CS 646 - IR

P1

Patrick Verga

September 24, 2014

1 Implementation Choices

I parsed the raw xml using the BeautifulSoup library. I first broke each book into pages and then represented the page as all text within a <line> tag, not including subtags of line which held metadata. I then removed punctuation, converted all text to lowercase, and finally split the page on whitespace.

2 Runtimes

Dataset	Runtime (H : MM : S)
tiny	0 : 00 : 9.95
small	0 : 03 : 21.5
medium	0 : 58 : 27.75
big	2 : 34 : 24.57

3 Differences in Dataset Statistics

Even though big was about 30x larger than small, the top words and the probabilities of those words are roughly equivalent. The same is true of medium. The number of unique words does not increase as drastically as the number of total words between the data sets, as shown later in Figure 6.1.

Having the larger data does however give better special word associations. For example, small has spurious previous word statistics such as handsome church which do not occur in the larger datasets.

4 Implications For Designing an IR System

These results show that it does not take an extremely large amount of data to generate relavent statistics about word frequencies. This also demonstrates the importance of things like stop word removal. This also shows the benefit of using idf term weighting for similarity calculations.

5 Zipf's Law

Zipf's law has to do with occurances of a word relative to its rank. Zipf's law as to do with occurances of a word relative to its rank. The bulk of the total words is contained in the few top most frequently occuring words. There is also a drastic drop off from rank 1 word to rank 2 to rank 3 etc.
Zipf's law clearly holds for this data set as shown in 5.1

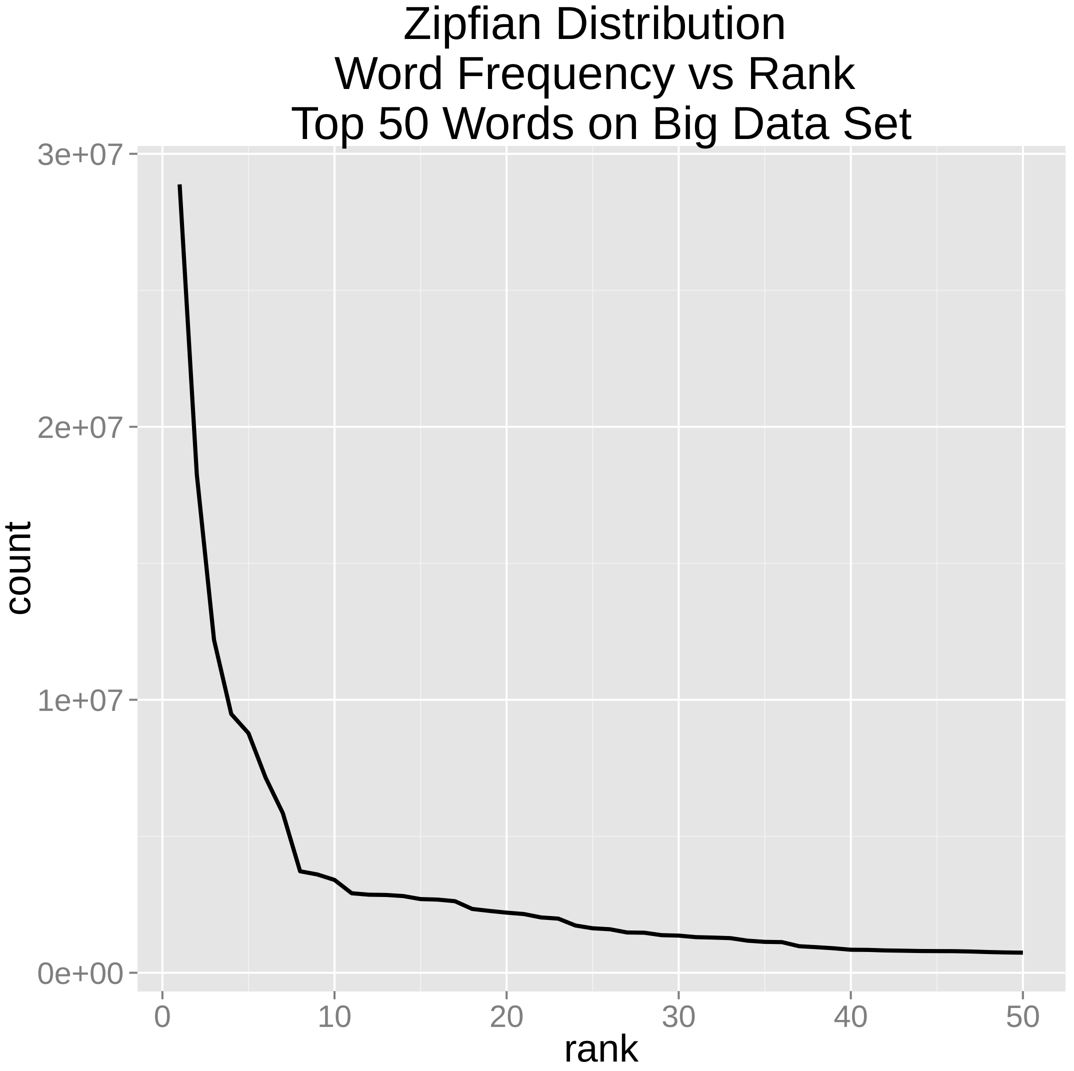

Figure 5.1: Here we plot the top 50 most occuring words on the big data set by their count.

6 HEAP'S LAW

Heap's law deals with the relation between total words and unique words in a corpus. Initially there is a rapid increase in the number of unique words. However, the increase in unique word drops off quickly as the corpus grows larger.

