CMPSC 497B/CSE 597B HW4. DUE 10/27

Instructions:

You cannot share code with other students or view code written by other students. Do not post code online. Violation = Academic integrity violation.

Your submission will be a zip file containing:

- Your pig script(s)
- Your answers in the submission text box on ANGEL.

1. Version for CSE 597B and CMPSC 497B Honors Option

Problem 1. *One common task in social network analysis is to count the number of triangles in a graph. A common algorithm is to first create an undirected representation of the graph and join it to itself to get all undirected paths of length 2 (where the first and last nodes are different). It is easy to see that the paths:* $A - B - C$, $A - C - B$, $B - C - A$, $B - A - C$, $C - A - B$, $C - B - A$ *exist if and only if* A, B, C *form a triangle. Thus the algorithm proceeds to find triples of nodes* (A, B, C) *that together participate in 6 paths. [Hint: this algorithm will use many of the Pig constructs you have learned]*

- *Implement this specific algorithm in Pig.*
- *Apply it to the dataset graph.txt and put the answer in the submission textbox on ANGEL.*

Problem 2. *The file docdata.txt has the format [document id:][tab][words]. Write a Pig script that computes that top 10 documents (i.e. the 10 longest documents) and for each word in these top 10 documents, count how many of these top 10 documents contain that word. [hint: look at additional functions/commands in the Pig documentation]*

- *Write a pig script that computes the answer. The output should be sorted in descending order by count, and ties are broken in alphabetical (increasing) order.*
- *In the zip file, include the answer you get on the file docdata.txt*

2. Version for CSE 497B

Problem 3. *The file docdata.txt has the format [document id:][tab][words]. Write a Pig script that computes that top 10 documents (i.e. the 10 longest documents) and for each word in these top 10 documents, count how many of these top 10 documents contain that word. [hint: look at additional functions/commands in the Pig documentation]*

- *Write a pig script that computes the answer. The output should be sorted in descending order by count, and ties are broken in alphabetical (increasing) order.*
- *In the zip file, include the answer you get on the file docdata.txt*