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 40 marks** of the final COMP34412 final mark.

Tasks

1) Named-entity recognition [10 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 [30 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 below) 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 (e.g. *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 outcomes: how many of the proposed adjectives have been properly classified (according to your judgement of their *typical* polarity *or* using the MPQA lexicon Subjectivity Cues Lexicon, see below). Provide explanations for any errors. Your report should be around 1 page long.

b) The two files in corpus 2 represent positive and negative examples of movie reviews. As a baseline, implement a classifier that classifies movie reviews based on the numbers of positive or negative words appearing in a review (use the MPQA lexicon Subjectivity Cues Lexicon). Then, build two machine-learning sentiment classifiers: one based on a bag-of-words approach, and one using word2vec embeddings. Consider expanding the feature set to include other possible features: e.g. try incorporating negation. Evaluate the methods using accuracy and k-fold cross validation, and compare to the baseline. Discuss the pros and cons of your approaches. What other approaches you may consider using in your classifiers (you don't need to implement these). For training your classifiers, use any available machine-learning framework (e.g. scikit-learn (http://scikit-learn.org/) or Weka (http://www.cs.waikato.ac.nz/ml/weka/). Explain briefly what you have done and discuss the results in the report (~2 pages).

Submissions

The deadline for submissions is **6pm on Tuesday May 4ᵗʰ 2021**. Your submission should be a single 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 and discussions. Please submit all write-up parts in a single pdf (up to 4 pages). You should also submit your source code (e.g. as a Jupyter notebook). Any useful output of either your code or of a third-party tool you have used (where applicable) can be included. The README file should clearly specify how to run your programs.

Please add your name to the report and make it look professional. Please provide useful comments in the code. All code and reports will be checked for academic malpractice.

Data

Task 1:

Corpus 1 – this is the inaugural address corpus, available in the Blackboard (and in nltk).

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 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 for training and evaluation.

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 gripping
offensive	thrilling
stupid	intriguing
dishonest	satisfying
r-rated	entertaining
rough	stylish
unsuccessful	funny
unfunny	emotional
repetitive	naturalistic
sappy	romantic
dull	resonant
dry	brilliant
mush-hearted	absorbing
predictable	fresh
creepy	lyrical
neurotic	honest
disturbing	clever

Marking schema and indicative rubrics:

Task 1 [total of 10 marks]
- NER processing [total of 4 marks]
    - o Corpus processed with NLTK (2 marks)
    - o Corpus processed with Stanford NER (2 marks)
- Comparison and discussions [total of 6 marks]
    - o Boundary detection discussion (3 marks)
    - o Agreement between tools (3 marks)

Task 2 [total of 30 marks]
- Bootstrapping sentiment lexicon [total of 10 marks]
    - o Basic patterns for finding new adjectives (3 marks)
    - o Additional patterns implemented (2 marks)
    - o Reconciliation (2 marks)
    - o Discussion and evaluation of the extended dictionary (3 marks)
- Classifier implementations [total of 11 marks]
    - o Baseline classifier (2 marks)
    - o BoW classifier (3 marks)
    - o word2vec classifier (3 marks)
    - o Additional features implemented (3 marks)
- Evaluation and analysis of classification results [total of 9 marks]
    - o Accuracy and cross-validation (2 marks)
    - o Comparison to baseline (2 marks)
    - o Overall discussions of results, including pros/cons (3 marks)
    - o Other ideas (2 marks)