|  |
| --- |
| ***Team Analytics Project Proposal – Second Draft*** |
| ***Part 1. General Information*** |
| Team Name (optional):  Team Members:  Yi-I Chiu (yic211), Po-Chih Lin (pcl322), Yu-Chih Yu (ycy247)  Project Title:  Semantic Information Retrieval System  Project Description (short description of your project):  The objective of this project is to build a system which can perform search on the semantic content of texts (using Wikipedia).  Semantic retrieval differs from normal information retrieval in the following aspect: Normal information retrieval system only returns the documents (contents) which contains the exact query words. Semantic retrieval systems, on the other hand, use the content of top ranked result to perform another round of search.  For example, if the query word is ``White House’’, a semantic retrieval system could also results include only ``President Obama’’ but not `White House’’.  The project will be divided into four main parts: (1) parsing the html data to text (2) create the inverted-index of the content for retrieval (3) creating the PageRank (4) performing semantic expansion (5) online querying procedure.  (1) can be a simple Map function which take the (document id, html content) and produce (document id, text)  (2) is an offline MapReduce procedure which the Mapper takes (document id, text) and produce (word, document id), the Reducer takes (word, document id) and produce (word, vector(document id, term frequency))  (3) is an offline MapReduce procedure that generate the PageRank of the Wikipedia pages.  (4) is an offline MapReduce procedure that takes each document in the database, calculate the similarity of the document and expand the inverted index set.  (5) is an online procedure which perform DataBase like retrieval from the stored PageRank and inverted-index. |

|  |  |  |
| --- | --- | --- |
| ***Team Analytics Project Proposal – Second Draft*** | | |
| ***Part 2. General Data Source Information*** | | |
| **Data Sources**  - E.g. tweets | **Data Source Description (brief)** | **Data Size**  Estimate size, e.g. MB? GB? TB? |
| 1. Tweets | Tweets downloaded from Twitter | GB ~ TB (Depend on how much data we download) |
| 2. Wikipedia | The English Wikipedia | 48 GB (The size of Wikipedia Corpora) |
| 3. |  |  |
| 4. |  |  |
| 5. |  |  |
| 6. |  |  |
| ... |  |  |

|  |  |  |
| --- | --- | --- |
| ***Team Analytics Project Proposal – Second Draft*** | | |
| ***Part 3. Detailed Data Source Information*** | | |
| **Data Sources**  - From Part 2 above | **Data Characteristics**  - Is data source a realtime source?  - Is it realtime and stored (e.g. a log)?  - Is it statically loaded data (e.g. historic)? | **Data Frequency**  - If realtime data, what is the frequency? |
| 1. Tweets | Realtime Source. Not realtimely stored (Need to crawl). | Every milli-second (but we won’t get data at this rate) |
| 2. Wikipedia | Non-realtime Source (Or changing very slowly upon modification, but changes is not the main target to model.) | N/A |
| 3. |  |  |
| 4. |  |  |
| 5. |  |  |
| 6. |  |  |
| ... |  |  |

|  |
| --- |
| ***Team Analytics Project Proposal – Second Draft*** |
| ***Part 4. Technologies***  List technologies. Will your project make use of MapReduce? Pig? Flume? HBase? Hive? Impala? Spark? Mahout? Other? |
| Our project will be using MapReduce, HDFS, Hive and Spark on creating the inverted index and performing query. |
| ***Part 5. References***  References – Please add references to all papers/articles read by the team (should be at least two references per team member). |
| 1. Baban, Hediyeh, S. Kami Makki, and Stefan Andrei. "Comparison of Different Implementation of Inverted Indexes in Hadoop." *The Second International Conference on E-Technologies and Business on the Web (EBW2014)*. The Society of Digital Information and Wireless Communication, 2014. 2. Gu, Chunhao, and Yang Gao. "A content-based image retrieval system based on Hadoop and Lucene." *Cloud and Green Computing (CGC), 2012 Second International Conference on*. IEEE, 2012. 3. Logothetis, Dionysios, and Kenneth Yocum. "Ad-hoc data processing in the cloud." *Proceedings of the VLDB Endowment* 1.2 (2008): 1472-1475. 4. Chu, Cheng, et al. "Map-reduce for machine learning on multicore." *Advances in neural information processing systems* 19 (2007): 281. 5. Lin, Jimmy, et al. *Of Ivory and Smurfs: Loxodontan MapReduce experiments for web search*. MARYLAND UNIV COLLEGE PARK DEPT OF COMPUTER SCIENCE, 2009. 6. Busch, Michael, et al. "Earlybird: Real-time search at twitter." *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE, 2012. |