
UNSW SYDNEY

# COMP 6714 Project Part I

**Student ID: z5197332**
**Student Name: Jiling Yang**

**Project Description:**
This project is required to compute $TF\text{-}IDF$ score of a document given by an input query and a dictionary of entities.

**Solution Description:**
Class InvertedIndex has three functions to implement this project.

Index_documents function uses only one for loop to generate token and entity index in the document. The index of the token is generated firstly followed by generating the index of the entity. Finally, I delete the token index with zero values during formalizing the index of the entity.

Split_query function utilizes four steps from specification to achieve splitting query by using combination of the query and a filter subset. According to the specification, four steps were used to get final query results. One edge case of these 4 steps is that for each subset the token count cannot exceed the corresponding token count in query. A flag is set to figure whether final entities can append the subset of filter subsets or not by using counter method.

Max_score_query function generates tf_idf_entity and tf_idf_token dictionaries to store $TF\text{-}IDF$ score of entity and token and calculates the max score of all splitting queries.

**Major Implementation Issues:**
The most difficult part of this project is to reduce total time cost. I import time library to calculate runtime of each function and optimize time complexity. The first submission raises TimeoutError because of many nested loops. The first solution I have tried is to swap the sequence of indexing entities and tokens and reduce nested loops by only using one outside loop.

**Lessons Learned:**
It is not difficult to output the results required for the entire project. The first thing I have learned is that when the dataset is huge, the main problem is to focus

on how to reduce the time complexity to optimize the time the project takes to run the entire project. Next, the edges cases need to be considered when completing the project. According to specification, indexing document can be processed in two steps, and splitting the query can be processed in four steps. Finally, by reducing the use of for loops, the efficiency of programming can be greatly improved.