Text Mining Practical: Project Ideas

Behrang QasemiZadeh
The Chair of Digital Libraries and Web Information Systems
Prof. Dr. Siegfried Handschuh

Winter Semester, 2014-2015

1 Introduction

This document suggests a number of project titles for the text mining project course (winter semester). 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 use 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 emphasis on 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 (as it is used in this document). For your final project report, you are expected to extend your project proposal document with a method and discussion section followed by a conclusion. A good start to your project writing is the study of relevant chapters of Day et al. [2011].

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

2 Author Profiling: Project Ideas

Author profiling is a an area of text analytic that tries to uncover characteristics of authors of texts. Examples of these characteristics are the age, gender, native language, or even more sophisticated socio-psychological aspects of the writer's character such as personal traits, feeling and behaviours. Author profiling algorithms predict these characteristic based on the style of written text, i.e., for instance lexical or grammatical choices that are made by authors. In authorship profiling, therefore, the hypothesis is that authors with same linguistic profiles share similar characteristics, e.g. people of the same age group use the same kind of slang. Consequently, the use of language is taken as an evidence to classify authors into categories (which is the subject of study in forensic linguistics).

Author profiling can be used in a number of applications. In marketing, companies are interested to know more about people who write positive or negative product reviews. Similarly, a company might be interested to know about certain characteristics aspect of its customers, e.g. by analysis of their blogs, tweets and Facebook comments as well as reviews that are left for their product. The output of these algorithms therefore can be used in trending applications such as personalisation, in recommender systems and personalised advertising. In security and forensics applications, author profiling can be employed to ensure children's safety in online platforms, e.g. in chat-rooms or other social web platform. In this context, methods of author profiling can be used to identify people with false identity, e.g. to protect children from sexual predators who hide their true selves in these communities. Similar techniques can be used to find the unknown author of a text, i.e. the method used to uncover the identity of J. K. Rowling from her recent book "The Cuckoos calling".

We offer a number of projects in the area of author profiling in the areas of prediction and data collection.

2.1 Personality Trait Recognition from Micro-blogs

In psychology, the *Big Five personality traits* are five broad domains or dimensions of personality that are used to describe human personality: openness, conscientiousness, extraversion, agreeableness, and neuroticism[1]. Project work involves processing the shared task dataset of the computational personality recognition workshop [Celli et al., 2013][2]. Project topics on this dataset included, but are not limited to, the following areas:

- Prediction of the personality traits from Facebook status updates using text-based features: the project involves the use of machine learning methods for the automatic identification of personality traits using textual features. The project work involves both implementation of the system as well as its evaluation using the provided data.

[1] See "Big Five personality traits" Wikipedia article: http://en.wikipedia.org/wiki/Big_Five_personality_traits

[2] Data is available from here: http://mypersonality.org/wiki/lib/exe/fetch.php?media=wiki:mypersonality_final.zip

- Feature analysis for the prediction of personality traits: the project involves the definition of a large number of text-based feature and study their relevance to the prediction task. This project topic is similar to the the prediction task, however, the emphasis will be on the study of correlation of text-based features and certain type of personality traits.

2.2 Identification of Age and Gender from

For this topic, students are expected to implement a system for the identification of the age and gender of the author of a text. The dataset for this project is the corpus from the Author Profiling task in the PAN workshop 2014[3]. Students who choose this topic will have a chance to compare the performance of their methods with the official participants of the task. Similar to Personality Traits Project, the focus is on author profiling in social media.

