Language Modeling and Text Generation

Group CFX (Clark, Florian, Xiong) Project 1

Submission 2: How to Run Our Code

ECU CSCI 6040 Spring 2020

Professor Gudivada

Participants

Kelle Clark
Andrew Florian
Xinyu Xiong

1. Start JupyterLab https://mybinder.org/v2/gh/jupyterlab/jupyterlab-demo/try.jupyter.org?urlpath=lab
2. Upload the following files:
    - Generate.ipynb
    - Unigramfile.dat (the most frequent 300,000 unigrams from 1.5 GB of Gutenburg texts)
    - Bigramfile.dat (the most frequent 270,000 bigrams from 1.5 GB of Gutenburg texts)
    - Trigramfile.dat (the most frequent 230,000 trigrams from 1.5 GB of Gutenburg texts)
    - Quadgramfile.dat (the most frequent 190,000 quadgrams from 1.5 GB of Gutenburg texts)
3. Run Generate.ipynb and scroll down to see the searches, decisions, and finally the 100 words that our model generates in about 10 seconds on a typical PC given the word ‘the’
4. Try different options. Example: to generate 5 words with “what” as a seed, use generate(“what”,5)
5. See our other report explaining our design decisions and approach to solving the problem including smoothing graphs and quality analysis. We went through 5 phases of development and due to inefficiencies in NLTK and the size of some of the text files, much of our code took hours (and sometimes days) to run so is not suitable for a quick demo, but can be made available via our GitHub repository if needed.