but the occur everywhere in the articles. The most frequent word have twice the count of the second most frequent word. From the second to the ninth most common word, the word count decreases by around 1000 to 2000 for every word.

b) The 10 most frequent words in the dataset using smarter(nltk) tokenization are : (’,’, 27915), (’the’, 26491), (’.’, 18649), (’of’, 13084), (’to’, 11292), (’a’, 10526), (’and’, 9623), (’in’, 8741), (’that’, 5943), (”””, 5026)

These words are almost the same as the most frequent words using whitespace tokenization, but there are two major differences: 1) their word count increases slightly; 2) punctuation are also counted, and their count is high. One reason is that by whitespace tokenization, we cannot split words and punctuation, so they are connected and counted as different words.

c) 1) If we use whitespace tokenization, we many count ”great.”, ”great!”, and ”great?” as three different words, which may increase the dimension of bag-of-words vectors, and their weights are calculated separately. But in fact they are the same word.

2) We can split abbreviations like ”wasn’t” into ”was” and ”n’t”, so the computer can find the connection between ”wasn’t” and ”was” instead of considering them as two unrelated words.

Q7

a) See Figure 1.

b) According to the plot, the relationship between the most common word and the second most common word does not hold Zipf’s Law. From the second most common word to the following ones, we can see the line is almost straight, so Zipf’s Law holds in general. But there are some exceptions, for example, from 5th to 8th most common word, as their count are greater than what Zipf’s Law expected.

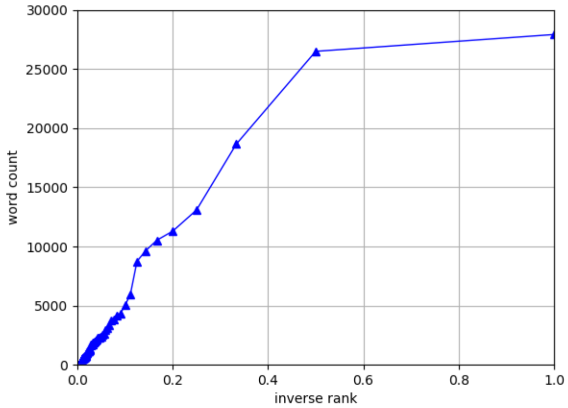

Figure 1: Inverse rank vs. Word count

c) ”Mr.”, ”percent”, ”Dole”.

1) Mr.: In written English, Mr. should not be the tenth most frequent word. People may not use so many names when writing. Also, ”Mrs.” and ”Miss” are not even in top 100.

In news articles, however, writers would mention some important people frequently like officers, spokesmen and entrepreneurs (and a large proportion of them are male), so they use ”Mr.” a lot.

2) percent: ”Percent” is used when describing a ratio or proportion, and news articles will often use it to give precise statistics in fields like medical, population, agriculture, etc.

3) Dole: Actually as a foreigner I don’t even know what it is. I googled it and find out that it’s a food company, and also the name of a U.S. Senator. Maybe it is because news regarding the influential company and the famous senator add up in a period, leading to the result.