COMP34412
Natural Language Systems

Coursework 2

In this coursework, you will explore corpus-based approaches to named-entity tagging and sentiment analysis.

The coursework is worth a total of 10 marks of the final COMP34412 final mark.

Tasks

1) Named-entity recognition [3 marks]

a) NLTK provides a classifier that has been trained to recognise several types of named entities (see Section 5 at http://www.nltk.org/book/ch07.html). Use the function nltk.ne_chunk( ) to process corpus 1 (see Data below).

b) Use the Stanford named-entity recogniser (https://nlp.stanford.edu/software/CRF-NER.shtml) to process the same corpus.

c) Compare the outputs of the two NER methods for the ORGANIZATION class. Explore the differences between the two approaches – which of the tools seems better in getting the bounders of named entities right? Provide some examples. In how many cases the tools fully agree between themselves on the mentions of named entities (*exact* match), and in how many cases they have a *partial* overlap. Your report should be around 1 page long.

2) Sentiment analysis of movie reviews [7 marks]

a) With the popularity of social media, building and maintaining a sentiment/polarity lexicon is a huge challenge. In the first part of this task, you will build a sentiment lexicon using a semi-supervised approach by bootstrapping the process, starting from a small lexicon of adjectives (see at the end of the document) and corpus 2 (see Data below). Write a program that will collect more adjectives to populate the lexicon and assign them with the likely polarity. For example, adjectives conjoined by "and" are likely to be of the same polarity (e.g. *corrupt and brutal*), while adjectives conjoined by "but" are not (*fair but brutal*). Consider (and implement) other possible patterns; consider how you could assign a polarity if an adjective appears several times in the corpus (and, for example, you have conflicting polarity signals). Evaluate the outcome – how many of the proposed adjectives have been properly classified (according to your judgement of their *typical* polarity)? Provide explanations for any errors in the report (~1 page).

b) The two files in corpus 2 represent positive and negative examples of movie reviews. As a baseline, implement a classifier that simply counts whether there are more positive or negative words in a review (use the MPQA lexicon Subjectivity Cues Lexicon, see below). Then build a machine-learning sentiment classifier using a simple bag-of-words approach. Expand the feature set to include other possible features: e.g. whether any of the words in the review come from a polarity lexicon (e.g. MPQA), whether they are negated, etc. For training your classifier, use any available machine-learning framework (e.g. scikit-learn (http://scikit-learn.org/) or Weka (http://www.cs.waikato.ac.nz/ml/weka/). Evaluate the method using k-fold cross validation and compare to the baseline. Explain briefly what you have done and discuss the results in the report (~1 page).

Submissions

The deadline for submissions is **6pm on Friday March 27th 2020**. Your submission should be a zip file uploaded via Blackboard. For each task you should submit a write-up that clearly explains what you have done and presents the results. Please submit both write-up parts in a single pdf. You should also submit your source code, and the output of either your code or of a third party tool you have used (where applicable). The README file should clearly specify how to run your programs. You can also submit your coursework as a Jupyter notebook.

Data

Task 1:

Corpus 1: the inaugural address corpus is available in the Blackboard.

Task 2:

Corpus 2 (see Blackboard) contains two files, one with positive and one with negative examples of movie reviews that have been collected for sentiment-analysis experiments (see http://www.cs.cornell.edu/people/pabo/movie-review-data/). For the first part (a) of Task 2, merge the two files and use it as a single corpus. For the second part (b), you will use them as separate files with positive/negative examples.

The MPQA sentiment lexicon is available at:
http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

Initial lexicon of adjectives (also available in the Blackboard):

Negative	***Positive***
silly	inspiring
sour	exotic
cynical	good-looking
amateurish	effective
offensive	gripping
stupid	thrilling
dishonest	intriguing
r-rated	satisfying
rough	entertaining
unsuccessful	stylish
unfunny	funny
repetitive	emotional
sappy	naturalistic
dull	romantic
dry	resonant
mush-hearted	brilliant
predictable	absorbing
creepy	fresh
neurotic	lyrical
disturbing	honest
	clever