Abstract

# Ci siamo concentrati sulla sfida di Kaggel denominata “[Bag of Words Meets Bags of Popcorn](https://www.kaggle.com/c/word2vec-nlp-tutorial)”, la quale consiste in una classificazione supervised di un insieme di recensioni.

# Il programma è stato scritto in Python utilizzando tre librerie fondamentali:

# [Sklearn](http://scikit-learn.org/stable/index.html): per i classificatori, la vettorializzazione delle stringhe e la dimensionality reduction.

# [Pandas](http://pandas.pydata.org/pandas-docs/stable/index.html): per la creazione e manipolazione dei dataset.

# [MatplotLib](http://matplotlib.org/): per la creazione dei grafici.

# La classificazione è binaria, recensione positiva o negativa, e il nostro obbiettivo è stato usare molteplici classificatori, quali: [Random Forest](http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html), [SGD](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier), [Naive Bayes Bernoulli](http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html) e la [Naive Bayes Multinomial](http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html) per constatare la differenza nelle predizioni.

# Per prima cosa abbiamo sottoposto i dati ad un pre-processamento, essendo questi stringhe prima abbiamo rimosso le così dette “stop words”, ovvero le parole che appaiono spesso nella lingua inglese ma che non sono utili al fine della classificazione, per esempio “the, it, and…”. In seguito abbiamo rimosso la punteggiatura e i caratteri non alfa-numerici, poiché le recensione portavano con se alcuni tag provenienti dal linguaggio HTML.

# Fatto questo le stringhe sono state convertite in vettori utilizzando l’algoritmo [TFIDF](http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfVectorizer) (term-frequency-inverse-document-frequency), che ci ha fornito un dataset con oltre 70000 features. Lo step successivo è stato la dimensionality reduction tramite l’utilizzo della funzione [Select-K-Best](http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html) utilizzando come stima il [Chi2](http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html#sklearn.feature_selection.chi2).

# Una volta ottenuti dei dati opportunamente trattati siamo passati alla classificazione con i precedenti citati algoritmi, ottenendo i seguenti risultati facendo una media dell’errore nelle predizioni:

# Random Forest: 0.85596

# Bernoulli: 0.82704

# Stocastic Gradient Descent: 0.88324

# Multinomial: 0.84184