Report for search engine project:

My project consists of four parts. The first one is Parser. Parser is used to read all files from the raw source and tokenize the resource and save those files into my own format. The second part is indexer, which used to read all the source that produced by Parser and build the index by tf-idf. The third part is judge. This is the most import function in my project. Judge will read the index that created by indexer and other file by parser to produce Cosine similarity by terms. Parser also get query and produce the result. Tinker is the part doing I/O and implement the web interface. The whole project is done by myself and the only third party API I used is beautiful soup to tokenize the source from the raw content.

Alan Xie