CS123 Wikipedia Project: Data and Hypothesis

Julian McClellan, Bobby Adsumilli, Andy Zhu

May 21, 2016

Data 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.

Hypothesis

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

Imagine the simplest case. A page only has one inlink, IL. We propose a linear model for the traffic of that page as:

$traf\_page_t = \beta_0 + \beta_1 traf\_IL_t + \beta_2 bratio\_IL_t + \beta_3 bytes\_page_t \epsilon_t$

Where $bratio\_IL_t = \frac{bytes\_IL_t}{bytes\_page_t}$. That is, the bytes ratio is the ratio of the page size in bytes, of the inlink IL to the page.

We would be most interested in the coefficients β_1 and β_2. Respectively, these can be interpreted as the change in page traffic when the inlink traffic increases by one unit holding the bytes ratio constant (β_1) and vice versa (β_2).

However, we do not simply believe that the traffic of a page is a function of the attributes related to the link (or links) to that page. Given the low dimensions of our data set we use the predictor bytes_page, the size of the page in bytes as an "internal" cause of the traffic to that page. Bytes can be thought of as a direct way to measure the amount of content the page contains.

Once the actual data has been more thoroughly explored, we might be inclined to introduce an interaction term between the bytes ratio and inlink traffic. Doing so complicates a simple explanation of what the regression coefficients mean, but suffice to say, a statistically significant interaction term means that the effect on page traffic that the traffic to the inlink IL has depends on the value of the bytes ratio.

Now, to generalize the model, say a page has i inlinks ($2 \leq i < \infty$), then the model generalizes to:

$traf\ page_t = \beta_0 + \beta_1 traf\_1 + ... + \beta_i (traf\_i) + \beta_{i+1} bratio\_1 + ... + \beta_{2i} bratio\_i + \beta_{2i+1} bytes\_page + \epsilon_t$

In this case, we would be interested in all β's with subscripts greater than 0. With this linear regression we can test a number of hypotheses:

Does the traffic from each inlink affect the page traffic in the same way?

Null | $\beta_1 = \beta_2 = ... = \beta_i$

Do the bytes ratios have any effect on page traffic?
Null | $\beta_{i+1} = \beta_{i+2} = ... = \beta_{2i} = 0$

Do the bytes ratios all have the same effect?

Null | $\beta_{i+1} = \beta_{i+2} = ... = \beta_{2i}$

etc.