Lab 5: Analogies

October 12th 2017

In this lab you'll use word embeddings to solve analogies for you. I have included some basic starter code to show you how to structure your data in an SVD-friendly way. The plot at the end of the code was mostly just a sanity check. It should be deleted before you turn in your assignment. This assignment is due by midnight on 10/18/17.

Relevant Resources:

- http://www.nltk.org/book/ch02.html
- http://bit.ly/2i3baH9

Part 1: 30 points

Use the Brown corpus (included in nltk) to construct a matrix of word embeddings (aka word vectors). Save this matrix as its own file that can then be loaded into your solutions for Part 2 and Part 3. Set your context window to be equal to 7.

Part 2: 30 points

Write a program that prompts the user to submit an analogy, which will be of the form:

```
fly is to plane as drive is to
```

You will then use your matrix to cough up the word that is most likely to complete the analogy. This can be done using the equation from the famous word2vec paper:

$$||w_x - w_{drive} + w_{fly} - w_{plane}||^2$$

The word from your matrix that minimizes this equation is the winner. Your program should allow the user to continue to enter analogies until they type `quit`. In a `readme` file include three analogies that your program attempted to answer. For your first analogy, start with one of medium difficulty. If your program gets it wrong, try to think of an easier one for your second analogy. If it gets it right, try to think of a harder one. Same thing for your third analogy. You should be able to come up with one analogy that your program can answer.

If you use the example analogy that I've already used, you will get a 0 on this entire assignment.

Part 3: 30 points

Use SVD to transform your large, sparse matrix into a smaller, dense one. Experiment with a few different values of *k* (as in the *k* most significant columns). Then, just as you did for Part 2, test your dense matrix on three different analogies that get harder or easier, depending of whether your program is successful. You should be able to bump up the difficulty using your dense matrix. Include these three analogies in your `readme` as well.

Part 4: 10 points

Come up with a hilarious analogy that would completely stump a computer. Be thankful that you're human and not such a ridiculous hunk of circuits and linear algebra.

Step 7

Submit your matrix, code, and readme using the command:

```
provide comp150nlp analogy analogy.zip
```