CS4000
Homework # 6: Hadoop Streams and Twitter n-grams

due Thursday, April 18th, 2019, 11:59 p.m.
(50 pts.)

Introduction

During the 2016 election season, a number of state-sponsored organizations were apparently attempting to influence the election by spreading specifically targeted, and often false information via Twitter. In 2018, Twitter released the following statement: "In line with our principles of transparency and to improve public understanding of alleged foreign influence campaigns, Twitter is making publicly available archives of Tweets and media that we believe resulted from potentially state-backed information operations on our service." For this assignment, we will be using Hadoop streams and map-reduce to analyze the data provided by Twitter. In particular, we will be calculating *n-grams* from this collection of data to see what patterns exist in the data provided by Twitter (see here for the full data set).

N-grams

According to Wikipedia, an n-gram is "a contiguous sequence of n items from a given sample of text or speech." Google captures n-gram frequency from a variety of sources (and languages) and uses them for a variety of purposes (e.g., natural language processing). Google's n-gram viewer (see here) allows users to plot how often words or phrases appear as a function of time. For this assignment, our speech sample will be filtered tweets where hashtags and web-references are removed, and words are split up via the standard punctuation.

The Twitter Data

Twitter provided several very large `.csv` files containing information about the tweets associated with the purported election interference. In this data, the data in each field is enclosed by beginning and ending quation marks, and fields are separated via commas. The text of an individual tweets appears in column 12 of the the `.csv` file. (Note that fields may contain newlines in their text. Hence, you cannot read the data using `getline`. However, you can use the `quoted` function in C++ to read each field.)

The assignment

For this assignment, you will use map-reduce to calculate the frequency (number of occurrences) of n-grams in the released twitter data. You will write two programs, a mapper and a reducer. The mapper will take a single input parameter from the command-line (n), it will read the Twitter data from the standard input and it will produce key/value pairs to the standard output. The reducer will take those key value pairs from the standard input and produce key value pairs where the key is an n-gram, and the value is the number of times that n-gram appears in the Twitter data.

You can run your program as follows: `cat twitter.csv | map 2 | sort | reduce`.

Submission

Submt both your mapper and reducer as single files via blackboard. Make sure that your code is adequately documented.