2.3 Data Collection App for LR construction

A major bottle neck in the development of the the above mentioned systems in real world applications is the lack of language resources. We invite students to build tools for automatic construction of language resources. Project topics on this dataset included, but are not limited to, the following areas:

- Exploiting Emoticons for collecting text with an affective state: for this project, students are expected to use emoticons in order to create a database of affective statements from social networks. The targeted states are:
  - Emotion: angry, sad, joyful, fearful, ashamed, proud, elated;
  - Mood: subjective feeling such as cheerful, gloomy, irritable, listless, depressed, buoyant;
  - Interpersonal stances: affective stance toward another person in a specific interaction e.g. friendly, flirtatious, distant, cold, warm, supportive, contemptuous;
  - Attitudes: states towards a person or an object such as liking, loving, hating, valuing, desiring;
  - Personality traits: enduring beliefs such as nervous, anxious, reckless, morose, hostile, jealous.

  The collected data must be used to build frequency profiles of adjective, nouns and verbs that are common with each of the above stated category of Emoticons.

[3] `http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html`

- Facebook app for creating language resources for author profiling: myPersonality Project[4] is an example of a successful Facebook app that has been used for the collection of data from more than 4 millions of Facebook users. In this project, Facebook users are invited to do a psychological test that reveals certain characteristics of them. The result of the test, together with a lot of other information, e.g. Facebook status updates are collected for research purpose. We invite interested students to design similar app to collect and verify emotional states of user's and their status or Facebook post. Similar idea can be used to annotate any other kind of textual data. For example, you may develop a Facebook app that automatically generates Emoticons/Emotional state from a given Facebook status, in which there will be the possibility to collect user's input on the suitability of the generated Emoticons.

3 Text Classification

Text classification is a classic task in natural language processing. Given a text document and a set of categories, the text classification algorithm must assign the text document into one or more of the given text categories. We discuss classic text classification techniques in our course. As a project title in the domain of text classification, we suggest the task of short text classification, which is more challenging than classic document classification. In the proposed project title, we use the "Twitter Political Corpus"[5]. Given a tweet, the task is to distinguish political tweets from non-political ones.

4 Distributional Semantics and Word Meanings

Distributional semantics embraces a set of methods that decipher the meaning of linguistic entities using their usages in large corpora. n these methods, the distributional properties of linguistic entities in various contexts, which are collected from their observations in corpora, are compared to quantify their meaning Q. Zadeh and Handschuh [2014]. We suggest two project titles in this area.

4.1 Find the missing word in every sentence from multiple choices

Given a sentence with a missing word, the aim is to develop a method that automatically completes the sentences in a meaningful and coherent manner. The participants for this project are expected to develop a vector space model from a large corpus and evaluate a number of variables in this model. We use the MSR Sentence Completion Challenge dataset in this project Zweig and Burges [2011].[6]

[4] http://mypersonality.org/wiki/doku.php

[5] http://www.usna.edu/Users/cs/nchamber/data/twitter/

[6] http://research.microsoft.com/en-us/projects/scc/

4.2 Paraphrase and Semantic Similarity

The ability to identify paraphrase, i.e. alternative expressions of the same meaning is useful for a number of natural language processing applications. We suggest a paraphrase identification task based on the Microsoft Research Paraphrase Corpus[7]. Given a pair of sentences, the aim is to classify them as paraphrases or not paraphrases. A list of the state of the art publications on this subject can be found on ACL Wiki[8].

5 Data-driven conversation simulation: a chat-bot

The development of natural interactive systems has been a research goal in the area of human language technology and natural language processing since 1960s. The research in natural language dialogue systems aims to develop methods and models that can understand and generate linguistic expressions, in which cooperation and planning of tasks requested by users are inherent parts. The research topic in this area embraces a set of technologies amongst them spoken dialogue systems, Natural Language Database Queries, Simulated Conversation [see McTear, 2004, chap.1, for a brief history].

Simulated conversation is an area of research that focuses on the development of systems that simulate conversational interaction, which is dated back to Alan Turing's "imitation game". ELIZA, perhaps, is the most well-known simulated conversation system that uses a number of patterns, keywords and associated responses to them for the simulation of a conversation between a patient and a psychotherapist. Recently, Nio et al. [2014] and Banchs and Li [2012] suggest a data-driven approach for the development of a conversational system that exploits a large dataset of TV dialogues. In the same line of research, we suggest a similar project title to exploit the Movie-Dic dataset [Banchs, 2012]. Movie-Dic is a movie dialogue corpus collected for research and development purposes. In this project, we suggest the development of a statistical conversational systems.

