# Project 3: MapReduce—Social Network Mining

COP 4610 — Operating Systems Principles

## 1. Summary

In the third project, we will write a MapReduce program to analyze the data collected from social networks. Dr. Z has a research project named EMPANADA, which stands for EMergency Preparedness Against NAtural DisAsters. The idea is to use the information available on social networks to help people be better prepared for disasters. For example, if there is flooding in a certain area, when people talk about it on a social network, that information can be automatically discovered and sent to other people so they can avoid the flooding area. The challenge is how to correctly and comprehensively discover the relevant information. You are charged to help Dr. Z address this challenge by creating a social network mining program using MapReduce. Although top projects will not get prizes as in the Cloud Hackathon, they can help Dr. Z's research and will be acknowledged in his future publications.

## 2. Description

Your will be provided two tweet datasets which were collected by Dr. Z's group last year as the input to your social network mining program, one is from the Hurricane Arthur, and the other is from the King Fire. You can look up their information on Wikipedia. A set of VMs will be provided to you, including one master VM and three worker VMs, which have Hadoop MapReduce preinstalled. Your program should satisfy the following requirements:

**Requirements:**

1) Your program should be named SocialMining. If you need two versions to handle Hurricane Arthur and King Fire separately, name them SocialMiningHurricane and SocialMiningFire.
2) Your program should identify tweets relevant to disasters (e.g., hurricane, flood, evacuation etc. for Arthur, and fire, smoke, evacuation for King) from the input data and save them in the output. Note that simply using grep to find a single keyword from the data will not get you far in this project because a word can bear different meanings depending on the context.
3) Your program should take a single input file stored on HDFS and store the output in a single file also on HDFS.
4) The output should be in text format and include only the text body of a single tweet per line. Do not include any other field of the raw tweet in the output.

## 3. Submission

You need to submit your project by May 3rd, 11:59:59pm. No late submissions will be accepted.

**Requirements**:

1) Submit a single zip file on Moodle. The file should be named by the full names of your

project team members. In the zip file, include the following:

a. The Java source code of your SocialMining program. Do not include any code that is not written by you.

b. A plain-text README file explaining how your program works, where it is stored on the master VM, and how to run it on the master VM that is provided to you.

2) Make sure that your assigned VMs are ready to be used to run your program by the deadline. Make sure to include the path to your program and instructions for running your program in the README file in your zip file submitted to Moodle.

## 4. Policies

1) Late submissions will absolutely not be graded (unless you have verifiable proof of emergency). It is much better to submit partial work on time and get partial credit for your work than to submit late for no credit.

2) Each group needs to work independently on this exercise. We encourage high-level discussions among groups to help each other understand the concepts and principles. However, code-level discussion is prohibited and plagiarism will directly lead to failure of this course. We will use anti-plagiarism tools to detect violations of this policy.