Introduction to NLP (Course 2017-2018)

Final Project

1. Main task: build a system for (binary) sentiment analysis classification

2. Task specifications:
    a. **Dataset:** the NLTK movie_reviews corpus (1000 positive, 1000 negative)
        i. The corpus is annotated (positive negative)
        ii. The corpus is tokenized and sentence segmented
        iii. Each review contains multiple sentences
        iv. The corpus is NOT split into test/train
    b. **Classifier/Architecture:** Any supervised algorithm available in nltk or scikit-learn
    c. **Mandatory pre-processing and/or features:** None. You can use any pre-processing step shown in the class. You can use any features that you see fit
    d. **Submission deadline:** 15.06

3. Objectives:
    a. Build a system for a "real" NLP task
    b. Use (some) of the tools and concepts learned in class
    c. Write a complete program from scratch
    d. Look for information in out-of-class resources (tutorials, articles)

4. Suggested Pipeline:
    a. Load the corpus from nltk
    b. Split train/test **(75% for training, 25% for testing)**
    c. (Pre)-process the corpus
    d. Define and extract features
    e. Plug the features in a classifier and evaluate
    (optional)
    f. Use different test/train to cross-validate
    g. Use different sets of features to determine the importance of the different (sets of) features

5. Evaluation:
    a. Understanding the task and obtaining positive results
    b. Using (some of) the tools and skills shown in class (frequencies, statistical association, n-grams, co-occurrence, POS tagging, word embeddings)
    c. Code quality and readability
    (bonus)
    d. Using external resources
    e. Feature analysis and comparison of different features and/or architectures