UNIVERSITEIT TWENTE.

Data Science [201400174]

Course year 2019/2020, Quarter 2A

DATE
January 31, 2020

TEACHERS
Maurice van Keulen
Christin Seifert
Mannes Poel
Karin Groothuis-Oudshoorn
Elena Mocanu
Faiza Bukhsh
Nicola Strisciuglio

COURSE COORDINATOR
Christin Seifert
Maurice van Keulen

PROJECT OWNERS
Faiza Bukhsh
Karin Groothuis-Oudshoorn
Maurice van Keulen
Elena Mocanu
Mannes Poel
Michel van Putten
Mohsen Jafari Songhori
Luc Wismans

Project 5

Project 5: Text classification or Named Entity Recognition [TCNER]

5.1 Introduction

Project owner: Maurice van Keulen
Primary topic: IENLP
Well combinable with DM or SEMI.

In the realm of information extraction and natural language processing, there are two suggested projects to select from

- **Text Classification**
  Recommend a conference to a researcher given the title of his new article
- **Named Entity Recognition**
  Extract and classify named entities from tweets

In case that a group wants to suggest a different project, the group should submit an initial project proposal to be reviewed. In this case the group is encouraged to meet with the project owner / topic teacher to discuss project ideas.

The suggested projects represent two challenges. For each challenge, you will be given a training set to train and tune your system on. The models should only be optmized using the training data. The test set then gives you a good estimate how your method behaves on unseen data.

Note that the project grading is not related to the achieved results, but rather on the methodology and the soundness of the evaluation.

5.2 Description of data set

The data for *Text Classification* is Conference Proceedings training data from the paper below.

Reference: Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1 (March 2002), 1-47.http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf.

The data for *Named Entity Recognition* is NER Twitter training data available on Canvas.

References: (1) Nadeau, David & Sekine, Satoshi. (2007). A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes. 30. 10.1075/li.30.1.03nad. (2) Sharnagat, R. (2014). Named entity recognition: A literature survey. Center For Indian Language Technology. http://www.cfilt.iitb.ac.in/resources/surveys/rahul-ner-survey.pdf.

5.3 Description of challenge

5.3.1 Text Classification

- Design and implement a system to recommend a conference to a researcher given the title of his new article.
- The system should use the provided Conference Proceedings training data. You should implement the sub tasks (feature extraction, dimensionality reduction, and classifier) by yourself.
- You are free to select the algorithms you prefer for each sub task. However it is recommended that you test and compare multiple methods.
- Evaluate you system on the training set by using the cross-validation approach. Provide the confusion matrix of your system output.
- Evaluation should be done in terms of Micro-average precision, recall and F1 measures.
- Once you fount the best model on the training set, evaluate your model on the test set and report the results.

5.3.2 Named Entity Recognition

- Design and implement a system to extract and classify named entities in tweets.
- The system should use the provided NER Twitter training data. You should implement the sub tasks (feature extraction, and classifier) by yourself.
- You are free to select the algorithms you prefer for each sub task. However it is recommended that you test and compare multiple methods.
- Evaluate you system on the training set by using the cross-validation approach.
- Evaluation should be done in terms of micro-average of precision, recall and F1 measures.
- Evaluate your best model you found on the training set on the test set. Report your results.

5.4 Tips and suggestions