

# Read Me

There are two Python files need to run. One is **Preprocess\_PositionalIndex.py** while the other is **SearchEngine.py**

- **Preprocess\_PositionalIndex.py**

Put collection file (trec.5000.xml) with this Python file into same folder. Then just run: python Preprocess\_PositionalIndex.py in the shell. Then wait for about 4-5 mins, you will get a json file called **pos\_inv\_stem\_index\_5000.json** which contains word positional inverted index needed by **SearchEngine.py**. Note that this json is not the same as index.txt with their content. The **json file generated is for searching** while the **index.txt is just for submitting**.

- **SearchEngine.py**

Put **pos\_inv\_stem\_index.json** with **SearchEngine.py** in the same folder and run: python SearchEngine.py. Then you need to input the number of queries you want to input (How many queries for Boolean Search and Proximity Search; How many for TFIDF Ranking, separately). And input the query one by one, you will get what you want.

PS1: I contain **pos\_inv\_stem\_index\_5000.json** in my submission folder so you **do not need to generate it again** from collection by using **Preprocess\_PositionalIndex.py**.

PS2: **STRONGLY RECOMMENDING** you to run my search engine by tying:  
python SearchEngine.py