Program Characteristics:

- It took the program 18.25 seconds to run completely.

- It took 0.537 seconds to create the index. This was trivial as I created many HashMaps to reduce the overload of creating the index. Each term was taken from each HashMap to create both the compressed as well as the uncompressed index.

- The size of the uncompressed index is 1522639 Bytes.

- The size of the compressed index is 259279 Bytes.

- There is a reduction of 82% in the index size.

- The number of inverted lists in the index is 5713.

- The Df, Tf and the inverted list length (in bytes) for the words given in the homework as follows:

Term: Reynolds
Df: 200
Tf: 384
Inverted List Length: 411 Bytes

Term: NASA
Df: 145
Tf: 148
Inverted List Length: 298 Bytes

Term: Prandtl
Df: 63
Tf: 80
Inverted List Length: 135 Bytes

Term: flow
Df: 730
Tf: 2080
Inverted List Length: 1468 Bytes

Term: pressure
Df: 551
Tf: 1382
Inverted List Length: 1113 Bytes

Term: boundary
Df: 467
Tf: 1185
Inverted List Length: 946 Bytes

Term: shock
Df: 239
Tf: 737
Inverted List Length: 487 Bytes

Output Description:

- The program has the following output; it creates two folders in the current directory, **Results** and **FinalIndices**.

- The contents of **Results** are as follows:

CompressedPostings.txt: It contains the postings list that has been compressed using the gamma and delta codes.

Dictionary.txt: It contains the dictionary containing <Term, TotalNoOfDocuments, TotalNoOfTerms> information.

DocumentCount.txt: It contains <Term, NoOfDocs> information.

DocumentList.txt: It contains <Term, <List of DocId's>> information.

MostFrequentStem: It contains <DocID, Term, Frequency> information. It has the most frequent stem for each document.

Postings.txt: It contains the <Term,<List of DocId's, List of Frequencies>> information. It is the postings list.

TermsInEachDocument.txt: It contains <DocID, <ListOfTerms, FreqOfEachTerm>> information. It lists the terms and frequency of the term for each document.

Token-FrequencyInformation.txt: It contains the <Token, Frequency> information.

TotalWordOccurrences.txt: It contains <DocId, TotalWords> information. It has the number of tokens in each document.

WordOccurrencesIncludingStopwords.txt: It contains <DocId, <Term, Frequency>> information. The terms also include all the stopwords.

- The contents of **FinalIndices** are as follows:

CompressedIndex.txt: It is the compressed index with the dictionary and compressed postings.

Index.txt: It is the index with the dictionary and postings.

- There are three arguments for the program. The first is the location of the corpus. The second is the location of the stopwords file, which contains a list of the stopwords which are to be removed while building the index. The third is the location of the Word file, which contains the words for which Df, Tf and inverted list length (in bytes) is to be calculated. Extra terms can be added in subsequent lines in the file.