Sneha Shankar Narayan
CS646 — Fall 2014

Report on P 1

snehas@cs.umass.edu

1. Implementation choices

Coding language: Python

Parsing: I've used the cElementTree XML Parsing library for parsing the tags. I used the streaming implementation such that no on the fly validation takes place and data is processed faster.

Tokenization: I split the words on the occurences of spaces and later stripped all punctuation and converted the strings to lower case for easy processing.

Design decisions: Code is made to run in parallel. Files are processed in parallel and they output to a queue which again is processed in parallel to consolidate the stats. Garbage collection is done.

Setup: In order to run the code, make sure the data set directory is specified properly in P1/dataDirNames.py and then run using "./main.py datasetname". For example to run the tiny data set all you have to do is say "./main.py tiny"

2. Time Analysis

- Tiny: Run on a 8GB RAM, i5 processor: Time taken in minutes: 0.0470937132835
- Small: Run on a 8GB RAM, i5 processor: Time taken in minutes: 1.60628538529
- Medium: Run on a 4 core processor with 8GB RAM: Time taken in minutes: 18.4011251012
- Large: Run on a c2.4xlarge Amazon EC2 instance with 16 cores and 30GB RAM: Time taken in minutes: 24.2061160843

3. Compare the statistics of small with those of the other run(s) you did. What are the differences? Can you explain them?

- Small has 14548011 tokens and 285963 unique tokens.
- Medium has 56353507 tokens and 874753 unique tokens.
- Large has 377895692 tokens and 3307432 unique tokens.

4. What are the implications of what you found to designing an IR system? For example, how might it affect your use of term weights in an algorithm?

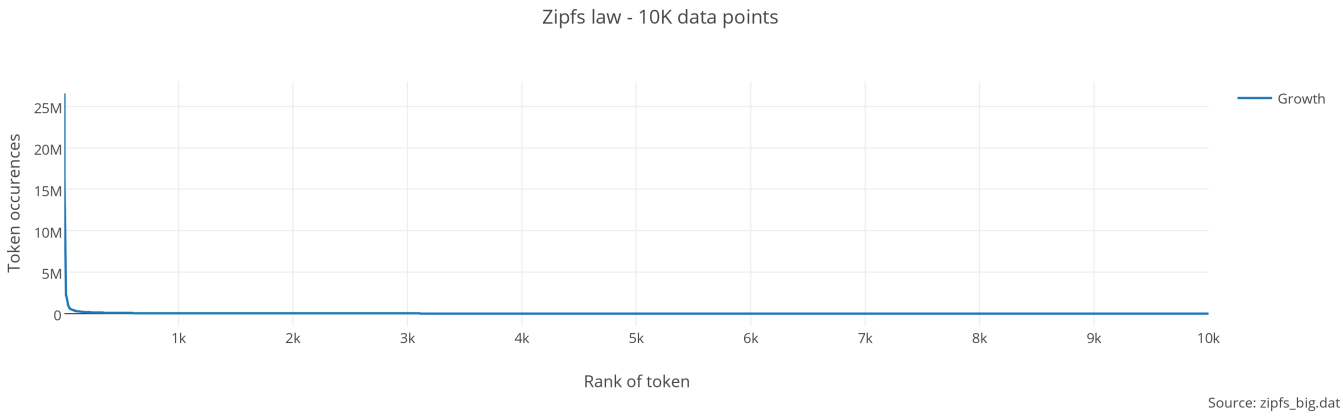
I made the following conclusions:

- The IR System has got to deal with scale. Making sure the big data set runs leads to a lot of thought about the architecture of the system
- Just by analyzing the window and adjacent words an understanding of the corpus can be known.

Large collection

1. Graph the Zipf-related data and explain whether Zipf's Law holds on this data

For the big data set, I'm plotting the rank vs Number of occurences graph for the top 10000 words.



As you can clearly see in the graph, the most frequent words occur the highest number of times and the growth just lies on the X axis after a point of time.

