COSC 589 - Web Search and Sense-Making

Assignment 2
Due Wednesday 2/1/17, 11:59pm
Submit to Blackboard

Task: WordCount and Standalone Program

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
```

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

6. Submit the job to your Spark Master.

What to Submit:

- Your code
- Screen capture of the results, including your commands “sbt package” and “spark-submit”, then the results.
- The two output files located at wcOutput/part-00000 and wcOutput/part-00001

Where to submit:

- Blackboard

When:

- Due on 2/1/2017, 11:59pm.