

wikipedia-rankings

=====

Support files for [\[TIME's ranking of the prominent people on Wikipedia\]](https://time.com/109947/web-ranking/).
(<https://time.com/109947/web-ranking/>).

Data was collected over several days in May using [\[node-wikipedia\]](https://www.npmjs.org/package/node-wikipedia).
(<https://www.npmjs.org/package/node-wikipedia>), a Node.js module maintained by
[@wilson428](https://github.com/wilson428) (<https://github.com/wilson428>).

We considered eight data points for each entry:

- + Number of words
- + Number of links to other Wikipedia pages
- + Number of external links (which are typically references)
- + Number of categories the person is in
- + Total number of revisions to the page
- + Number of unique individuals who have edited the page as a signed-in editors
- + Number of anonymous edits
- + Number of vandalisms, as identified in editing notes

Data for the top 100,000-or-so people is available as a [\[15MB CSV file\]](#)(/people.csv).

Analysis

Using out-of-the-box R functions, we reduced these eight variables to their principal components (using [\[this handy guide\]](http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html#principal-component-analysis)(<http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html#principal-component-analysis>)). As you can see, a huge amount of the variance is contained in the first PC:

! [\[variance\]](https://raw.githubusercontent.com/TimeMagazine/wikipedia-rankings/master/variance.png)(<https://raw.githubusercontent.com/TimeMagazine/wikipedia-rankings/master/variance.png>)

You can rerun the principal component analysis like so:

```
RScript wikipedia.r
```

(This may require installing the relevant libraries first).

By trial and error, the ranking that most satisfied our anecdotal sense for "influence" in the real world was PC1 + PC2, which becomes the ``score`` for each person.