Q1. (5 pts) Consider the following fragment from a collection of documents. Assume that these three documents are the only documents in the collection that contain the words “dog” or “cat”.

Document id	**Document Text**
234569	The phrase "fight like cats and dogs" reflects a natural tendency for the relationship between the two species to be antagonistic. However, sometimes the two species can be friends.
234578	Dogs and cats can have a bad relationship.
234839	Cats are furry.
234879	Dogs are man's best friend.

1. Write the inverted index posting lists for terms “cat” and “dog”.
2. Write the compressed form of the posting list for dog.
3. Calculate the cosine similarity of each of the above documents with query
   q: cat and dogs.
   Use stemming and remove stop words from the documents and query. Use only TF (ignore IDF).

Q2. (4 pts) Design MapReduce algorithms to take a very large file of integers and produce as output:

(a) The largest integer.
(b) The average of all the integers.
(c) The same set of integers, but with each integer appearing only once.
(d) The count of the number of distinct integers in the input.

Q3. (3 pts) Write MapReduce algorithms for computing the following operations on bags R and S:

(a) Bag Union, defined to be the bag of tuples in which tuple t appears the sum of the numbers of times it appears in R and S.

(b) Bag Intersection, defined to be the bag of tuples in which tuple t appears the minimum of the numbers of times it appears in R and S.

(c) Bag Difference, defined to be the bag of tuples in which the number of times a tuple t appears is equal to the number of times it appears in R minus the number of times it appears in S. A tuple that appears more times in S than in R does not appear in the difference.

Q4. (3 pts) Suppose we take a star join of a fact table $F(A_1, A_2, \ldots, A_m)$ with dimension tables $D_i(A_i, B_i)$ for i = 1, 2, …, m. Let there be *k* Reduce tasks, each associated with a vector of buckets, one for each of the key attributes $A_1, A_2, \ldots, A_m$. Suppose the number of buckets into which we hash A_i is a_i. Naturally, $a_1a_2\ldots a_m = k$. Finally, suppose each dimension table D_i has size d_i, and the size of the fact table is much larger than any of these sizes. Find the values of the a_i’s that minimize the cost of taking the star join as one map-reduce operation.

Q5. (6 pts) Implement the following tasks using RDDs in PySpark.

Consider the movielens small dataset; see: https://grouplens.org/datasets/movielens for a description of the dataset.

Here is the zip file: http://files.grouplens.org/datasets/movielens/ml-latest-small.zip

1. (2 pts) Find the average rating for each user and movie.

   Use the “ratings.csv” file.

2. (4 pts) Find the average rating for each genre.

   Use the “movies.csv” file to find the genre of each movie.

   This file is assumed to fit in the memory of a machine; it is cached locally.

   A movie can belong to multiple genres.

   A rating of a movie should be used for all the genres the movie belongs in.

Submit all your source code and your outputs.