Text Mining and Sentiment Analysis

Data Science and Economics (DSE), 2019/2020, The University of Milan

A. Ferrara [1], *E. Guidotti* [2]

Course Overview

The course provides a complete overview of the state of the art and research perspective in the field of text mining and sentiment analysis, with an introduction to some relevant and correlated problems such as emotion detection and opinion mining.

The course program is articulated in two main modules of 20 hours each. The first introduces the main notions needed to understand text processing, foundations of natural language processing, text classification, and topic modeling. The second module addresses sentiment analysis in the context of opinion mining and introduces rule-based models and machine learning models, including statistical language models and neural networks. Emotion detection is also discussed. Finally, the lectures will be used also to drive students towards the choice of a topic for their final project and short paper by means of several case studies.

Course Content

1. **Introduction** (0:30h)

   Course introduction, logistic issues, course requirements and Python installation.

2. **Natural language processing** (3:30h)

   Basic techniques in natural language processing: tokenization (bag-of-words and n-gram models), stopwords and punctuation, stemming and lemmatization, part-of-speech tagging, chunking, regular expressions and named entity recognition. Public NLP toolkits such as NLTK and SpaCy will be introduced to gain hand-on experience in Python.

3. **Document representation** (2h)

   The Vector Space Model and tf-idf weighting: representing unstructured text documents with appropriate format and structure to support later automated text mining algorithms. PCA as dimensionality reduction technique.

[1] alfio.ferrara@unimi.it (Sentiment Analysis)
[2] emanuele.guidotti@unine.ch (Text Mining)

4. **Text classification** (5h)

   Feature selection and text categorization algorithms: Naive Bayes, k Nearest Neighbor (kNN), Logistic Regression, Support Vector Machines and Decision Trees. Evaluation of text classification: precision and recall, confusion matrix, F-score.

5. **Text clustering** (3h)

   Clustering algorithms, i.e., connectivity-based clustering (a.k.a., hierarchical clustering) and centroid-based clustering (e.g., k-means clustering). Evaluation of text clustering: purity and Rand index.

6. **Topic modeling** (4h)

   Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. Two basic topic models will be covered: Probabilistic Latent Semantic Indexing (pLSI) and Latent Dirichlet Allocation (LDA).

7. **Document summarization** (2h)

   It refers to the process of reducing a text document to a summary that retains the most important points of the original document. Extraction-based summarization methods will be covered.

8. **Introduction to sentiment analysis and emotion detection** (1h)

   Definition of the sentiment analysis problem. Differences between sentiment analysis and emotion detection.

9. **Lexicon-based approaches to sentiment analysis** (4h)

   Survey of the main approaches that exploit dictionaries, ontologies, and specialized corpora for detecting the sentiment polarity in texts.

10. **Machine learning approaches to sentiment analysis** (4h)

    Sentiment and polarity detection as a classification problem. Overview and comparison of the main unsupervised and supervised models on a case study.

11. **Overview of neural network architectures for sentiment analysis** (2h)

    Design and implementation of a case study based on a neural network for sentiment detection and polarity evaluation.

12. **Affect and emotion detection** (1h)

    Survey and definition of affect and emotion detection in texts. Discussion about the differences between the tasks of detection of sentiment, feelings, emotions, and opinions.

13. **The language of emotions** (4h)

    Methods and techniques for modeling the language of emotions using neural networks and statistical language models. Application to a case study.

14. **Multimodal approaches to emotion detection** (1h)

    Survey on the exploitation of multimodal data (e.g., face and body language in video and audio recordings) in combination with text to detect the language of emotions.

15. **Hands on a real case study for design to implementation** (2h)

    Students will be provided with a real case study on sentiment analysis and emotion detection. During the lesson, the case study will be studied to the end of design and implement a solution.

16. **Recap and conclusion** (1h)

    Recap on the main course topics. Open discussion of the project work chosen by the students as their exam assignment.

Tools

Reference programming language: **Python**. Main modules:

1. NLTK
2. scikit-learn
3. spaCy
4. Gensim
5. TensorFlow and Keras

References

1. NLTK Book: https://www.nltk.org/book/
2. Aggarwal, C. C., & Zhai, C. (Eds.). (2012). Mining text data. Springer Science & Business Media.
3. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.
4. Munezero, M. D., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. IEEE transactions on affective computing, 5(2), 101-111.

5. Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on affective computing, 1(1), 18-37.