6 Information Extraction

Information extraction (IE) is a natural language processing task that aims to automatically extract structured information from unstructured text. In their simple form, IE tasks involves the extraction of named entities and relationships amongst them. Named entities are text units such as the names of persons, organizations, locations, expressions of times, quantities, etc. A variety of techniques can be used to extract named entities. A relationship extraction task then deals with the detection and classification of semantic relationships between named entities, e.g. `has-birthday` can be a relationship between a `person` and a `date`. We offer one project title in this area.

[7] http://research.microsoft.com/en-us/downloads/607D14D9-20CD-47E3-85BC-A2F65CD28042/default.aspx

[8] http://aclweb.org/aclwiki/index.php?title=Paraphrase_Identification_%28State_of_the_art%29

6.1 Do it your (DIY) own Google Knowledge Graph: Extracting a Knowledge Graph from Wikipedia

Knowledge graphs are emerging as fundamental tools for building advanced intelligent applications such as question answering and semantic search. If you type a query in Google such as: Who is the wife of Barack Obama? you will get Michelle Obama as a straight answer. Behind the Google search engine there is a knowledge graph, a data structure which represents factual world knowledge at scale. The project consists of an information extraction platform which aims at extracting a knowledge graph from Wikipedia text. The goal is to extract entities and facts from Wikipedia, organizing them into a knowledge graph in the RDF (Resource Description Framework) format. In order to achieve this goal, different tasks in NLP will need to be addressed such entity recognition and linking, relation extraction, word sense/entity disambiguation, among others. In this task we will be focusing on the English Wikipedia as a corpus.

The final project outcome in an interesting and meaningful output: a platform for extracting knowledge graphs out of factual text and the Wikipedia knowledge graph as well. If you project succeeds you can help other people to build their own intelligent applications.

7 Project Assessment

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

- Project Codes (45% of final mark)
- Project Report (40% of final mark)
- Presentation (15% of final mark).

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 (15% of final mark).

The deadline for the completion of the project works is the end of January, 2015. You can work in a team of up to two students.

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.

At the end of the semester, 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 be graded 5. 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 and involved people will be graded by 5.

References

Rafael E. Banchs. Movie-dic: A movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 203–207, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2390665.2390716`.

Rafael E. Banchs and Haizhou Li. Iris: A chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 37–42, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2390470.2390477`.

Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. Workshop on computational personality recognition: Shared task. In *Proceedings of the Workshop on Computational Personality Recognition*, 2013.

Day, A. Robert, and Barbara Gastel. *How to Write and Publish a Scientific Paper*. Cambridge University Press, 7th edition edition, 2011.

Michael F. McTear. *Spoken Dialogue Technology: Toward the Conversational User Interface*. Springer, 2004.

Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, Mirna Adriani, and Satoshi Nakamura. Developing non-goal dialog system based on examples of drama television. In Joseph Mariani, Sophie Rosset, Martine Garnier-Rizet, and Laurence Devillers, editors, *Natural Interaction with Robots, Knowbots and Smartphones*, pages 355–361. Springer New York, 2014. ISBN 978-1-4614-8279-6. doi: 10.1007/978-1-4614-8280-2_32. URL `http://dx.doi.org/10.1007/978-1-4614-8280-2_32`.

Behrang Q. Zadeh and Siegfried Handschuh. Random manhattan integer indexing: Incremental l1 normed vector space construction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1713–1723, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1178`.

Geoffrey Zweig and Christopher J.C. Burges. The microsoft research sentence completion challenge. Technical Report MSR-TR-2011-129, December 2011. URL `http://research.microsoft.com/apps/pubs/default.aspx?id=157031`.