Gray Davidson

Minimum Viable Product
Project Fletcher
Metis SF Spring 2018

Domain: A brief description of the domain you're working in, and a discussion of your familiarity with this domain

I will use two (or more) large corpuses of text data from distinct authors (whole bibliographies) to train multiple embeddings with weights unique to the authors. An AWS P3 (or similar) instance will be employed when the pipeline is ready. Once the embeddings are trained they will be used to complete a 'mad-libs' style task, using a generic sounding sentence as the base case and 'adjusting' it based on the unique style and lexicon of the different embeddings. If I cannot find sufficiently large datasets restricting to one author then I will expand to incorporate multiple authors with very similar genres. I will be restricting myself to English language authors to avoid the stylistic alterations of human translators.

Data: A table with variables name and variable type fields, listing the variables you'd like to utilize

Name	Type	Description
three_sentence_chunks	str	From the original corpus I will break the data into three-sentence long strings, strip punctuation and make all characters lowercase.

Known unknowns: A list of items with an unclear level of effort, or which will require special attention

The use of AWS is intimidating. I believe that following Chris Albon's tutorial:

https://chrisalbon.com/software_engineering/cloud_computing/run_project_jupyter_on_amazon_ec2/

I will be able to instantiate a virtual machine but this is untrod territory for me. I am less concerned about spinning up the instance as I am about loading my data into the virtual machine and getting my saved model back out. Helen suggested this could be done with a simple scp command but I haven't tried it yet (ever).

Moonshot:

- I would like to go on to study Variational Auto Encoders and how they relate to this project.
- I would like to download and use a standard embedding to see how a 'plain-english' embedding would handle the mad-lib challenges.
- I would like to scrape and use (or borrow from a classmate) a dataset from a very different source, *e.g.* Twitter, to see how "The modern internet" conceives of certain word definitions.