This is the readme for the final project of the CS582 course by Piervincenzo Ventrella.

In the folder I provide:

Crawler.py file, this program crawls the pages from the uic domain. I provide also the notebook version, if you wish to take a look at the code you're free to run the executable or open the notebook.

Inverted_Index.py file, this program builds the inverted index. I provide also the notebook version, if you wish to take a look at the code you're free to run the executable or open the notebook.

Finally, I provide the **Retrieval.py** file that will load all the needed documents from file (such as Inverted_index, document norm dictionary, urls dictionary, …), These documents are created in the executables mentioned above but they will require time, so I already provide those files to directly run the Retrieval.py executable. In this file the 2 WordToVec models I used for query expansion are loaded and they may require a couple of minutes to load. After that a Graphical User Interface is launched and will be possible to interact with the System. Other useful information concerning the execution of the program are printed on the command line.

For the project I used 2 Word2Vec models: 1 is provided by google "GoogleNews-vectors-negative300.bin" and the other one is trained on the collection. In the notebooks PreBuiltWordToVecModel.ipynb and OwnWordToVecModel.ipynb I experiment them and save into file. Yoa are free to take a look at them but I already provide the 2 models, so it will be not necessary to run this notebooks.

You will probably need to install the nltk library for python and the genism library that is responsible for the WordToVecModels. Initially I had some issue in downloading and using the model provided by Google but now I directly provide it so there should not be any problem. In case there is some problem with the model offered by Google, I found useful information in the link below:

https://radimrehurek.com/gensim/models/word2vec.html

After inserting the query, you are free to use 3 different options for the search (check the report). After the matching document containing at least one word are retrieved, the System starts to rank them (if the documents are a lot it could take a while), After the system complete this operation a new GUI is displayed in order to show the results to the user.

Important note: because of the size of the model I coud not load the Google model on blackBoard so, to run the Retrieval.py file you need first to download "GoogleNews-vectors-negative300.bin" from the folder provided below and put it in the same folder of the project.

Link folder:

https://drive.google.com/open?id=1-G13h3NY3eBSv2XLP3i8kGAndvNAYhNk