The word ranked 1 occurs 26554000 times and rank 2 appears 16653809 times. Total words in the corpus: 377895692

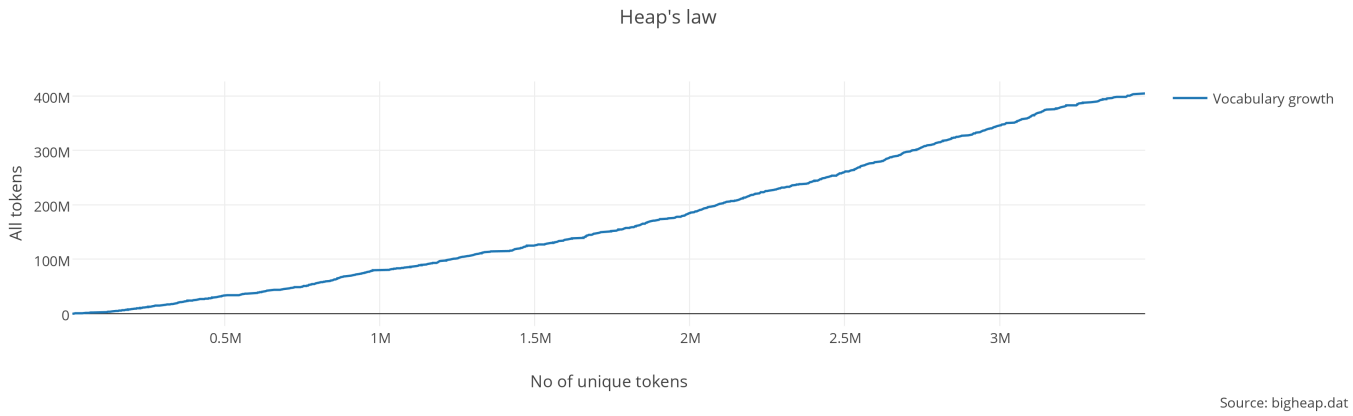
Word rank1 occurs = 26554000/377895692 = 0.07 = 7 % times Word rank2 occurs = 16653809/377895692 =0.044 = 4.4 % times

The total is approximately 10% which means that Zipfs law holds.

2. Graph the vocabulary growth and explain whether Heap's Law holds

From wikipedia: Heaps' law is an empirical law which describes the number of distinct words in a document (or set of documents) as a function of the document length (so called type-token relation).

For the big data set I've noted down the number of tokens, and the number of unique tokens after processing each book. Therefore I get 2521 data points, and I've plotted them in the following graph. The X axis contains the number of unique tokens and the Y axis contains the total number of tokens.



3. For each of these words, list the 10 words with the strongest association and discuss: powerful, strong, butter, salt

- powerful

page	window	adjacent
one, 15900	most, 3457	than, 597
more, 14308	more, 2626	influence, 538
great, 13573	one, 1347	army, 337
into, 12965	than, 1291	enough, 257
other, 12864	great, 1187	fleet, 202
these, 12720	very, 1140	nation, 199
than, 12492	these, 897	effect, 180
most, 12422	influence, 882	man, 160
time, 12124	her, 849	force, 154
some, 11830	would, 810	party, 152
only, 11667	against, 775	nations, 135
made, 11227	into, 763	king, 128
first, 10943	some, 756	tribe, 124
would, 10856	upon, 747	telescopes, 123
upon, 10847	will, 724	enemy, 112

- strong:

page	window	adjacent
one, 39333	very, 3770	enough, 2522
more, 31650	one, 3111	position, 1344
into, 30524	enough, 2515	force, 1013
other, 29334	too, 2358	man, 585
great, 29328	force, 2226	hand, 492
some, 29259	will, 2220	desire, 422
time, 28694	her, 2171	hold, 421
these, 28072	made, 2158	men, 413
than, 28067	men, 2148	drink, 396
made, 27847	position, 2089	feeling, 381
would, 27125	against, 2084	one, 330
upon, 26744	upon, 2048	line, 312
only, 26735	man, 2019	body, 291
two, 26597	if, 2006	party, 290
if, 26531	would, 1984	current, 288

- salt:

page	window	adjacent
one, 6611	water, 2122	lake, 1095
into, 5783	lake, 1134	water, 840
other, 5672	solution, 930	solution, 317
some, 5465	2, 686	sea, 233
these, 5075	into, 676	works, 179
more, 5005	sea, 673	springs, 165
two, 4868	common, 644	fish, 139
great, 4853	great, 606	river, 131
may, 4732	one, 583	meat, 124
very, 4520	some, 566	marshes, 120
than, 4476	other, 552	pork, 116
about, 4444	acid, 529	marsh, 114
only, 4435	city, 479	lakes, 108
water, 4383	fresh, 474	beef, 69
made, 4326	if, 467	pond, 57

