Natural Language Processing of Les Miserables using MapReduce - Brief ReadMe

Tahmeed Tureen

Fall 2017

This document briefly highlights what I did for this small scale project regarding the manipulation of text data and the usage of MapReduce. This is a coding assignment from SI 330: Data Manipulation, a course offered at the University of Michigan.

Purpose: To use MapReduce to count the number of bigrams in Victor Hugo's *Les Miserables* with the goal of reporting the frequency counts of the top 50 bigrams.

Definition: A bigram is a pair of consecutive written words.

Steps:

1. Download .txt file from Project Gutenberg
2. Convert the sentences in the book into single lines using nltk package in Python
3. Split the sentences into words and eventually collect the bigram pairs
4. Emit each bigram using the nltk package again
5. Count the number of times each bigram shows up in the text using the reducer
6. Sort the bigrams based on count
7. Exclude bigrams that include *stop* words like the, did, she, he, where etc.

(Q) : What are the top 50 bigrams, excluding those that contain stopwords, from Victor Hugos *Les Miserables*?

(A) : Refer to the .txt readme uploaded on this code repository