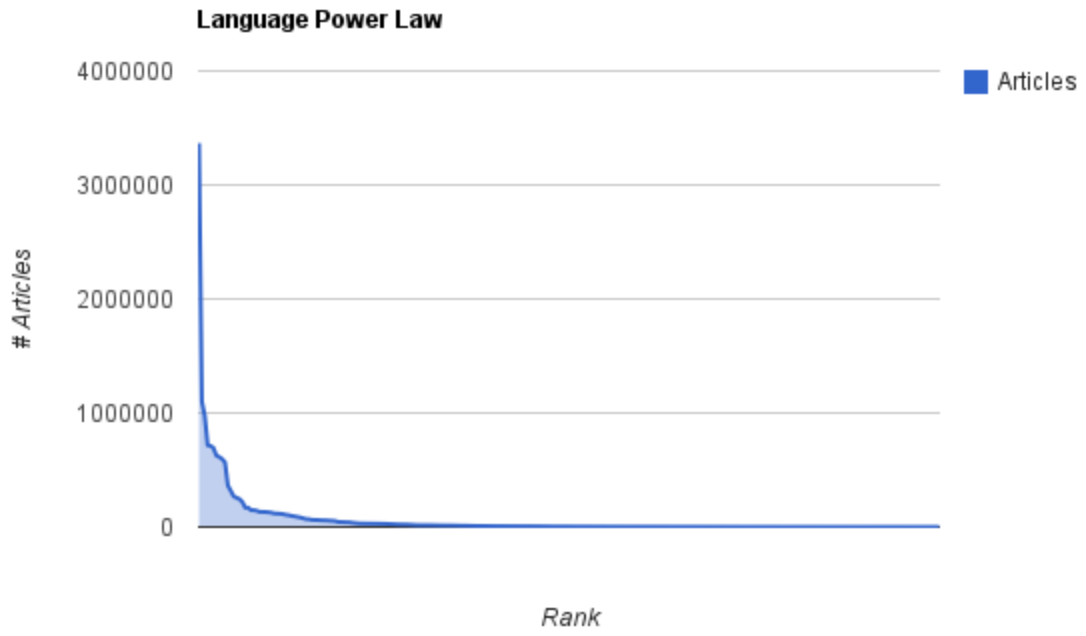
Galen Panger
3/7/2012

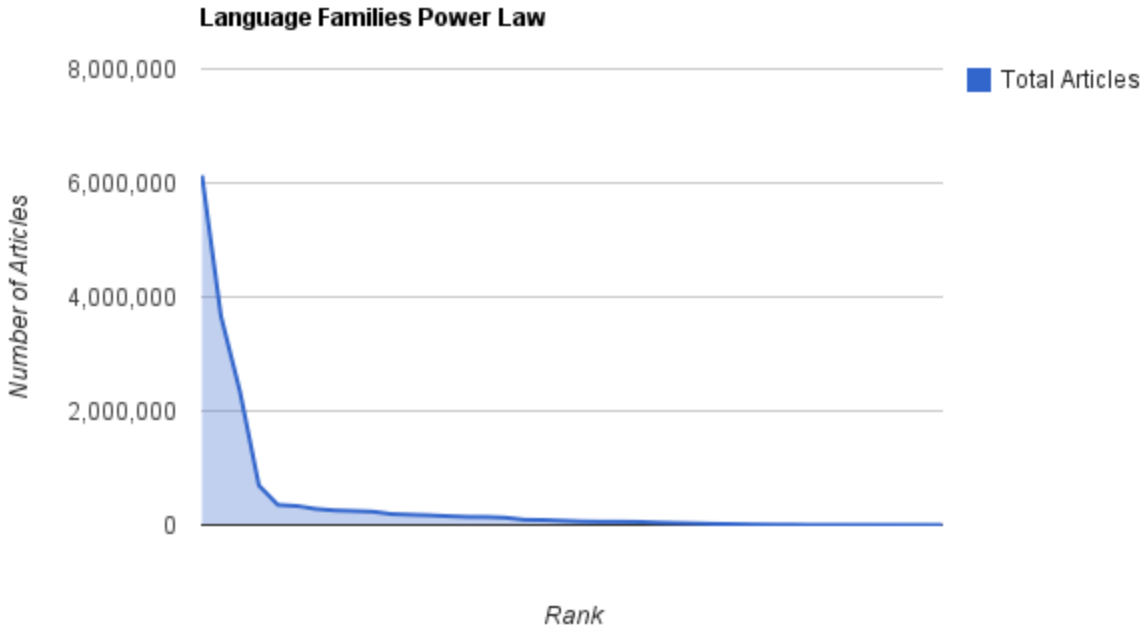
Link: http://people.ischool.berkeley.edu/~gpanger/i247/3/

