Information Retrieval Systems

Lab Practical and date – Practical 9, Thursday 5th May 2022

Name and Roll Number- Naman Thaker , 20BCE529

Practical Objective- Develop GUI based mini search engine.

Steps Involved

1. The corpus consisted of 10 documents, with first 5 documents consisting of text regarding blockchain while the next 5 contained information regarding AI
2. All the information is read and preprocessing steps are applied, these steps include tokenization, stemming, case conversion, word count frequency and stop word removal
3. The term-frequency and inverse term-frequency is calculated and stored in a json file
4. A GUI is made using TKInter in python which takes the input search query from the user and gives the search engine output
5. The preprocessing step are performed on the input string as well and the is compared to the TF-IDF data stored in the json file
6. The similarity index is calculated, and the relevant documents are printed in a ranked manner

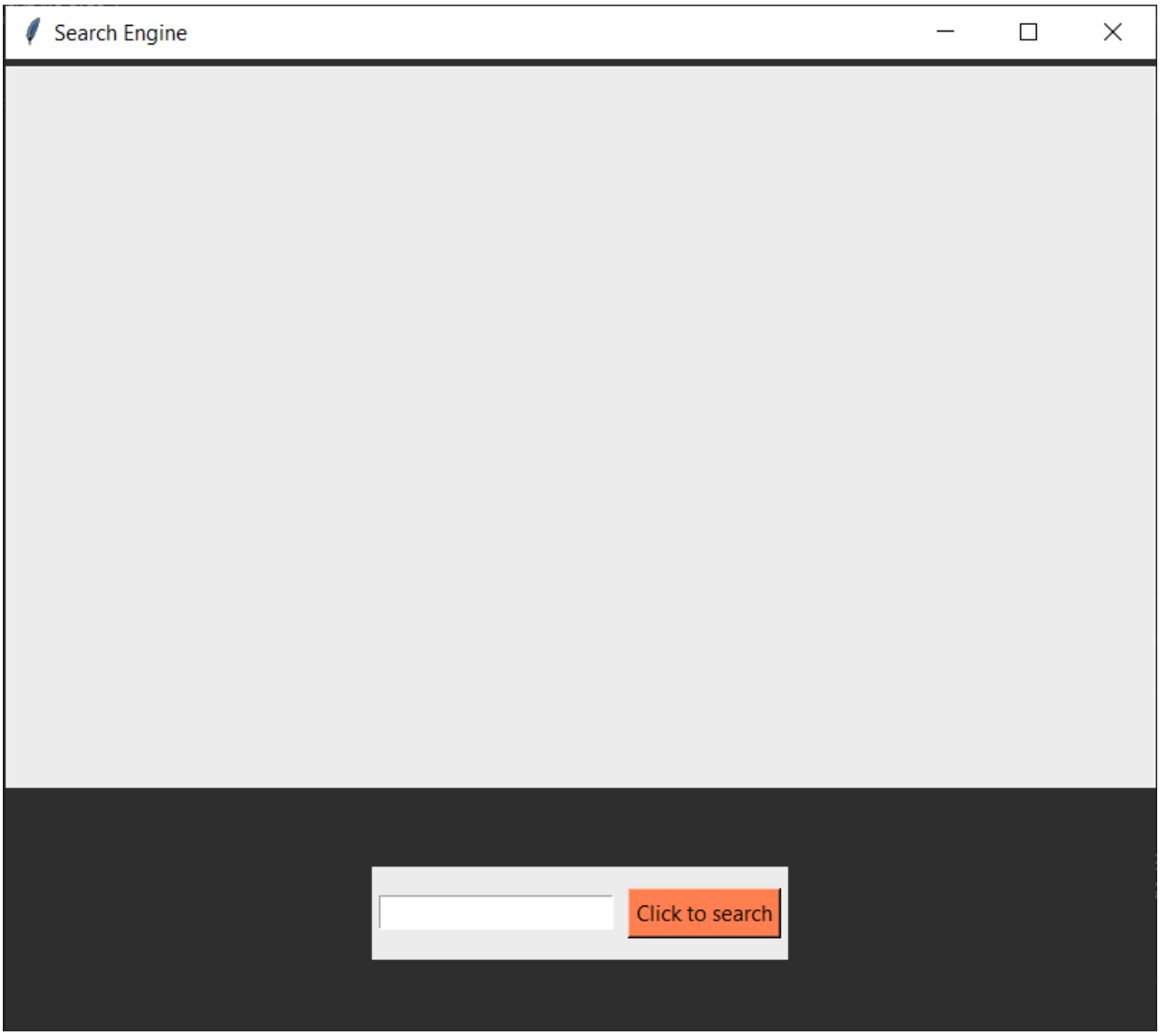
Python Package Used

- Tkinter-Tkinter is a Python binding to the Tk GUI toolkit. It is the standard Python interface to the Tk GUI toolkit, and is Python's de facto standard GUI. Tkinter is included with standard Linux, Microsoft Windows and Mac OS X installs of Python. The name Tkinter comes from Tk interface
- NLTK- The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

