IMDB Reviews

Text Mining and Search

Agazzi Ruben 844736, Cominetti Fabrizio 882737

Abstract
In this project, user reviews from the IMDB platform were analyzed through the use of text mining techniques. After carrying out an initial phase of text processing and text representation, the project continued with the classification of the reviews, through some text classification techniques - such as Support Vector Machines (SVM), Multilayer Perceptron (MLP), and Logistic Regression. Next, a text clustering phase was carried out through the use of two algorithms: DBSCAN and k-means.

Keywords
Text Mining — Text Classification — Text Clustering

University: University of Milan-Bicocca

Contents

1. Introduction

The project aims to analyze the dataset "IMDB reviews" [**empty citation**] through text mining techniques, specifically through *Text Classification* and *Text Clustering*. The dataset contains a total amount of 50000 user-released reviews on the IMDB platform, divided in half between training and testing. The dataset is ideal for performing a binary sentiment classification task, the first objective of our analysis. Next, we decided to exploit Text Clustering techniques with the goal of identifying different clusters within the text.
Text Classification is the activity of predicting which data items belongs to a predefined finite set of classes. There are many types of classification, in our case it is *Binary Classification*, where each item belongs to exactly one class in a set of two (positive or negative).
In addition, text classification may be performed according to several dimensions ('axes') orthogonal to each other. For example, by topic (the most frequent case), by sentiment - our case -, by language, by type, by author, by native language, by gender, and more.

2. Data

As stated before, the dataset used contains a total of 50000 user reviews on the IMDB platform, a platform that describes itself in the following manner :"IMDb is the world's most popular and authoritative source for movie, TV and celebrity content. Find ratings and reviews for the newest movie and TV shows."

Large Movie Review Dataset.

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided.
From an initial exploration of the data, we can observe that the dataset does not provide information about the date and reference film of the review, or any other indication, but contains only the text of the review and the extracted sentiment - positive or negative.
The data, initially divided into training and testing, but also between positive and negative sentiment, were merged, so that there would be a single dataset for the 25000 reviews to be used in the training phase and the 25000 reviews to be used in the testing phase.
Finally, the dataset contains precisely 12500 reviews labeled as positive and as many labeled as negative, both training and testing.

3. Text Processing

Having obtained the starting dataset, a series of *Text Processing* operations were performed:

- *Remove Numbers*, x;
- *Remove StopWords*, x;
- *Remove Punctuation*, x;
- *Remove Extra Space*, x;
- *Tokenization*, x;
- *Lower Case*, x;
- *Lemmatization*, x.

4. Text Representation

...

5. Text Classification

...

5.1 SVM

...

5.2 Multilayer Perceptron

...

5.3 Logistic Regression

...

5.4 Evaluation

...

6. Text Clustering

...

6.1 DBSCAN

...

6.2 K-means

...

6.3 Evaluation

...

7. Summary

...