It seems like Heap's law may hold for this data, though it is not extremely clear in this plot possibly due to using only 4 datapoints. The plot would have most likely been clearer plotting, for a single data set, the relation between unique and total words as the dataset is processed.

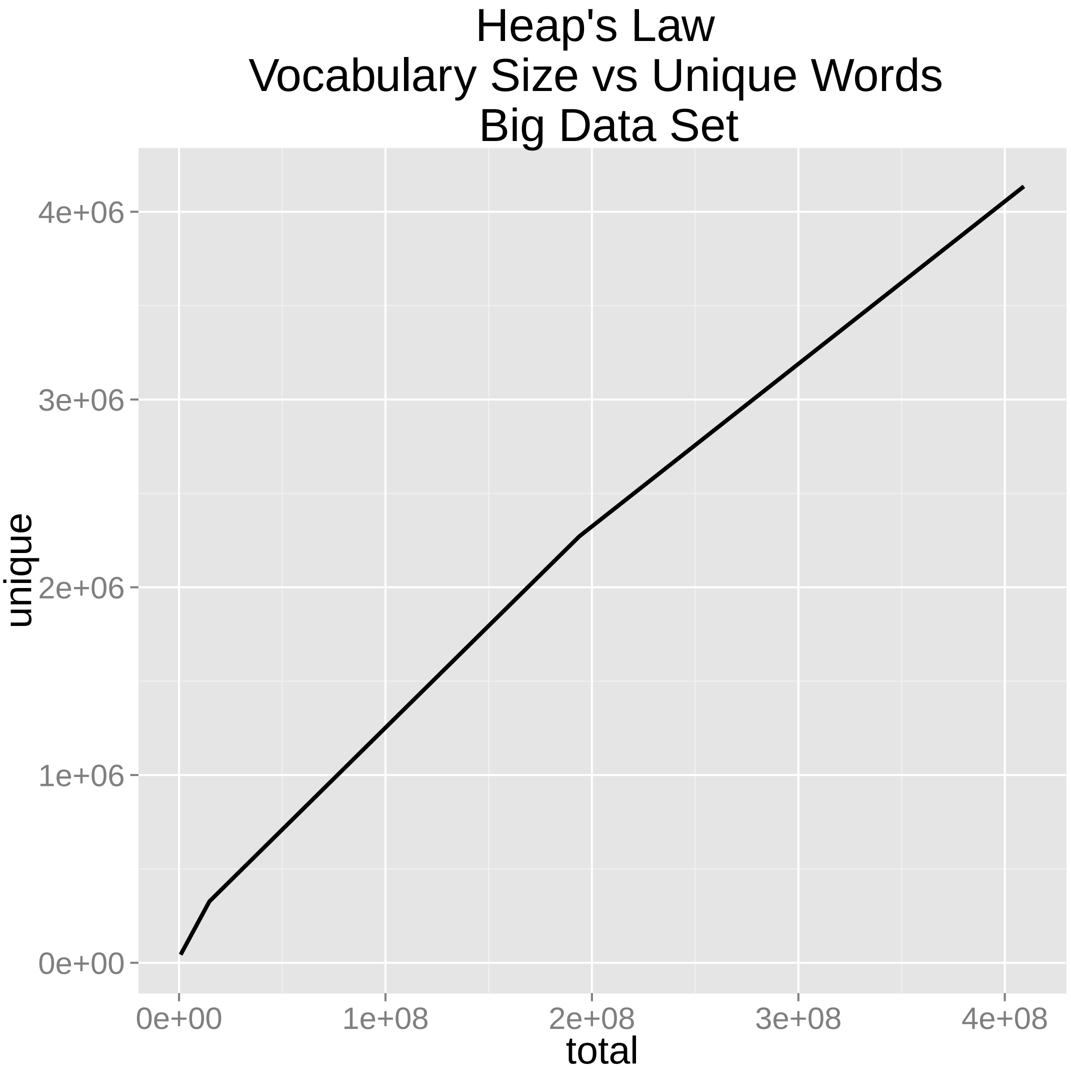

Figure 6.1: Plotting the total words vs unique words for the 4 data sets: tiny, small, medium, and big. Note the differences in axis size

7 Strong Word Associations

For this we look at the word following the special word. For example: **salt** water.

Word	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
powerful	than	influence	army	enough	nation	fleet	effect	man	force	party
strong	enough	position	force	man	hand	hold	desire	men	drink	feeling
salt	water	lake	solution	sea	springs	works	fish	meant	marshes	pork
butter	cheese	flies	eggs	fat	made	fly	milk	maker	making	worth

8 Words Likely to Follow

Here we look at the top 5 words likely to follow the 3 special words as well as their entropy:
P(w2|w1)

Word	Rank 1	Ent. 1	Rank 2	Ent. 2	Rank 3	Ent. 3	Rank 4	Ent. 4	Rank 5	Ent. 5
washington	d	0.089451	dc	0.033986	city	0.019926	county	0.018546	george	0.017597
church	dedicated	0.040684	history	0.011985	music	0.011067	where	0.010856	built	0.008748
james	ii	0.046639	river	0.032846	madison	0.021453	g	0.020119	monroe	0.014899

9 Words Likely to Proceed

Here we look at the top 5 words likely to proceed the 3 special words as well as their entropy:
P(w2|w1)

Word	Rank 1	Ent. 1	Rank 2	Ent. 2	Rank 3	Ent. 3	Rank 4	Ent. 4	Rank 5	Ent. 5
washington	general	0.07240	george	0.06514	gen	0.026457	fort	0.02605	president	0.01720
church	parish	0.05653	catholic	0.04256	presbyterian	0.034152	episcopal	0.03145	christian	0.02821
james	sir	0.06872	st	0.06693	king	0.049857	rev	0.03045	mr	0.02445