Sample Input/Output

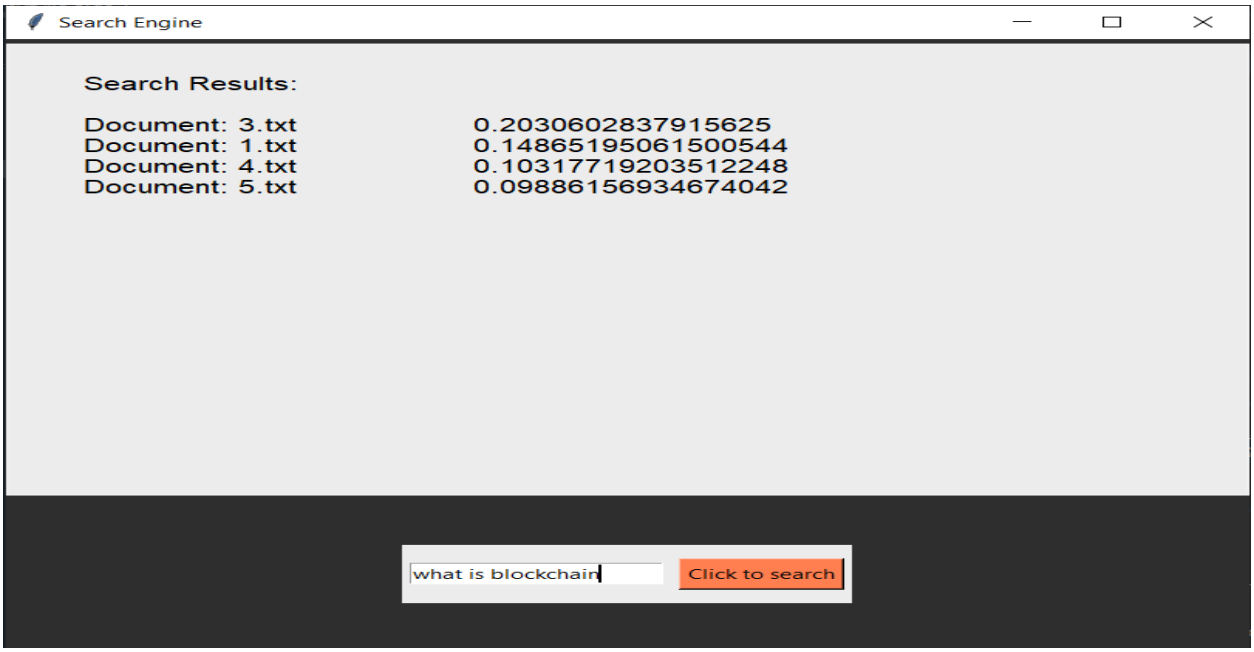
The input is in the form of corpus of a files on which the text preprocessing was done. The code in by the name of search engine.py

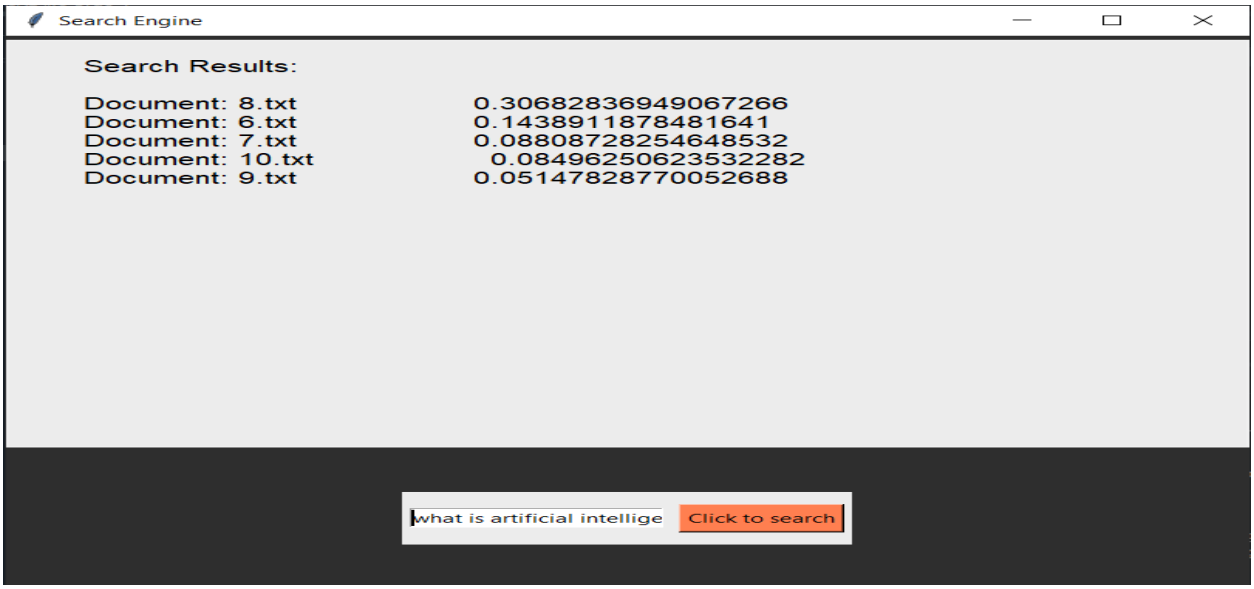
The output is presented in the form of ranked documents which contain information about the search query



Sample GUI-Window

Sample Search Operation on String of Words





Conclusion

In this practical, we used to create a GUI and implemented a mini-search engine in python using TF-IDF. We learned GUI programming in python and reading and outputting the files