-------------------------------------------------------------- READ-ME---------------------------------------------------------------

1. Place all the source files are under a single directory
2. Make sure you put the correct path in open commands in the code
3. Please install all the necessary API’s like NLTK, Matplotlib in your Python environment
4. We split the source dataset from SNAP repository (<http://snap.stanford.edu/data/web-Movies.html>) using gSplit software (available for free) into 179 file of ~51 MB each.

Date extraction:-

1. Now we ran Histogram technique on these 179 files generated to get the productID of movies with maximum reviews. These 179 files are present on box in zip file named Movie\_Data.zip (https://iu.box.com/s/9wfo2x7e83mcnu2hwvmvkhl3ttc3hkf1)

**File with code**: test.py

**Prerequisite files**: 179 data files and file\_list.txt (contains a list of 179 names line by line, we created this file manually)

**Output:** gives a dictionary containing list of highest reviewed productID with their frequency or count.

1. We selected the two movies/productID on the basis of output generated from test.py and generated two files named IronMan\_data.txt and Cars\_data.txt with their reviews in them.

**File with code:** generateData.py

**Prerequisite files:** 179 data files and file\_list.txt (contains a list of 179 names line by line, we created this file manually)

**Output:** Generates two Files IronMan\_data.txt and Cars\_data.txt with their reviews in the

Analysis of Data:-

You can directly run the code using data files IronMan\_data.txt and Cars\_data.txt **one by one** for running the analysis. Change the file name in line no 98, You might need to enter complete path of the data file for execution.

**File with code**: final\_analysis.py

**Prerequisite files**: IronMan\_data.txt and Cars\_data.txt

**Output:**

1. Provides output in the Python console for Accuracy and most informative features.
2. Also generates a graph for trend analysis.