# Online News Popularity prediction

## Ehsan RahimiNasab

1. **Introduction:** There is a Dataset provided in UCI Machine Learning Repository which summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict the number of shares of these articles in social networks (popularity) [1].
2. **Summary of Data:** The Dataset consists of 39,644 observations, with 61 variables. Each observation is related to an article published by Mashable in a period of two years. A summary of the description of each variable is provided in *OnlineNewsPopularity.names*.
   Clearly, if there is an article with reference to another article in the set, then the two corresponding observations are not independent, because the one's number of shares will be affected by the other one's popularity. Hence, I have selected a subset of the observations which do not have any self-referenced articles. This new subset has 5350 observations.
   The density histogram, scatter plot, and box plot of the response variable shows that the distribution of number of shares is highly positively skewed and there are heavy outliers in the distribution. The *boxcox* transformation and *qqplot* shows that the response variable can be assumed approximately normal with log transformation.
   The features that I want to consider in my model building and my intuition is that they have the most effect on the response variable are *n_tokens_title* (X1), *n_tokens_content* (X2), *n_unique_tokens* (X3), *num_imgs* (X4), *num_keywords* (X5), *data_channel_is_lifestyle : data_channel_is_world* (X6:X11), *is_weekend* (X12), *global_subjectivity : global_rate_negative_words* (X13:X16), *abs_title_subjectivity* (X17), and *abs_title_sentiment_polarity* (X18).
   The most correlated predictors among the above features are *global_subjectivity*(X13) and *n_unique_tokens* (X3) with 0.9 Pearson coefficient. Then *global_rate_positive_words* (X15) and *global_rate_negative_words* (X16) with *global_subjectivity* (X13) with coefficient 0.8 and 0.7 respectively. Also *global_sentiment_polarity* (X14) and *global_rate_positive_words* (X15) have 0.7 Pearson coefficient.
   There is no missing values in the dataset.

3. **Methods:** I want to use multiple linear regression method in order to build a model for this dataset which enables me to predict the number of shares of each observation according to their features values. I will divide the dataset to three parts of training, development, and test set in order to choose the best linear model and the best predictive subset of the features which have the best result on the development set. Then I will report the prediction accuracy on the test set. I will try to see what happens if I add other features to the model which I neglected based on my intuition at the beginning.

4. **References:**

[1] K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.