I've long been fascinated by concepts of "soft power," or the kinds of non-military means that countries and cultures use to influence one another. So for this assignment, I decided to look at data on the size of Wikipedia's various encyclopedias, which are organized by language, as a way to get at the sort of soft power that cultures and their languages deploy by way of online knowledge production.

Though perhaps it should have been obvious, I was struck by the way that the various Wikipedias exhibited the traditional power law shape found in many online communities, meaning that their size drops off precipitously. It's not just user contributions that exhibit the power law, though, it's whole languages on Wikipedia. After cleaning up my data and looking at the histograms (see "hist" folder), I made this chart showing the power law for the different language versions of Wikipedia:



Interestingly enough, the power law still holds even when you group by language family:



To me, this signals that not only are some languages more influential online, some of them *dominate* online. I looked at a number of other variables (edits, users, etc.) and they too exhibit power laws and a distribution indicating a few dominate the many (see "dominance and power laws" folder). I decided to base my visualization around the power law and its corresponding implications of *dominance*.

Visually, I think dominance is most strongly connoted by area; dominance connotes the geographic spread of conquering empires. You can see on page 1 of my sketches that I considered a bar chart, but I didn't think that would be as compelling. So the basic unit in my visualization is a circle. In addition, to focus on the *appearance* of dominance, I used apparent scaling (Flannery[1]) rather than actual.

On page 1 of my sketches, you can see that I considered clustering the language families, centered on the Germanic family, which is the largest by number of articles. However, I wanted the data to still mimic a power law shape, and so I decided to orient the circles along the x-axis (by rank according to number of articles). This orientation was useful for also sorting by attributes other than number of articles, and you'll see that my final visualization has a number of sorting options. The position of the circles changes according to the sort, but the area of the circle doesn't. This helps with comparison.

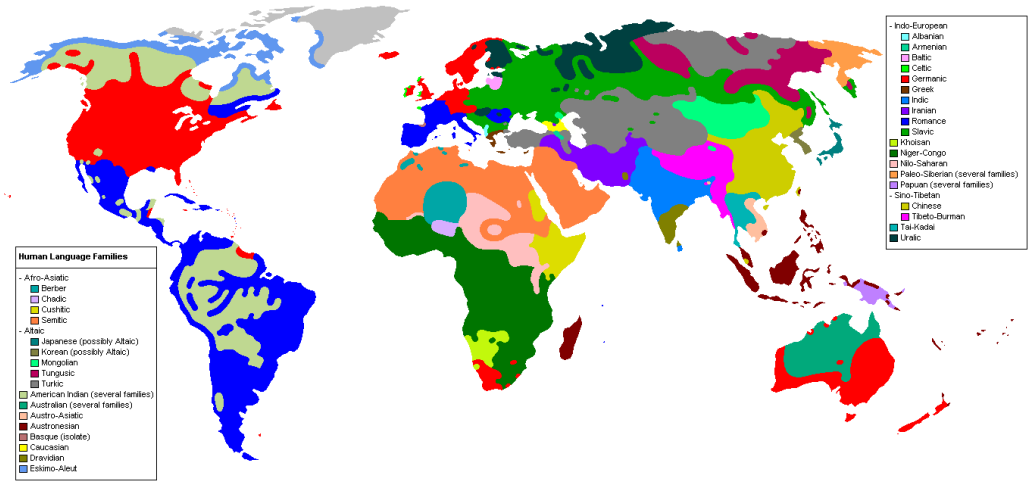
It also allows users to see basic correlation between number of articles and all other variables. I didn't want to use scatterplots for this because of the potential for data occlusion and because I wanted a simpler layout and presentation. So the correlation analysis afforded by my visualization is general-level but still fairly clear: languages, articles, edits, admins, users and images are well-correlated with one another, while the other variables are surprisingly uncorrelated, or correlated only at low levels (see

