HOMEWORK 2 BIG DATA COMPUTING

ANDREA MORELLI 1845525

I attached the jupyter notebook that can be executed on colab.
It can also be reached here :
https://colab.research.google.com/drive/1dMrW_DrpzEwoS69RgTZruEL6Wfau1H0X#scrollTo=8mftCoHcI5D-

maybe there is the need to ask for the access.

The code is all commented and the testings are also detailed with descriptions.
To execute the code just put it on colab and get a key from kaggle to download the dataset directly.

All the code until the last part where there are the tests on the entire dataset can be executed sequentially.
Once you reach the last part there is the need of all the memory so you have to restart the session and reread the dataset preprocessed that is saved previously.
To execute the last function that calculate the entire set A there is the need to restart again the session as described above.

To summarize i implemented the lsh both in a normal and in a vectorized way.
The vectorized class implements the functions to calculate B and A in a vectorized way and are waaaay faster than the normal counterpart.

The hash function used in the vectorized lsh is similar to the one proposed but i just take the sum of the elements obtained by the matrix multiplication in order to vectorize everything.

On the entire dataset if the threshold is >= 0.1 than we can evaluate all the A and if we chose r and b that dont take a lot of documents with low cosine similarity we can also compute B.
There is this problem because if we have a lot of similar elements for every document than the memory consumed is a lot.

Just in the vectorized way its possible to do those calculations because the normal implementation is orders of magnitude slower.

the time to compute A vectorized is 74 seconds.
the time to compute B vectorized is something like 1800 seconds.

it’s easier to compute A just because in B i couldnt vectorize everything.
A is computed with block matrix multiplication because the entire test matrix couldn’t be used to calculate the multiplication so i just did it in blocks of 300 elements of the test set.

Every detail on the implementation and a guide to the tests done are in the jupyter notebook that can be loaded in colab where everything works fine.

For every doubts on the execution of the code contact me

Andrea Morelli 1845525