Artificial Intelligence 2020/2021

Exercise Sheet 7: Natural Language Processing

7.1 Software/Library Installation

Besides the usual python libraries for machine learning, which include *pandas* and *sklearn*, for this exercise sheet you will need some libraries to work with text, including regular expression matching operations (*re*), and the natural language toolkit (*nltk*).

After installing the needed libraries, please open the example Notebook available at Moodle.

7.2 Restaurant Reviews Dataset

The restaurant reviews dataset includes 1000 single-line English reviews on restaurants, associated with a polarity (1 = positive review; 0 = negative review). The text included in reviews can be seen as noisy, in the sense that not every token corresponds to a word in English or a punctuation mark.

In this Notebook, we will develop sentiment analysis models for predicting the polarity of restaurant reviews.

a) Unzip the file with the dataset and the example notebook available at Moodle, and open the notebook.

b) Use the pandas library to Import the restaurant reviews dataset.

c) Using the regular expressions library (re), perform some simple cleanup on the text. For example, you may consider only alphabetic character sequences, and convert the whole text to lowercase.

d) Tokenize the text, use NLTK's Porter stemmer to stem the obtained tokens, and remove stop words using NLTK's stop word list for English.

e) From the cleaned up and tokenized corpus, create bag-of-words features, using sklearn's CountVectorizer. Now you should have obtained a structured dataset, where each restaurant review is represented by a list of 0's and 1's with the size of the vocabulary.

f) Split the data into train and test sets.

g) Try to *fit* a Naïve Bayes classifier to the training set, and *predict* its test set results. Analyse the confusion matrix and the classification scores (accuracy, precision, recall, F1).

h) Try out the model by prompting the user to input a restaurant review and predicting its class.

i) Experiment with other classifiers and see if you can improve on the performance of the classification model.