[1] radius = (beta * # articles) ^ 0.5716, where beta is scaled such that all circles fit on the canvas

Appendix for correlation table).

Ultimately, my hope had been to allow users to drill down into each language family and do the same kind of sorting on languages within each family. Unfortunately, though I spent several hours making this attempt, I was unsuccessful. Selectively appearing/disappearing/filtering and sorting the resulting sets proved unworkable for me. The other (minor) problem I had was the fact that many of my numbers had been formatted in my spreadsheet to have commas (as in 1,000,000), which when sorting values produced odd orderings that I did not immediately detect and did not know how to deal with. After some work, I was able to sort values correctly on every dimension while still displaying a comma-formatted value upon mouseover, which was my goal.

Lastly, a word about color and opacity. There are 39 language families in Wikipedia's dataset, which is way too much to apply categorical color values, though Wikipedia itself has done this in its graphic of language family distribution:



Rather than follow Wikipedia's bad example, I made every circle the same color. However, I did make them semi-transparent, because through experimentation I found that it was easier to follow a circle's movement that way. For visual polish, I animated transitions smoothly and attractively, but otherwise I kept the styling and decoration dead simple.

Tools used:

- Google ImportHTML
- Google Spreadsheets for almost all data processing and graphing
- Stata for histograms, scatterplots and correlation analysis
- Mr. Data Converter for JSON conversion
- d3
- TextMate

Appendix: Correlation table

```
. pwcorr articles edits admins users images editsarticle articlesadmin articlesuser
imagesarticle usersadmin, sig

             | articles    edits   admins    users   images editsa~e articl~n
-------------+------------------------------------------------------------------
    articles |   1.0000
             |
             |
       edits |   0.9479   1.0000
             |   0.0000
             |
      admins |   0.9277   0.9939   1.0000
             |   0.0000   0.0000
             |
       users |   0.9012   0.9865   0.9893   1.0000
             |   0.0000   0.0000   0.0000
             |
      images |   0.9048   0.9815   0.9788   0.9709   1.0000
             |   0.0000   0.0000   0.0000   0.0000
             |
editsarticle |   0.0760   0.1084   0.0989   0.1067   0.1041   1.0000
             |   0.2246   0.0827   0.1138   0.0879   0.0959
             |
articlesad~n |   0.0959   0.0180   0.0015   0.0076   0.0101  -0.1752   1.0000
             |   0.1438   0.7844   0.9813   0.9084   0.8782   0.0072
             |
articlesuser |   0.0226  -0.0290  -0.0273  -0.0371  -0.0291  -0.2824   0.8552
             |   0.7182   0.6438   0.6635   0.5536   0.6420   0.0000   0.0000
             |
imagesarti~e |   0.0483   0.0743   0.0765   0.0736   0.1207  -0.0535  -0.1162
             |   0.4407   0.2353   0.2219   0.2400   0.0532   0.3926   0.0761
             |
  usersadmin |   0.3605   0.2372   0.1960   0.2397   0.2002  -0.0060   0.3014
             |   0.0000   0.0003   0.0026   0.0002   0.0021   0.9277   0.0000
             |

             | articl~r images~e usersa~n
-------------+---------------------------
articlesuser |   1.0000
             |
             |
imagesarti~e |  -0.1101   1.0000
             |   0.0781
             |
  usersadmin |   0.0177   0.0086   1.0000
             |   0.7875   0.8954
             |
```