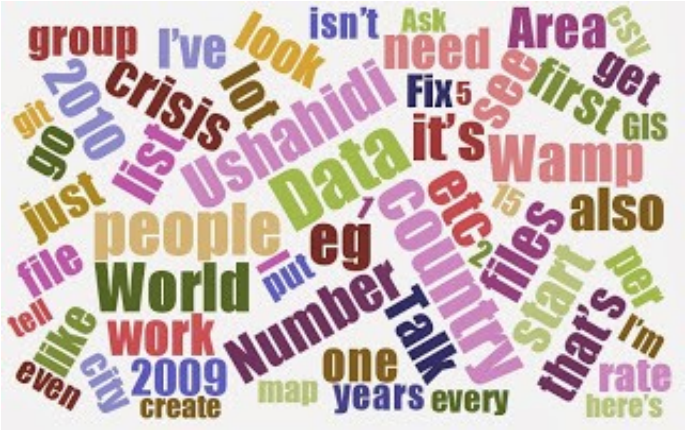
Wut 2 Do Wif Werdz

[Cross-post from IcanHazDataScience]

One of the first cool things beginning data scientists do with a bunch of words is create a word cloud. There are several ways to do this, but it's important to know about tools like D3 (one of the most powerful software libraries for data visualisation), so we'll use that. Before you run off screaming because it's a javascript library, it's okay: there are a whole bunch of example pages on the D3 website, that you can play with your own data with. One of these is the word cloud generator. Go there and play with it now... here are some clouds for both crisismapping and my own sites, with some comments...

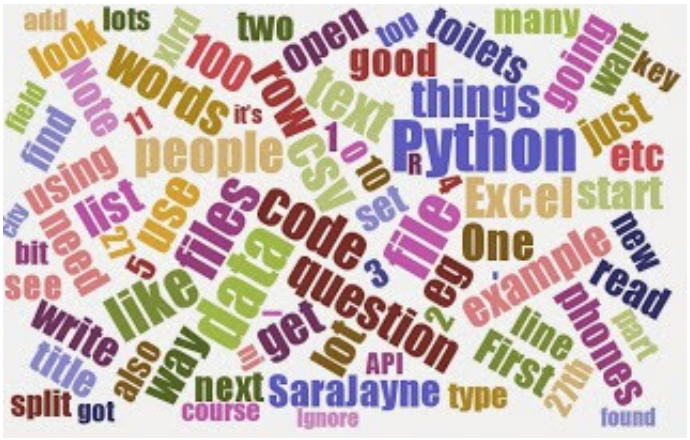


blog.overcognition.com - my work blog. I seem to have a thing about countries, Wamp, Ushahidi and Data (which really reflects the last 3 posts that I wrote). Now because I used Jason's example code, it's using every word on the site, including the stopwords (it's, get, also etc), so they're quite significant here too. But it's a pretty cloud.

Blog.standbytaskforce.com - the SBTF blog. Nothing really shouts out in this wordcloud - except perhaps 'jQuery' and function - which is odd. It's possible that the code is picking up words that are in the html file but not on the page itself (right-click on the page, "view source" to see what I'm talking about).

www.crisismappers.net - lots of good words like "humanitarian" and "Technology" (note the capital letters: something else we'd remove if we processed the data before feeding it to our own app). Also "Jen" and "Ziemke", which makes sense because Jen Ziemke is the main organiser for this site.

icanhazdatascience.blogspot.com. Oh whoops: that didn't work so well. It's full of font names (Helvetica, Ariel etc) and colour codes (#009EB8, #333333 etc: go to http://html-color-codes.info/ if you want to know which colours these are). No fear... if I just cut-and-paste the page (thanks, control-A!) contents into the app, the picture gets a bit clearer:



Better, huh? Icanhazdatascience.blogspot.com appears to be obsessed with Python, interested in code and people, and likes questions (and someone called SaraJayne). The moral of this little story: be very careful to check that what you *think* is the data going into an app, actually *is* the data going in; oh, and that wordclouds can be a cool way to double-check that.

The last wordcloud is http://www.opencrisis.org/ - again, there's some non-visible html creeping into the wordcloud, but the basic cloud looks good for what we do - which is inform people about ways to process crisis data, and volunteer groups etc who can help.

Oh heck... just because I can... the opencrisis.org cut-and-paste wordcloud. Crisis is big, Data is big, and we like months (there's a list of upcoming events on the front page).

More D3 notes later - in the meantime, please go and play with the examples at https://github.com/mbostock/d3/wiki/Gallery and start thinking about how they could be applied to your data!