

NLTK classification task 1

Joanna

10 stycznia 2023

1 Exercise 1

Work with the movie review corpus which is appropriate for sentiment analysis.

1. Check the corpus, size, how many reviews there are, and how many of them are positive/negative. (If you think of any other information you can obtain, please do, use examples from previous classes.)
2. Use a short review "Mr. Matt Damon was outstanding, fantastic, excellent, wonderfully subtle, superb, terrific, and memorable in his portrayal of Mulan." See how the classifier classifies this short and fake movie review.
 - (a) Define a variable storing the review.
 - (b) lowercase the text
 - (c) tokenize the text
 - (d) generate word feature dictionary using `document_features` function
 - (e) Classify (`classifier.classify`) prepared features - save results
 - (f) Calculate and save the classification probability to class 'pos' (`classifier.prob_classify`).
 - (g) Calculate and save the classification probability to class 'neg' (`classifier.prob_classify`).
3. Change "Matt Damon" to "Steven Seagal". See how the classifier classifies this short and fake movie review. Compare with previous results. Save and comment on them.
4. Try short review "Mr. Matt Damon was outstanding, fantastic.". Save and compare results - comment on them. Hint: the classifier learns from 2,000 presence/absence word features.
5. Using the movie review document classifier generates a list of the 30 features that the classifier finds to be most informative. Please explain why these particular features are informative? Are they surprising?

Submit the code file including also your comments.