Preprocessing

Mandatory assignment 04 - report by Frédéric Charon

This assignment is about preprocessing data for a model. The data consists of review texts for movies, labelled as positive and negative and we have to decide which preprocessing methods to use before storing the changed data into a dataframe.

I decided to use five functions: lowercasing and removing punctuation, stopwords, urls and emoticons.

Frequent and rare words aren't removed since they may contain valuable information about the sentiment, for example the word "like" is used a lot – removing it would result in a different outcome for the model since it is a keyword to get the authors opinion about the movie.

Lowercasing could also be removed since it may contain sentiments, for example if a word or a sentence is written in all caps. However, I've decided to implement it because I think the profit of lowercasing seems higher than the amount of sentiments ignored this way.

When removing the punctuation I decided to leave the symbols ! and ?, each use of them might express a sentiment for example a lot of exclamation marks can mean that the author was very excited or frustrated about the film.

In my code emoticons are removed but not emojis, this is because the files are in .txt format using UTF-8 so they may contain emoticons, but they can't contain emojis.