Programming Project – TF-IDF Computation Using Map-Reduce

Advanced Data Intensive Computing (CSCI 8790)

In this project, you will implement TF-IDF computation using the map-reduce paradigm. TF-IDF is a very common metric used in information retrieval to estimate the importance of a word to a particular document (https://en.wikipedia.org/wiki/Tf-idf). TF-IDF and its variants are used in search engines, document mining etc.

As the name suggests, TF-IDF is a combination (product) of two different scores. TF (term frequency) indicates how many times a particular term or word appears in a document. IDF indicates the inverse of how many documents a particular term appears in. Formally, suppose the corpus has N documents $\{d_1, d_2, \ldots, d_n\}$. Let $tf(t_x, d_y)$ represent the number of times the word tx appears in document d_y and let $df(t_x)$ represent the number of documents that t_x appears in the corpus. $TF\text{-}IDF(t_x, d_y)$ is calculated as follows.

$$Tf\text{-}IDF(t_x, d_y) = tf(t_x, d_y) * \log_2(N/df(t_x))$$

map-reduce-corpus.tar.gz contains 556 files containing Imdb movie reviews. This is a small subset of files from the Imdb data set (complete dataset is available at http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz)

Each file is named with the following format "pqr_xyz.txt", where pqr and xyz are numbers. The directory also includes a vocabulary file vocab.txt. Please note however, that since the corpus we are providing is a small subset of files, all the words that are included in vocab.txt may not appear in the corpus.

In your project you will write a map-reduce program that produces the following output. Let tx be an arbitrary word in the corpus and let dy be an arbitrary document in the corpus. Your project should produce t_x d_y $TF\text{-}IDF(t_x, d_y)$. Please note that if a term is not in the corpus (even if it is in the vocabulary file), your program should not include it in the final output.

You can choose to merge the input files or include a listing of the input files as an additional input to your program. Many aspects of the project description are intentionally kept vague. This is because, I want you to make suitable assumptions and design your project.

What to submit and how to submit:

You will submit a directory containing your program, the output file and a detailed documentation. The documentation will explain the high-level logic of your program (including the key-value pairs used etc.). It will also contain all the assumptions made in your program.

Project submissions will be done via elc. You will also do a demo (a signup sheet will be made available) to me.

Project Due Date: October 14, 2016.

Points to note:

1. If your program does not confirm to the map-reduce paradigm, your project will get a zero (irrespective of the correctness of the output).
2. No unauthorized assistance of any form (using code from the Internet, copying code from someone else, etc.). Any violations will be regarded as academic dishonesty. When in doubt talk to the instructor.