Text Mining Practical: Project Ideas

Prof. Dr. Siegfried Handschuh, Dr. Jelena Mitrović
Chair of Digital Libraries and Web Information Systems
Natural Language and Semantic Computing Group
Summer Semester, 2018

1 Introduction

This document suggests a number of project titles for the text mining project course. Individual students or small groups must read the following document and prepare a brief proposal describing what they want to do. We highly recommend the use of LaTeX for documentations.

Your proposal contains the following sections: introduction, targets/plan and references. In the introduction section, you must define and outline your proposed project goals briefly. Introduction can be used to explain why you are interested in this topic and what you would like to learn. For instance, this can be achieved by extending the given project ideas. You must cite at least two references related to the project you are going to do. Introduction can be about 300 to 600 words long. You are encouraged to use figures to explain your idea.

In the target section, you will list your targeted topics/skills to study. For instance, in your project, you may want to emphasise a particular machine learning technique. Consequently, you must list a number of characteristics of the learning method that you are going to study. Alternatively, the emphasis can be on a specific application. In this case, you must list a number of important characteristics of this application and explain how you are going to study these aspects. In any case, your project works must be accompanied by quantitative measures.

The documents or external resources for your project work must be listed in the bibliography section. Please use the name and year citation style. For your final project report, you are expected to extend your project proposal document with a method and discussion section followed by a conclusion.

In the remaining parts of this document, project ideas for this semester are listed.

2 Topics

2.1 Aspect Based Sentiment Analysis

- `http://www2.aueb.gr/users/ion/docs/SemEval2016_ABSA_overview.pdf`

- http://alt.qcri.org/semeval2016/task5/
- http://m-mitchell.com/NAACL-2016/SemEval/pdf/SemEval62.pdf
- http://m-mitchell.com/NAACL-2016/SemEval/SemEval-2016.pdf

2.2 Affect in Tweets

- http://alt.qcri.org/semeval2017/task1
- https://aclweb.org/anthology/S/S16/S16-1081.pdf
- https://competitions.codalab.org/competitions/17751
  http://www.aclweb.org/anthology/S18-1000

2.3 Topic and Trend Detection in Scientific Papers

http://dblp.uni-trier.de/pers/hd/o/Osborne:Francesco
http://2017.eswc-conferences.org/program/accepted-papers

2.4 Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification

http://search.proquest.com/openview/5f2a5f4401187b61868b9d537109ec97/1?pq-origsite=gscholar&cbl=18750&diss=y

2.5 Towards Automatic Extraction of Epistemic Items from Scientific Publications

http://dl.acm.org/citation.cfm?id=1774377&CFID=945064051&CFTOKEN=80622791

2.6 Extraction of Epistemic Items from Legal text Corpora

https://dl.acm.org/citation.cfm?id=1774377

2.7 Automating Legal Research through Data Mining

https://pdfs.semanticscholar.org/4d49/2d103672723d5683e4fc5b468e49ffaece3b.pdf
http://ceur-ws.org/Vol-710/paper23.pdf

2.8 Classification of Proposition Types in Legal Texts

https://cgi.csc.liv.ac.uk/~katie/jurix16a.pdf
http://ceur-ws.org/Vol-1341/paper3.pdf
http://dl.acm.org/citation.cfm?doid=1568234.1568246
References from workshops at legal informatics conferences:
ICAIL https://icail2017evidencedecision.wordpress.com/
JURIX http://jurix2016.unice.fr/?page_id=130
IRIS http://www.obs.coe.int/en/legal

2.9 Argument mining in Legal Texts

Corpus available, natively in German, or a translation to English
https://www.ukp.tu-darmstadt.de/data/argumentation-mining/argument-annotated-essay

2.10 Detection and Analysis of Legal Reasoning

http://www.uni-weimar.de/medien/webis/publications/papers/stein_2016q.pdf

2.11 Detection and analysis of political argumentation in the Spiegel corpus

http://ceur-ws.org/Vol-1341/paper6.pdf
https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/12164/12050

2.12 Detection and Classification of Argumentative Discourse Units in the Spiegel Corpus

http://www.uni-weimar.de/medien/webis/publications/papers/stein_2016q.pdf (distinction of 6 types of argumentative discourse units: common ground, assumption, testimony, statistics, anecdote, other)
https://pdfs.semanticscholar.org/ee55/b972fbf8398ae9dc32c64b367551c7277682.pdf?_ga=2.257218763.1134470293.1496402075-678269664.1496402075
http://www.aclweb.org/anthology/W14-21#page=76
https://www.research.ibm.com/haifa/dept/vst/papers/Evidence2015.pdf

2.13 Ontology Learning from the Spiegel Corpus

https://protege.stanford.edu/publications/ontology_development/ontology101.pdf
http://jens-lehmann.org/files/2014/pol_introduction.pdf
http://www.jlcl.org/2005_Heft2/Chris_Biemann.pdf

