INLS 613: TEXT MINING

Term Project

Objective

The goal of the term project is for students to gain practical experience with a particular text mining task. Example tasks include topic categorization, meta-data extraction, information extraction, sentiment analysis, and text-based forecasting. You can choose any text mining task as far as it involves some kind of predictive or exploratory analysis of text. This is a semester-long project, so you are encouraged to choose an area that will really interest you!

Overview

Overall, the project team is expected to: (1) select a particular text mining task; (2) conduct a literature survey of some of the most recent and successful solutions to the problem; (3) find or produce a dataset that can be used for experimentation; (4) design a program or use an existing toolkit to test one or more hypotheses; (5) do error analysis; (6) report your findings. The goal is not for your team to invent a new algorithm, but rather to apply existing techniques to a new text mining application, or explore different feature representations on an existing task.

Guideline

- Form groups of 2-3 people. If you wish to work individually, please consult with me first. All groups will be evaluated using the same standards regardless of the number of members in the group. Thus, working alone means more work.
- As a group, you are expected to divide the workload evenly among your team members. Please note the type of contribution (e.g., literature review, data cleaning, and programming) and the degree contribution (e.g., 50%) of each team member in the final term report. If a student is reported as doing disproportionately less work than the rest of the team, I will schedule a meeting with the individual and/or with the group before taking any action. We need to avoid such situations.
- The term project accounts for 40% of your grade and is associated with three deliverables: a project proposal (5%), a project report (25%), and an in-class project presentation (10%). These are described in details later. All due dates are specified in the course syllabus.

Project Proposal (5%)

The goal of the project proposal is to provide early feedback on your project plans and to determine whether the scope of your project is appropriate and doable. The project proposal should include the following information.

- A description of the problem you wish to investigate.

- A list of 6-8 papers you plan to survey as part of your literature review. For each paper, provide a brief justification for why you selected it.
- A description of the purpose of your experiments. What are you testing?
- A description of the dataset you will use to conduct your experiments. If you are using an existing dataset (highly recommended), provide a short description and a hyperlink so that I can review it. If you plan to collect our own dataset, provide a short description of how you will collect it (e.g., By scraping text from the web? Are you going to use Twitter API?)
- A description of any risks or limitations associated with your proposed plan

Project Report (25%)

Your project report is the main deliverable and should be about 8-10 pages single-spaced or 16-20 pages double-spaced. As a guide, your project report should contain the following sections and address the following questions.

- **Introduction**: What text mining application are you investigating and why is it important or useful? What is the purpose of your experiments? What are you testing? Why are your experiments interesting from a text-mining perspective?
- **Related Work**: How have others attempted to solve the problem you are addressing? How is your chosen approach different from these methods?
- **Approaches**: Please describe the details of your approaches.
- **Evaluation Methodology**: What and how are you measuring performance?
- **Experimental Results**: What are your results?
- **Discussion**: Pick a few of the most meaningful and/or puzzling results from the previous section and try to explain why they happened. Error analysis is better than speculation. If you can only speculate, try to be as specific as possible.
- **Conclusion**: What did you find and why is it important?

Project Presentation (10%)

The goal of the project presentation is to give you experience with public speaking and to give us the opportunity to learn from what you did. Your presentation should be between 15-20 minutes. Don't try to fit everything you did during the semester into your presentation. Please just highlight the most important details.

As a guide, the outline of your presentation should resemble the outline of the paper: introduction (1-3 slides), overview of related work (1-3 slides), description of the approaches tested (4-5 slides), methodology (2-3 slides), results (3-5 slides), discussion (2-4 slides), and conclusion (1-2 slide).

Example Topics

The following are some example topics that would be appropriate for the project.

- **Opinion Mining**: Choose a particular product or service, for example, books or restaurants. Find a dataset of reviews for that product. Try to learn a model to detect whether a review expresses

a positive or negative opinion. Explore different feature representations and discuss what works, what doesn't work, and why.

