ASSIGNMENT 1 :-

Code file 1 :- invidx_cons.py
Functions :-

1. compressionC1 :- Compresses the postings lists using the C1 method and stores it in the .idx file and the dictionary corresponding to it in .dict file, it takes the postings list of C0 method as a refrence and then it compressess ater reading from that helper posting list of C0 method.
2. compressionC2 :- Compresses the postings lists using the C2 method and stores it in the .idx file and the dictionary corresponding to it in .dict file, it takes the postings list of C0 method as a refrence and then it compressess ater reading from that helper posting list of C0 method.
3. compressionC3 :- Compresses the postings lists using the C3 method and stores it in the .idx file and the dictionary corresponding to it in .dict file, it takes the postings list of C0 method as a refrence and then it compressess ater reading from that helper posting list of C0 method.
4. compressionC4 :- Compresses the postings lists using the C4 method and stores it in the .idx file and the dictionary corresponding to it in .dict file, it takes the postings list of C0 method as a refrence and then it compressess ater reading from that helper posting list of C0 method. “k” is given in the merge function and then stored in the dictionary after compression.
5. decodeC0 :- It returns the list of the doc numbers in which the input term is present after looking at the dictionary and posting list formed using C0 method.
6. decodeC1 :- It returns the list of the doc numbers in which the input term is present after looking at the dictionary and posting list formed using C1 method.
7. decodeC2 :- It returns the list of the doc numbers in which the input term is present after looking at the dictionary and posting list formed using C2 method.
8. decodeC3 :- It returns the list of the doc numbers in which the input term is present after looking at the dictionary and posting list formed using C3 method.
9. decodeC4 :- It returns the list of the doc numbers in which the input term is present after looking at the dictionary and posting list formed using C4 method.
10. write_to_file :- It keeps on writing posting lists in different files (which we can get from the variable filenum) into a directory known as FilesToMerge which is removed after merging all these files, it writes in the posting lists using strategy C0.
11. final_merge :- First of all it takes the files and merges them just following C0 in a helper file, now it looks at the compression technique from the dictionary and writes in the main index file and dictionary file with the corresponding compression technique using the helper index file and dictionary file. At the end it removes the helper files.
12. process_string :- To see the place where 0 occurs in Unary string, so as to find the number represented by it.
13. helper_function1/2/3/4 :- Functions just to see whether the result of all the info retrieval using various compressions is same or not.
14. Clean :- To clean the filesToMerge directory

Code file 2 :- search.py
Functions :-

1. intersection_of_2_lst. :- Takes intersection of 2 integer lists, where integers are the doc numbers
2. intersection_of_lst. :- Takes intersection of a list of integer lists, where integers are the doc numbers
3. query_search. :- Searches for all the queries in the query file using suitable decode methods and stores the result in the result file
4. query_result :- Returns a list of the docids in which the query is present

METRICS OF INTEREST :-

1. **Index Size Ratio** :-

Collection Size	C0	C1	C2	C3	C4
200 MB	0.280	0.122	0.143	0.231	0.118
100 MB	0.301	0.145	0.158	0.240	0.123
10 MB	0.354	0.157	0.174	0.277	0.136

2. **Compression Speed** (time taken in milliseconds) :-

Collection Size	C1	C2	C3	C4
200 MB	67000	81034	18431	58300
100 MB	31000	34056	9876	29700
10 MB	2471	3413	701	2349

3. **Query Speed** (time to retrieve the query in milliseconds) :-

Collection Size	C0	C1	C2	C3	C4
200 MB	280.9	312.9	2304	2567	265.8
100 MB	25.8	29.9	290.5	311.3	25.6
10 MB	2.25	2.46	20.2	33.33	2.31

Submitted By :-

-> Nikita Bhamu
-> 2018CS50413