2.14 Comparison of ML Approaches on a Multilayered Scientific Corpus

http://sempub.taln.upf.edu/dricorpus

This topic includes building a similar corpus for NLP – a group of 4 can participate

http://sempub.taln.upf.edu/dricorpus

2.15 Mining Arguments from Legal Cases

http://wyner.info/research/Papers/WynerMochalesPalauMoensMilward2009.pdf

2.16 Machine Learning for Rhetorical Figure Detection

http://www.aclweb.org/anthology/W13-1605

https://www.lexalytics.com/lexablog/2013/sentiment-and-litotes-how-salience-deals-

https://competitions.codalab.org/competitions/17468

http://www.sciencedirect.com/science/article/pii/S0169023X12000237

Irony detection on microposts with limited set of features

http://bit.ly/2rvWPVG

http://www.ep.liu.se/ecp/131/005/ecp17131005.pdf

2.17 Rhetorical Figure Detection in Political Texts

http://computationalrhetoricworkshop.uwaterloo.ca/wp-content/uploads/2016/06/An-Annotation-Tool-for-Automatically-Detecting-Rhetorical-Figures.pdf

http://bit.ly/2raxVHW

http://computationalrhetoricworkshop.uwaterloo.ca/wp-content/uploads/2016/06/Rhetorical-Figure-Annotation-with-XML.pdf

2.18 Extra info files for Twitter related topics

- http://alt.qcri.org/semeval2016/task4/data/uploads/semeval2016_task4_report.pdf
- http://m-mitchell.com/NAACL-2016/SemEval/SemEval-2016.pdf
- http://alt.qcri.org/semeval2016/task4/

2.19 Rhetorical Relation Identification for English

Any sound and coherent text is not simply a loose arrangement of self-contained units, but rather a logical structure of utterances that are semantically connected so that the reader

is able to interpret the meaning of the message that was presented by the author. This challenge of uncovering coherence structures in texts is pursued in the area of Discourse Parsing, which aims to identify discourse relations that hold between textual units in the text. Using a parallel corpus of complex and syntactically simplified sentences (generated by Graphene), identify the rhetorical relation that holds between a core sentence and each of its associated context sentences. Example:

- complex input: "In February 2013, Obama said that the U.S. military would reduce the troop level in Afghanistan by February 2014."

- simplified core: "The U.S. military would reduce the troop level."

- Context: "This was what Obama said in February 2013." (relation: attribution)

- Context: "This was by February 2014." (relation: temporal)

- Context: "This was in Afghanistan." (relation: local)

References

- Mann and Thompson (1988): "Rhetorical structure theory: Toward a functional theory of text organization"

- Marcu (1997): "The rhetorical parsing of natural language texts"

- Feng and Hirst (2014): "A linear-time bottom-up discourse parser with constraints and post-editing."

2.20 Semantic Linking for Event-Based Classification of Tweets

`https://hal.inria.fr/hal-01529729/document`

Event extraction from Twitter

`http://bit.ly/2rKxnKL`

2.21 Political argumentation and prevalence of right-winged vs. left-winged speeches

A corpus of German political speeches `http://www.lrec-conf.org/proceedings/lrec2018/pdf/324.pdf`

2.22 Distributed word representations – Distributional semantics project

`http://www.lrec-conf.org/proceedings/lrec2018/pdf/721.pdf`

2.23 Distributional term expansion

`http://www.lrec-conf.org/proceedings/lrec2018/pdf/303.pdf`

2.24 Distributional semantics for discourse relations

https://www.cc.gatech.edu/~jeisenst/papers/ji-tacl-2015.pdf

2.25 A distributional-semantics approach to detection of semantic change in the Google Books Ngram corpus

http://www.aclweb.org/anthology/W11-2508

2.26 A distributional semantics model for idiom detection

https://msuweb.montclair.edu/~feldmana/publications/NLPinAI_2018_8_CR.pdf

2.27 Own Topics

You can propose own topics. But you need to be able to convince us that your topic is valid, interesting and challenging enough for the course, as well as equal to the given ones. We reserve the right to reject your topic based on our subjective assessment.

3 Project Assessment

The outcome of the project works are project codes, a report and a presentation. The assessment is based on

- Project Codes (50% of final mark)
- Project Report (30% of final mark)
- Presentation (20% of final mark).

  TOPIC ASSIGNMENT LINK

  https://docs.google.com/spreadsheets/d/1GBtRA7E5Y3h6lzyEPsULXqh9MsESHZyMSfcy0LA3
  edit#gid=0

The above mentioned output of your project work are assessed by the following criteria:

- Topic knowledge
- Technical soundness
- Originality and Creativity
- Organisation
- Communication performance (use of visual aid, stage performance, etc.)

- Presentation

The deadline for the completion of the project work is July 2018 or October 2018. The first date is for those who want to finish early; the second date is for those who want to go the extra mile and work during the semester break. Nevertheless, both will be marked and assessed with the same standard. Please note, that there will be only one registration in the examination system for both dates!

You can work in a team of 2, 3 or 4. Of course, we expect an accordingly better performance from a team of 4 than from the team of 3.

Copying code without mentioning the original source is forbidden and it counts as plagiarism. However, you can reuse codes provided that you state the original source in your code documentations. In the case of code-reuse, you must be able to verify its output and explain the algorithms that is implemented by the reused code. You would not get any credit, if you fail to answer these questions.

The codes must be modularized. It must be accompanied by proper documentation, i.e. for most of methods the input, output and functionality must be clearly stated in the source code.

Your report must state your general findings, justify your approach, and report an evaluation. If you do team work, the role of each student must be clear. Based on our impression of the individual performance of a team member, variations in the marking withing a team is possible.

At the end of the semester[1], each project will be presented by student(s). Students are expected to explain their project work, discuss their finding and answer questions from the tutor as well as other students.

Obtaining assistance is acceptable and encouraged. However, please be informed that **plagiarism** won't be tolerated. In the case of plagiarism, the involved students will fail. Similarly, if a student or a member of a group can not explain what he/she did in detail, it will be assumed that the work has been done by somebody else and involved people will fail.

[1] At begin of the new semester, in case of the second deadline