INLS 613 Text Data Mining
Term Project

Objective

The goal of the term project is to gain practical experience with a particular text mining task. Example tasks include topic categorization, meta-data extraction, sentiment analysis, and text-based forecasting. You are free to choose any text mining task as long 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 is interesting to *you*!

Overview

At a high level, your 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 you 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.

Guidelines

- 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 between your team members. Unless is it brought to my attention, and unless the case is very extreme, everyone in the group will be given the same grade. 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. Like you, I would much prefer 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 next. All due dates are specified in the course website.

Project Proposal (5%)

The goal of the project proposal is for me to give you early feedback on your project plans and to determine whether the scope of your project is appropriate. The project proposal should include the following information.

- A description of the problem area 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 (3-5 sentences).

- 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? Will you annotate it manually?).
- A description of any risks 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. Project reports that are too short or too long will be deducted points. 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 does your chosen approach compare to these methods?
- Approaches: What are the inner workings of your approach or the several approaches being investigated?
- Evaluation Methodology: 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 determine 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 all the opportunity to learn from what you did.

Your presentation should be between 10-12 minutes. Ten minutes is a not much time. In fact, it's very little. Don't try to fit everything you did during the semester into your presentation. Your presentation should just highlight the most important details.

As a group, you are allowed to take turns presenting, if you'd like, though not every group member has to present. If you decide to take turns, the transitions between group members should be smooth. 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 (2-4 slides), discussion (2-4 slides), and conclusion (1 slide).

Tip: Practice! Practice! Practice! Don't delude yourself into thinking that public speaking is purely a nature talent. Your presentation will drastically improve if you know what you want to say and how you want to say it ahead of time. When presenting, you are welcome to use notes if you wish, but please don't read from your notes. As an audience member, it's difficult to become engaged with a presenter that is looking down the whole time.

Example Topics

The following are some example topics that would be appropriate for the a 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 lawn mowers. 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? 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.

- **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 for 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 a 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 number of retweets? Why?

- **Predicting Stock Fluctuations:** Gather daily tweets about a particular company (I can show you how to do that) 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.

- **Hierarchical Text Classification:** Collect data from an on-line topic hierarchy such as the Open Directory Project. Build a model that classifies documents into a particular node in the hierarchy. Explore different ways of exploiting the hierarchical structure of the topic in order to improve classification accuracy.

- **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 9 (and in particular, Section 9.9) in the Witten, Frank, and Hall book 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 number of yearly conference that focus (or at least 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 (ACM) Digital Library. 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), ASIST (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 other resources available online.

- Chapter 7 in Opinion mining and sentiment analysis contains links to various publicly available datasets and resources within the broad area of opinion and sentiment analysis.
- 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.

Tips

- Form groups with diverse skills and interests. At least one member of your group should be a strong programmer, one member should have an interest for developing the literature review, and one member should be good at coordinating people and resources, putting things together, and ensuring that the group makes steady progress. 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.
- Be clear about what you know and what you don't know. If you make a claim, make sure that your results support it. It's great to provide plausible explanations for why something happens, but do not present such speculations as fact!
- Have fun.