Language Modelling in Python

Anirudh Kalla

October 14, 2020

Project Description: Analysing Language Models using graphical and numerical tools in Python.

Programming Language: Python 3

Distribution: Anaconda v2020.07

Text Corpora:

1. mother-goose-corpus
2. brown-corpus
   (http://www.sls.hawaii.edu/bley-vroman/brown_corpus.html)

Libraries and Packages used:

1. numpy: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
2. os: This module provides a portable way of using operating system dependent functionality. If you just want to read or write a file see open(), if you want to manipulate paths, see the os.path module, and if you want to read all the lines in all the files on the command line see the fileinput module. For creating temporary files and directories see the tempfile module, and for high-level file and directory handling see the shutil module.
3. SciPy: Open-source Python library which is used to solve scientific and mathematical problems. It is built on the NumPy extension and allows the user to manipulate and visualize data with a wide range of high-level commands.
4. matplotlib: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.
5. Operator: Exports a set of efficient functions corresponding to the intrinsic operators of Python.

6. re: Module provides regular expression matching operations similar to those found in Perl. Both patterns and strings to be searched can be Unicode strings (str) as well as 8-bit strings (bytes).

7. Statistics: This module provides functions for calculating mathematical statistics of numeric (Real-valued) data.

8. sklearn.preprocessing: Provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators. Standardization of datasets is a common requirement for many machine learning estimators implemented in scikit-learn; they might behave badly if the individual features do not more or less look like standard normally distributed data: Gaussian with zero mean and unit variance.

9. Seaborn: Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

10. sklearn.neighbors: Provides functionality for unsupervised and supervised neighbors-based learning methods. Unsupervised nearest neighbors is the foundation of many other learning methods, notably manifold learning and spectral clustering. Supervised neighbors-based learning comes in two flavors: classification for data with discrete labels, and regression for data with continuous labels.

11. math: This module provides access to the mathematical functions defined by the C standard.