**Project Report**

*ETL (Extract, Transform, Load)*

Will Payne & Polina Zasheva

**Extract:**

The data for our project came in the form of CSV files from Kaggle. Specifically, the N**ew Times Comments Dataset** published by Aashita Kesarwani: <https://www.kaggle.com/aashita/nyt-comments>

* The data contains information about the comments made on the articles published in New York Times in Jan-May 2017 and Jan-April 2018. We used datasets for April 2018
* The month-wise data is given in two csv files - one each for the articles on which comments were made and for the comments themselves.
* The csv files for comments contain over *2 million comments* in total with *34 features* and those for articles contain *16 features* about more than *9,000 articles*.
* The data set is rich in information containing comments' texts, that are largely very well written, along with contextual information such as section/topic of the article, as well as features indicating how well the comment was received by the readers such as editors.
* This data can serve the purpose of understanding and analyzing the public mood.
* The exploratory kernel here can be used for a review of the features of the dataset and the NB-Logistic model kernel for predicting NYT's pick can be used as a starter for building models on a range of ideas, some of which are:

**Transform:**

* We extracted multiple files - a set for comments and a set articles, across multiple months. They came in a CSV format
* Stored our files into a ‘Resources Folder’ and created a Jupyter notebook with python script to begin the data clean-up and transformation process
* Read in the CSV files as a dataframe into jupyter notebook
* There were over 50 data fields in each of our files so we selected only what we considered most interesting for potential future analysis, thereby dropping a number of columns
* We renamed the column headers to be more user intuitive and similar across our multiple data sources
* Using python, we merged the data together on Article ID thus creating a large dataset
* Removed rows with N/A values and cleaned the data

**Load**:

* We established a connection to our SQL server. For extra challenge we did this for both MYSQL and SQLite
* We created a new table called ‘article\_comments’ in our SQL server
* Loaded the data
  + In order to load the data successfully into MySQL, in python, we had to set the characters to utf-8 which enabled us to successfully use all valid code points in Unicode using one of four 8-bit bytes.
  + It was important to take that step due to a number unusual characters in our User Name and Comments data points which the default (latin) code setup of SQL wasn’t allowing us to accomplish.
* We chose to utilize a SQL database for our data load because of its robust querying capabilities. Leveraging that, we can answer a number of insightful questions about our dataset such as ‘Who are our most frequent visitors?’, ‘How many comments, on avg., are there by article?’ etc.