Honey Nut Clusters

Brice Scott, Brandon West, Colby Cartwright

The data that we will use is Wikipedia dumps. We have already downloaded the latest complete English Wikipedia dump that totals (when extracted) to 37.7GB. Figure/Table 1 shows the word frequency for the top twenty words and Zipf's curve of the entire Wikipedia file.

Because Wikipedia contains so many articles, our initial goal of performing hierarchical clustering on the entire Wikipedia might be unrealistic. Also because of the size we would have to store it on disk making the entire process of analyzing it extremely slow. Because of this we used Wikipedia's Special::Export tool to download a 13MB XML file that contains around 1500 articles. The articles downloaded are based on science, engineering, computer, mathematics and other science related categories. We will use this small file for testing purposes. Figure/Table 2 shows the word frequency for the top twenty words and Zipf's curve of this small file. By using only a small subset of Wikipedia we should be able to keep everything in RAM, speeding up the entire program, and only requiring the tf-idf vectors to be saved on disk.

We are making a modification to our Project by only using a smaller subset of Wikipedia. Instead of allowing users to browse through the entire Wikipedia we are going to focus more Wikipedia articles that are relevant to college students.

Table 1

the	160910158
of	102109943
and	67048607
in	66568469
to	59590186
is	29245795
for	25508367
on	21745345
was	21133251
as	19790864
by	18485800
that	16723028
with	16702919
it	16119454
from	14528676
at	13255091
this	12873397
be	11240966
he	10847206
an	10842524

Table 2

the	82678
of	57403
and	37979
to	28094
in	27742
is	23908
for	12271
that	11434
as	10965
by	9397
are	9164
be	8373
with	7827
or	7752
on	7088
an	7033
this	6281
it	6021
from	5225
can	5223

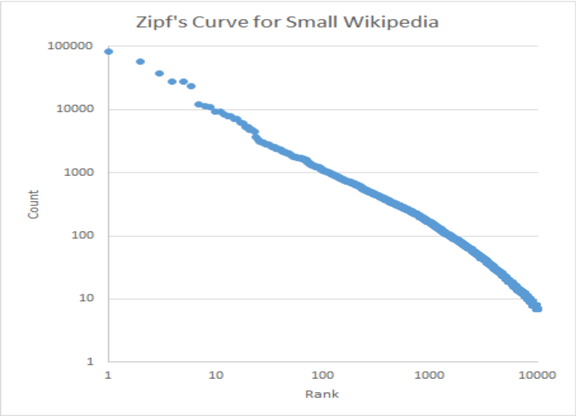

Figure 1

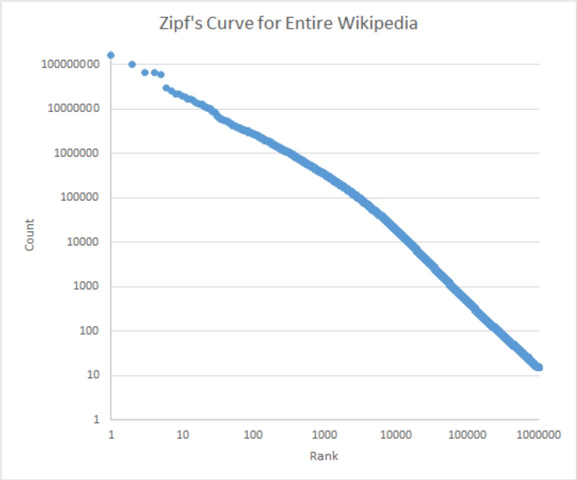

Figure 2