- **Opinion Mining across Domains**: Choose several domains, for example, laptop computers, cars, and smartphones. Find review datasets for all. Try to learn a model using reviews about one product and apply it to reviews about a different product. Does it work or not? Why or why not? Does it work better between some pairs of products than others? Why? Explore different feature representations that allow a model to generalize better from one domain to another.
- **Predicting Usefulness of Reviews**: Collect a data set of reviews and their usefulness/helpfulness votes. Try to produce feature representations and to build a model that predicts the usefulness/helpfulness of those reviews. Explore different feature representations and explain what works, what doesn't, and why.
- **Discussion Group Analysis**: Gather data from an on-line discussion/support group. Build a model to predict whether a post gets a response. Alternatively, build a model to predict the number of responses to a particular post. Explore different feature representations and discuss what works, what doesn't work, and why. Or, do the opposite: try to retrospectively predict whether a post will be the last post in its thread.
- **Twitter Retweets**: Collect tweets from news publishers and re-tweets about those tweets (I can show you how to do that). Try to predict the number of retweets for a particular tweet. What are useful features in predicting the number of retweets? Why?
- **Predicting Stock Price**: Gather daily tweets about a particular company as well as stock price data. Build a model that predicts whether the stock price will go up or down based on previous tweets about the company. This is a very difficult problem. However, the point is to learn about the problem and not necessarily to solve it.
- **Detecting Age-Appropriate Language/Content**: Find a data set of texts and their age-appropriate ratings. Try to build a model that predicts the age-appropriateness of a span of text. Explore different feature representations and discuss what works, what doesn't, and why.

Choosing a topic

Chapter 13 (and in particular, Section 13.10) in the Witten, Frank, Hall, and Pal book (Chapter 9 and Section 9.9 in the 3rd edition) discusses many forward-thinking data mining problems and applications and contains many references that you might want to consider as starting points. Additionally, there are a number of conferences that cover text data mining and related research areas.

All these conferences are held once a year and most of their yearly proceedings are available through the Association of Computing Machinery Digital Library (ACM DL). You should have access to the ACM DL from within the UNC network.

- **Search Engines and Search Technology**: SIGIR (Information Retrieval), CIKM (Information and Knowledge Management), WWW (World Wide Web), WSDM (Web Search and Data Mining), TREC (Text Retrieval), INEX (XML Retrieval)
- **Digital Libraries and Information Science**: JCDL (Digital Libraries), ASIS&T (Association for Information Science and Technology)

- **Natural Language Processing**: ACL (Computational Linguistics), NAACL (Computational Linguistics), HLT (Human Language Technologies), TAC (Text Analysis)
- **Human-Computer Interaction**: CHI (Computer-Human Interaction), Ubicomp (Ubiquitous Computing)
- **Computer-Supported Collaboration and Learning**: CSCW (Computer-Supported Collaborative Work), CSCL (Computer-Supported Collaborative Learning)
- **Social-Media**: ICWSM (Weblogs and Social Media and Text-based Forecasting)

Getting Data

Here are some data resources available online.

- Lillian Lee at Cornell University maintains several datasets related to opinion/sentiment analysis, discourse analysis, text summarization/simplification, and other NLP applications.
- UC Irvine maintains a large number of datasets, some related to text data mining.
- Charles Sutton at the School of Informatics at the University of Edinburgh compiled a list of datasets, some related to text data mining.
- Twitter API (I can share a Java code in a very primitive form)
- Reddit API (I can share a Python code in a very primitive form)
- Yelp Challenge Dataset (I can share a Parser that converts JSON to plain text)
- Kaggle
- AWS Public data sets
- Yahoo Research Data
- CrowdFlower
- List of Public Data
- Stanford Large Network Dataset Collection
- Cornell movie review data

Tools

Here are some tools available for free (if not, at least to students).

- LightSIDE, Carolyn Rose et al. at Carnegie Melon University provides a free and open text mining tool bench. This will be used for assignments and a tutorial will be provided.
- RapidMiner is a data science platform that offers an integrated environment for data preparation, machine learning, text mining, and predictive analytics. It's a commercial product, but students can use it for free.
- GATE, General Architecture for Text Engineering, from the University of Sheffield, is a commonly used software tool.
- NLTK is an open-source Python-based natural language toolkit.
- Stanford CoreNLP is one of the most acknowledged open-source natural language processing software based on Java.
- WEKA, a data mining tool that can be used for text mining. A tutorial will be provided.

Tips

- Form groups with diverse skills and interests. At least one member of your group should be a strong programmer and one member should have an interest in developing the literature review

and be good at coordinating work, resources, and schedule. Capitalize on the fact that everyone has strengths!

- When conducting the literature review, make sure you organize the material at a high level. Do the existing solutions to the problem fall under different general categories? Provide a bird's eye view before delving into the details. Try to avoid describing the existing approaches in the form of a list, without any higher-level organization.
- Make sure your literature review talks about evaluation. How are existing approaches typically evaluated and what metrics are used to measure performance and progress? Make sure that your evaluation methodology is consistent with how others have evaluated their solution to the problem.
- Error analysis is important. If you try something and it doesn't work, you can still make a big contribution by trying to determine why it doesn't work. Developing a prediction model which is very good is not a goal of this course. Instead, you need to learn from the error.
- Start early (this is most important because we only have 5 weeks) and have fun (this is also very important).