Text Mining

Course project description

Sentiment classification of movie reviews

Context Sentiment analysis is one of the many applications of text mining (Cambria *et al.*, 2013). It consists in automatically determining the polarity of a textual document. With the abundance of discussions, reviews, *etc.*, that are available online, sentiment classification provides an interesting way of summarizing information for readers, in addition to topical classification.

Data In 2004, Pang and Lee introduced the "polarity dataset version 2.0", that consists of 2,000 movie reviews, where each is associated with a binary sentiment polarity label (either "neg" or "pos").

Task Predict the polarity of a new, unseen movie review.

1 - Detailed project description

The goal is to conduct an end-to-end study about the prediction of the polarity of movie reviews. You should go through these stages, each of them being a section in a detailed R markdown notebook:

(i) Experimental design Present the problem, the data, and your evaluation protocol.

(ii) Data description and exploratory data analysis Present the data, provide some descriptive statistics (*e.g.* distribution and conditional distributions of the number of tokens in the reviews) and provide some insights about the content of the reviews (*e.g.* perform topic modeling with NMF and/or LDA and report the topics proportions, topic summaries, correlation between topics and polarity).

(iii) Experimental results Experiment and report results with various representations of the reviews (*e.g.* bag-of-words, weighted bag-of-words, bag-of-ngrams, topic mixture, average of word embeddings, combinations of differents representations), additional variables extracted from the text (*e.g.* number of tokens, frequency of punctuation signs) and various classifiers (*e.g.* multinomial naïve Bayse classifier, logistic regression, regularized logistic regression, LSTM).

(iv) Conclusion Sum up your experiments and recall what the best representation & classifier combo.

The notebook should be self-contained, meaning one should be able to get the basics and the intuition behind every technique you use. In other words, before applying a technique and commenting the results, you should briefly introduce this technique (*e.g.* state the related assumption(s), present the formalism). Also, it shouldn't be a collection of experiments, but, rather, a report with a logical organisation.

2 - Submission details

You work should be uploaded on EVA: https://eva.fing.edu.uy/course/view.php?id=1223#section-1

Format You have to upload a ZIP file containing (i) an R markdown file, (ii) the input data, and (iii) the HTML rendering of the notebook. The R markdown file should compile using the latest version of R.

Deadline You have to submit your work by Sunday, November 18 at 11pm, local time.

Oct. 2018 – Maestría en Ingeniería Matemática, Universidad de la República, Uruguay
Instructor: Adrien Guille

References

- E. Cambria, B. Schuller, Y. Xia, C. Havasi (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems 28 (2), pages 15–21.
- B. Pang and L. Lee (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278.