# COSC 4570/5010 Data Mining
# Spring 2020
# University of Wyoming
# Homework #5

**Due:** **May 11, 2020, 11:59 p.m. (Extended!)**

***Submission guideline*** You need to submit only one *.zip* file. Please name the file as "Your Net id_Homework5.zip".

## 1. Problems from the book (Mining of Massive Datasets 2nd Edition by Leskovec, Rajaraman, and Ullman)

Solve the following:

Chapter 2 (2.3.1)
Chapter 3 (3.2.1, 3.3.3, 3.3.6, 3.4.4)
Chapter 4 (4.3.2, 4.3.3, 4.4.1, 4.5.3)

## 2. Bloom Filter

Implement the Bloom filter. Use:

http://www.stopforumspam.com/downloads/listed_username_30.zip

as your set S. This is a set of usernames known to be spam for the last 30 days. Select a proper hashing memory size (n) and find the optimal number of hash functions (k). Use the spam usernames for the last 365 days:

http://www.stopforumspam.com/downloads/listed_username_365.zip

as your stream. Submit your (1) code, (2) optimal k for your n, and (3) the percentage of false positives. There is no need to submit your datasets.

**Note.** For hashing you can use

- murmurHash: https://sites.google.com/site/murmurhash/
- FNV: http://isthe.com/chongo/tech/comp/fnv/
- Jenkins Hash: http://www.burtleburtle.net/bob/hash/doobs.html
- Or a Hash Function of your choice.

## 3. Flajolet-Martin (FM) algorithm (Takes Time!)

Implement the Flajolet-Martin (FM) algorithm. Count the number of distinct quotes (quotes are denoted with lines that start with Q) in the MemeTracker dataset (all files):

https://snap.stanford.edu/data/memetracker9.html

Submit (1) the estimated number from FM and (2) your code. In your implementation, use the method discussed in Section 4.4.3 to provide more accurate results. Do not to submit your datasets!

Note: The dataset is about 50-60 GB uncompressed. Let me know if you face an issue downloading the same.