Advanced Database Systems

Project 1

Team members: PING-FENG LIN, pl2730, Weihan Chen, wc2681

- List of the file: IR.py, Readme.pdf, Query transcript.pdf
- Description of how to run the program:
  1. Under the directory enter: python3 IR.py <1 or 2> <query> <precision>
  2. Install all the dependent package (nltk, bs4, sklearn)
  3. Install all the package of nltk (only for the first time, latter you can comment nltk.download())
  4. Answer (Y/N) for all the results shown in the screen sequentially
  5. See the result

- Internal Design of the project:
  We use the information from the website we consider relevant to refine the query. Web crawling skill and tf-idf technique are utilized to achieve the goal. We also implemented a simple way to reach the desired performance, by only adding the most frequent word of the title and snippet of relevant websites.

- Description of the query modification method:
  We obtain the description and the website content from the results we think is relevant. Calculate a tf-idf matrix by using these contexts and find the word with the highest summation of weights. Add this word into the query.
  In detail, we tokenized the context and simultaneously reserve the original collection of words. When a candidate token is obtained by calculation, we check if it already exists in the query. If not, we then choose from the original collection of words to retrieve the most common word related to this token. Otherwise, we choose the token with second highest summation weights, so on and so forth. For the query order, we simply append the new word into the end of the current query.

  However, the performance is not as well as the simplest way ☺, so we provide 2 method; method 1 is the simple way, method 2 is the proposed way.

- Search Engine ID: 000568897140034501999:kyspcqzrszq

- JSON API key: AIzaSyBKlE390a0krI3_Oe-zk-_Pdf7J0J7Oo1I