CA675: Cloud Technologies

Assignment 1: Data Analysis

School of Computing, Dublin City University

Assignment 1: Data Analysis

1. Get data from Stack Exchange
2. Load them with PIG
3. Query them with Hive
4. Calculate TF-IDF with MapReduce
   (Note: plenty of versions of code online in both Java and Python, just acknowledge the source and the changes you had to do to it)

Assignment 1: Data Analysis

1. Acquire the top 200,000 posts by viewcount (see notes on Data Acquisition)
2. Using Pig or MapReduce, extract, transform and load the data as applicable
3. Using Hive and/or MapReduce, get:
   1. The top 10 posts by score
   2. The top 10 users by post score
   3. The number of distinct users, who used the word “Hadoop” in one of their posts
4. Using Mapreduce calculate the per-user TF-IDF (just submit the top 10 terms for each of the top 10 users from Query 3.II)

Submission via LOOP

Please submit a *single zip file* with the following:

- The source code (python, java, pig, hive) for each of the tasks, or a link to a github where the code can be found (to be added into your documentation).
- Evidence of the results for each task (e.g. screenshots) and documentation for your code (max 3 pages).

Assignment 1: Data Analysis

- Submission open: 24th October
- Due date: 8th November
- Submit 1 zip file as per instructions
- Worth 10% of the final mark
- Assessment criteria:
    a) Task completion quantity
    b) Task completion quality

Assignment 1: Assessment guidelines

Criteria	**0-12**	**25**	**36**	**50**
Task completion quantity	One task or less fully completed	Two tasks fully completed	Three tasks fully completed	All tasks completed
Task completion quality	Major errors	Several minor errors	Only a few minor errors	No errors