- butter:

page	window	adjacent
one, 1808	cheese, 521	cheese, 150
some, 1566	milk, 425	per, 90
other, 1547	bread, 331	worth, 67
into, 1484	per, 302	flies, 64
more, 1420	eggs, 233	eggs, 46
great, 1352	c, 167	fat, 45
two, 1334	other, 161	beef, 41
than, 1326	cream, 151	made, 38
very, 1319	beef, 147	milk, 33
these, 1284	made, 142	fly, 29
about, 1282	os, 137	field, 24
made, 1254	fat, 129	maker, 22
only, 1234	ib, 128	making, 21
time, 1207	fresh, 127	cross, 18
most, 1178	some, 127	nor, 16

Following conclusions can be made:

- Adjacent words of powerful contain words like army, fleet, nation which makes sense as 'powerful' is the adjective that can be used to describe stuff. Powerful also lies in the window of various quantifiable words like "more", "than".
- Not surprisingly, strong also occurs in the vicinity of similar quantifiable words but in a higher frequency.

- Strong also happens to appear with the word "men" which does not bode well with the feminist part of me :P
- Salt appears adjacent to words like pond, river which shows the salinity of water. It also appears in the window of solution and acid which suggests that the corpus is talking about chemistry
- Butter appears with a range of food related words like egg, milk and bread which suggests that recipes are being talked about. It also is in the window of words like 'fat', 'fresh' which shows the quality of butter.
- Surprisingly the words that appear the most in a page with "Strong" and "powerful" and with "salt" and "butter" are similar. Since strong and powerful are adjectives used in similar circumstances, this makes sense. The same applies to salt and butter which can both be used in describing recipes or something similar.

4. For each of these words, list the entropy of the word and the 5 words most likely to follow them – i.e., P(w2—w1): Washington, James, church

I am defining the entropy to be
p = Occurences of w1 followed by some word / Occurences of w1 in the corpus

$$Entropy = -p\log(p)$$

- James

  The number of occurrences of James in the corpus is: 132808

word	Number of occurrences	entropy
ii	3553	0.042072
h	2483	0.032312
river	2451	0.031999
m	2193	0.029428
w	1976	0.02719
madison	1897	0.026356
b	1608	0.02321
do	1426	0.021143
c	1289	0.019537
monroe	1281	0.019442

- Washington The number of occurrences of Washington in the corpus is: 90535

word	Number of occurrences	entropy
d	10396	0.086604
george	2674	0.034149
city	1252	0.019096
dc	1049	0.016606
county	794	0.013293
irving	593	0.010494
july	521	0.00944
june	402	0.007625
college	397	0.007546
where	378	0.007246

- Church The number of occurrences of church in the corpus is: 240921

word	Number of occurrences	entropy
dedicated	1913	0.026526
history	1151	0.017872
music	868	0.014279
where	856	0.01412
new	782	0.013131
oxford	725	0.012353
st	701	0.012021
government	667	0.011547
should	606	0.010681
property	597	0.010551

5. For the same words as in (4), list the 5 words most likely to precede the word.

- James
  1. sir 4788
  2. king 3042
  3. st 3038
  4. mr 1816
  5. rev 1733
- Washington
  1. george 2952
  2. general 2393
  3. 0 2155

4. 1779 1265
5. 65 1004
6. 1780 922
7. fort 860
8. near 664
9. gen 663
10. president 626

- Church

1. catholic 5365
2. parish 3883
3. presbyterian 3386
4. christian 3320
5. episcopal 3247

Conclusions:

- The word church is preceded most by the various types of churches which is understandable
- James is preceded by various titles like Lt and Colonel and Sir, and hence can be characterized as a name.
- Washington is preceded by "George" or "General" which also signifies it as a name
- By just merely looking at the above facts we can say that the corpus has sections related to the history of the United States, and the fact that various churches are mentioned we can understand that the corpus may have sections of the various types of churches.