CS123 Wikipedia Project: Hypotheses

Julian McClellan, Bobby Adsumilli, Andy Zhu

May 21, 2016

Files

wikidata/wikistats (650G)

Contains hourly wikipedia article traffic statistics dataset covering 16 month period from October 01 2008 to February 6, 2010, from raw anonymous logs provided by Domas Mituzas at http://dammit.lt/2007/12/10/wikipedia-page-counters/

Each log file is named with the date and time of collection:

pagecounts-20090430-230000.gz

Each line has 4 fields: projectcode, pagename, pageviews, bytes

```
en Barack_Obama 997 123091092
en Barack_Obama%27s_first_100_days 8 850127
en Barack_Obama,_Jr 1 144103
en Barack_Obama,_Sr. 37 938821
en Barack_Obama_%22HOPE%22_poster 4 81005
en Barack_Obama_%22Hope%22_poster 5 102081
```

wikidata/wikilinks (1.1G)

Contains a wikipedia linkgraph dataset provided by Henry Haselgrove. These files contain all links between *proper english language Wikipedia pages*, that is pages in "namespace 0". This includes disambiguation pages and redirect pages.

In links-simple-sorted.txt, there is one line for each page that has links from it. The format of the lines is:

```
from1: to11 to12 to13 ...
from2: to21 to22 to23 ...
...
```

where from1 is an integer labelling a page that has links from it, and to11 to12 to13 ... are integers labelling all the pages that the page links to. To find the page title that corresponds to integer n, just look up the n-th line in the file *titles-sorted.txt*.

File summary:

pagecounts-.gz -Contains n lines with 4 fields: projectcode (we want "en" only!), pagename, pageviews, and bytes -Dates range from 10/01/2008 to 02/06/2010 *links-simple-sorted.txt* -Contains all links between proper english language wikipedia pages.

titles-sorted.txt -Contains the names of all English Language Wikipedia pages.

Hypotheses

1. We will attempt to quantify the relationship between pages and inlinks to these pages.

e.g. Imagine the simplest case. A page only has one inlink, IL. Let's assume a linear model for the traffic of these three links as: