**Data Selection Proposal**

**(Deliverable 1)**

260860011 Takuya Ishii

**Project idea:**

In this project, I will attempt to create a text sentiment predictor that takes a phrase as input and returns its sentiment on a scale of 0 to 4 based on its positivity.

1. **Dataset**

Kaggle (<https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews>)

The dataset consists of train and test table. The following 3 columns are present in both tables: PhraseId, SentenceId, and Phrase. Only the train table contains Sentiment label column. For each phrase, sentiment has been measured on a scale of 0 to 4 with 0 being most negative and 4 being most positive.

1. **Methodology**
2. Data Preprocessing

* When training the model, all letters in the dataset should be converted to lower case. The association between a phrase and its sentiment label in the train dataset is case-sensitive. This makes the model’s prediction also case-sensitive, which may reduce the performance of prediction as fewer input phrases would be found in the train table.
* When evaluating user’s input phrase, all letters should be lower-cased.

1. Machine learning model

Sentiment prediction is a classification problem and the dataset allows supervised learning. Considering these, the following algorithms could be applicable:

* + Naive Bayes
  + Logistic Regression  
    - may not be suited for multi-class classification?

1. Evaluation Metric

Possible evaluation methods are:

1. Confusion matrix
2. Misclassification error
3. Final conceptualization

My provisional plan is to showcase this project as a web-app where users can type their phrase and see its sentiment score (using Flask in mind).