NYU CS6513 Big Data **Fall 2022**

Assignment 1 – Hadoop

80 Points (+ 20 points extra credit)

Due Date: Thursday, September 22, 11PM Eastern

Submit your homework **as a zip file, with your name on it: e.g. hw1-jcr365.zip**

Abstract: Show proficiency using HDFS and writing a MapReduce program, including submitting to Hadoop and getting results out of HDFS.

1. Hadoop command line HDFS - 15 points

For each item in the list**, submit screen grab** or suitable capture (a picture in jpg or other suitable format):

a. Login to Dataproc or Hadoop cluster and find the Hadoop version ($hadoop version)
b. Create a new directory path as 'hw1-<netid>/input' , where netid is your netID. For example, my path would be 'hw1-jcr365/input'. Show the directory.
c. Extract and copy the homework input files to HDFS in the 'input' directory of part b. Show the files.

2. MapReduce 65 Points

Modify the MapReduce WordCount code from class (Java or Python) to compute the **n-*grams*** counts. An **n-gram:** A sequence of n words from an input line.

Write separate programs or one parametized program to:

a) Output n-gram count for n=1 (single words)	5 points
b) Output n-gram count for n=2 (bigrams)	30 points
c) Output n-gram counts for n=3 (trigrams)	30 points

- All text should be lowercase
- Punctuation does not count; so the words is '(1991)' and '1991'are the same.
  You must parse and remove/replace with space all characters not in this set: [a-z0-9]
- If a line has less words than the n, then skip that line

For example, for the sentence "`Extract and (copy) the input from the`",
the output of the mapper will give the following 1-gram and 2-grams:

1-grams:	extract,1	2-grams:	extract and,1
	and,1		and copy,1
	copy,1		copy the,1
	the,2		the input,1
	input,1		input from,1
			from the,1

Input: hw1.txt (provided in Brightspace)

3. Extra Points – 20 points

Each line in the input data is of the form: **DocumentID Text**.

Furthermore, the Text has paragraph formatting commands embedded in it. Paragraph formatting commands are of the form: **<x>,** where x can be H,P, etc.

TODO: Repeat the bigrams in part Q2, but this time parse the input lines to only look at the text portion, with all formatting commands of the form <x> removed.

For example, for a line of text that looks like:

```
@@4004141 Guest Editors <p> There can only be disaster arising from unawareness of the causalities…
```

`The processed line should be:`

```
guest editors there can only be disaster arising from unawareness of the causalities…
```