COSC 589 - Web Search and Sense-Making

Assignment 2
Due Wednesday 1/31/18, 11:59pm

Task: WordCount and Spark UI

Introduction:

The word count problem is to count how often each word appears in a text document, or a collection of text documents.

```
val file = sc.textFile (“./testFile”)
val counts = file.flatMap(line => line.split(“ “)).map(word=>(word,1)).reduceByKey(_+_)
counts.collect()
```

In this assignment, you will play with the word count problem and extend it.

Instructions:

1. In Piazza, we provide two files for you, one.txt and two.txt. Download them to your local machine.
2. Write a WordCount.scala program to produce the following about the files.

   - Print the total number of words in one.txt
   - Print the total number of unique words in one.txt
   - Get the word counts for both files. That is, each word and each word’s number of occurrences, from both files. Save the word counts into an output file (which will actually be a directory), with the name “wcOutput”

   For instance, if the content of one.txt is ‘'I love Spark spark is cool'' and the content of two.txt is ‘'i am learning spark now’’, we expect to see the word counts are:

```
i 2
spark 3
love 1
is 1
learning 1
now 1
cool 1
am 1
```

3. Compile your program into a standalone package using “sbt package”.

4. Follow the lecture notes, set up your Spark master and worker node on your laptop. Show that you could monitor your program using Spark UI.

5. Print the content of the mapped RDD in the Lecture Notes in a nice format. In the lecture notes, we have shown that printing the content of a flatMapped RDD is easy. The codes are :

```
val lines = sc.parallelize(List (“hello”, “how are you”))
val words = lines.flatMap ( x => x.split(“ “))
words.foreach(println)  // what do you get?
val mapwords = lines.map ( x => x.split(“ “))
mapwords.foreach(println)  // what do you get?
```

Can you do it for a mapped RDD? Take a screen capture of your code and results

What to Submit:

- Your code
- Screen capture of the results
- Screen capture of the Spark UI of your jobs (see how to get Spark UI running in the lecture
- notes)
- The two output files located at wcOutput/part-00000 and wcOutput/part-00001

Where to submit:

- Canvas

When:

- Due 1/31/2018, 11:59pm.