Basic Materials: Werdz an Regulah Expreshuns

[Cross-posted from ICanHazDataScience]

Okay, that last post was a bit long for Emily... she fell asleep on my desk long before I'd finished typing. So today we're back to short and practical.

Data is not just numbers. Numbers are one of the basic types of data that appear again and again in data science. Two of those types are words (as in written text, like this blogpost) and networks (as in objects connected with links - like a diagram of your twitter friends and your friends' friends etc). Today we're looking at words.

In the last post, I was looking at a set of online job descriptions. We'll leave the basics of webpage scraping til later (but if you're curious, ScraperWiki's notes are good) and assume that what we have is a set of text files that we've used the "Processing all the teh Files in Directory" post with the commands