CS 4000

Homework # 6: Extra Credit — Secret Map Reduce Messages

due Friday April 23rd, 2021

(20 pts.)

Introduction

Email and other electronic media are often filtered for offensive language using fast, but relatively simple means (think finite automata). However, these simple filters are easily tricked. For example, if your filter was looking for the phrase "bad language," you could trick it by inserting characters between the letters, e.g., by writing "..b..a..d...l..a..n..g..u..a..g..e...". In this case, this text is easily recognizable by a human as the words "bad language", but harder (but not impossible) to recognize by a simple filter.

For this assignment, you are looking for potential "secret messages" embedded into tweets (both real and simulated). In particular, you are trying to find which Twitter users are posting "secret messages" to their followers. You discovered that, given a string, such as "secretMESSAGE", you can see whether those letters are embedded (in that order) in a tweet in $O(n)$ steps, where n is the number of characters in the tweet.

For this assignment, you will write a Hadoop streams map/reduce pair in C++ to process the data to look for secret messages in the Twitter users tweets. The purpose of your program will be to identify the Twitter users who tweet the secret messages, and how many times they tweeted those messages.

The Data Format

For this assignment, the Twitter data that you will be using has been pre-filtered to remove information about the Twitter user (`screen_name`) and the text of the tweet. Furthermore, tabs and newlines within the tweet have been converted to spaces. So, the data files for this assignment are provided in a line oriented format, where each line contains the screen name of a Twitter user and the text of tweet, where the screen name and the tweet are separated by a tab character `\t`.

The Mapper (10 pts.)

Write a C++ mapper that takes one command line argument — the secret message —, and outputs a key value pair ever time it finds a user how tweeted the secret message. Your program should read from the standard input and write a key/value pair to the standard output.

The Reducer (10 pts.)

Write a C++ reducer that reads key/value pairs from the standard input and writes key/value pairs to the standard output. You should assume that the key/values pairs are given in sorted order, and that you may receive values with more than one key. The final output of the reducer, when combined with the mapper should be the name of a Twitter user and a count of the number of tweets with the specified secret message.

Testing Your Code

Normally, we would run this on a Hadoop cluster. However, for this assignment, we will simply test the code in the UNIX shell by using the standard approach:

```
cat FILE | map SecretMessage | sort | reduce
```