Justin Melde
INFX 575 – Data Science III
04/26/2015

Problem Set #2

Introduction

Working off of a set of 10 text documents, the learning goals for this problem set include becoming familiar with the cloud command line interface; interacting with S3 (extract, move and store data); and exploring NLP basics (stemming, stop-words, n-grams).

Overview

The Natural Language Processing Toolkit is a free, open source platform in Python built for processing, stemming, and parsing human language. The package revolves around extracting textual content from different media, 'tokenizing' the words found therein, and then performing different language processing techniques to manipulate and make sense of the text.

I used NLTK to initially remove stop-words from the ingested data (along with punctuation using the String package), and then tokenize the remaining text into discrete elements. I then used NLTK to stem the tokenized words using the Porter stemming algorithm, and then produce uni, bi, and trigrams of the stemmed words. I lastly used NTLK to count the frequencies of the different ngrams, and then used Pandas dataframes to manipulate the data and remove unwanted characters before writing to CSV.

Query String Authentication

Please see the Query String Authentication.py file for the code used to produce the following URL (accessible for the next 10 days). This is a protected link on AWS S3 for the summary_trigrams.csv file:

http://jmelde-bucket.s3.amazonaws.com/ps2_output/summary_trigrams.csv?Signature=saKUBtY89EmJ0qmj9Ol039Yl1qE%3D&Expires=1430937803&AWSAccessKeyId=AKIAJIWN4ZEIRRRWVVSQ

Command Line Interface

Downloading text files from S3

```
C:\Windows\system32>aws s3 cp s3://uspto-patentsclaims/6334220.txt .
download: s3://uspto-patentsclaims/6334220.txt to .\6334220.txt

C:\Windows\system32>aws s3 cp s3://uspto-patentsclaims/6334221.txt .
download: s3://uspto-patentsclaims/6334221.txt to .\6334221.txt

C:\Windows\system32>aws s3 cp s3://uspto-patentsclaims/6334222.txt .
download: s3://uspto-patentsclaims/6334222.txt to .\6334222.txt

C:\Windows\system32>aws s3 cp s3://uspto-patentsclaims/6334223.txt .
download: s3://uspto-patentsclaims/6334223.txt to .\6334223.txt

C:\Windows\system32>aws s3 cp s3://uspto-patentsclaims/6334224.txt .
download: s3://uspto-patentsclaims/6334224.txt to .\6334224.txt
```

Uploading csv files to S3 using sync (note that the output in this and the above image is cropped to save space)

```
C:\Anaconda\Natural Language Processing>aws s3 sync outputs s3://jmelde-bucket/p
s2_output
upload: outputs\6334221_trigrams.csv to s3://jmelde-bucket/ps2_output/6334221_tr
igrams.csv
upload: outputs\6334223_trigrams.csv to s3://jmelde-bucket/ps2_output/6334223_tr
igrams.csv
upload: outputs\6334221_unigrams.csv to s3://jmelde-bucket/ps2_output/6334221_un
igrams.csv
upload: outputs\6334222_unigrams.csv to s3://jmelde-bucket/ps2_output/6334222_un
igrams.csv
```