Course Project

The course project objective is to give the students hands-on experience in solving some novel text mining problems. Teamwork is required.

Two projects types:

- Code-based project (recommended)
    - Grading criteria as discussed in the syllabus
- Survey paper that surveys at least 6 recent newly published research papers (no older than 2015) in the area of text mining- The papers should be related and discuss one topic proposing different techniques and algorithms on that topic or some variation of this.
    - The grading criteria will be similar to that described in the syllabus for the code-based project, but instead of having a grade for the code, it will be moved to the final report.

General steps

- Form a team
- Pick a topic
- Survey related work (for both projects types)
- Write a project proposal (1-2 pages) – Explain what is the problem, you are going to solve? what is your approach? how you will test your work? and set some deadlines for your team to complete the project.
- Write a report
- Submit your report and code
- Present the project

Form a team

You are **required** to work with other students as a team. **Teams may consist of up to three total students.** Teamwork not only gives you some experience of working with others but also allows you to work on a larger (presumably more important) topic.

Note that it is your responsibility to figure out how to contribute to your group project, so you will need to act proactively and on time if your group leader has not assigned a task to you. The instructor will believe all team members actively contribute to the project and the same grade will be applied to the group member (unless special treatment is required by the group members).

Pick a topic

You can either pick from a list of sample topics provided by the instructor or choose your topic. Leveraging existing resources is especially encouraged as it allows you to minimize the amount of work that you have to do and focus on developing truly your ideas.

When picking a topic, try to ask yourself the following questions:

- What is exactly the (research) problem that you want to solve?
- What kind of changes could your project make to the others?

- What would be the major challenge(s) in this problem? Any specific background or resource you have to solve the identified problem?
- What is the minimum goal to be achieved during this semester? (Try to drop everything non-essential and only keep the truly novel part.)
- How do you plan to demonstrate that method to be developed? Empirical experimentation and/or demo are required.

Survey related works

While choosing a topic, it is **very** important to be aware of whether the problem you would like to tackle has already been solved. If so, you may want to figure out where exactly your novelty is and whether novelty leads to any benefit to others. Your goal is to go beyond, rather than duplicate, the existing work. However, it is an option as well. To minimize your effort, you are encouraged to leverage existing algorithms, toolkits, and other useful resources as much as possible. The instructor can also help you check related work. Please feel free to discuss your plan with the instructor before finalizing your proposal.

Write a project proposal

You are required to write a two-page proposal before you go in-depth on a topic. In the proposal, you should address the following questions and include the names of all the team members as authors. The order among the authors' names does not matter.

- What is the problem identified in the project?
- Why is this problem important?
- Is there any related work? How different is your idea from theirs?
- What techniques/algorithms will you use/develop to solve the problem?
- How will you evaluate your work?
- List your potential contributions to this work.

Intuitively, the proposal should read like the introduction part of a regular research paper. Briefly state the background/motivation, what has been done, what is missing, how do you plan to solve it, how do you plan to prove the usefulness of your method, and summarize your contribution(s).

Work on the project

You should leverage any existing tools or methods as much as possible. For example, consider using the Lucene toolkit for indexing and searching in a large text corpus; using Stanford NLP parser or OpenNLP toolkit for text analysis; using MALLET or WEKA for classification or clustering. There are also many tools available on the Internet.

Consider documenting your work regularly. This way, you will already have a lot of things written down by the end of the semester. Besides, we strongly suggest using version control for your project! Nothing is more frustrating than losing a lot of your hard work, especially if it's close to a deadline.

Present the course project

At the end of the semester, each project team is expected to present their project in class. The purpose of this presentation is

- Let you know about others' projects.
- Give you some opportunity to practice presentation skills, which are very important for a successful career in both academia and industry.
- Obtain some feedback from others about your project.

In general, the structure of your presentation should be prepared like a conference presentation. So it should touch all the following aspects (text in parenthesis states the instructor's expectation):

- What is the background/motivation for your work? What research question will you address? *(Learn how to attract public attention.)*
- Why is this problem important? *(Learn to how to persuade others.)*
- Is there any existing work? How novel is yours? *(Learn how to sell your ideas.)*
- How did you solve this problem? *(Learn how to deliver your solution.)*
- How good was your method? *(Learn how to quantitatively/qualitatively evaluate your work.)*
- Any ideas for further improvement? *(Learn how to look ahead.)*

Think about how you can best present your work to make it as easy as possible for your audience to understand your main messages. Try to be concise, to the point. Pictures, illustrations, and examples are generally more effective than text for explaining your project. Try to show screenshots and/or plots of your experimental results. Watching some top conference presentations (e.g., KDD, SIGIR, ICML) on VideoLectures will be beneficial.

Write a project report

You should write your report as if you were writing a regular conference paper. You should address the same questions as those you have addressed in the proposal and presentation, only with more details. Pay special attention to the challenges that you have solved and your detailed solutions. Basic sections to be included in the report should be the same as those in a conference paper, e.g., abstract, introduction, related work, method, experiment and conclusion.

You report should be between 6 – 8 pages total excluding the references.

Here are the LaTeX files necessary to write the project report. And you are required to use **"ACM Standard"** or **"ACM Large"** for your report. In the same page, you can found the word format if you plan to use word **Interim layout.docx**

Topic ideas:

- Twitter Sentiment Analysis: Detecting whether a particular tweet contains negative emotions attached with it or not from a given dataset
- Sentiment analysis on Rap music albums
- Sentimental analysis on movie reviews
- Spam filtering, classifying email text as spam or not using deep neural network
- Language identification, classifying the language of the source text using deep neural network or machine learning classification algorithms.
- Genre classification, classifying the genre of a fictional story.
- Prediction of movie success using data mining
- Generating new article headlines.
- Generating suggested continuation of a sentence.
- Social media mining to get relevant information like women's or men's behavior in a social network.
- Knowledge/information extraction using data mining.
- Cybercrime prevention: The anonymous nature of the internet and the many communication features operated through it contribute to the increased risk of internet-based crimes. Today, text mining intelligence and anti-crime applications are making internet crime prevention easier for any enterprise and law enforcement or intelligence agencies.
- Customer care service: Text mining, as well as natural language processing, are frequent applications for customer care. Today, text analytics software is frequently adopted to improve customer experience using different sources of valuable information such as surveys, trouble tickets, and customer call notes to improve the quality, effectiveness, and speed in resolving problems. Text analysis is used to provide a rapid, automated response to the customer, dramatically reducing their reliance on call center operators to solve problems.
- Fraud detection through claims investigation: Text analytics is a tremendously effective technology in any domain where the majority of information is collected as text. Insurance companies are taking advantage of text mining technologies by combining the results of text analysis with structured data to prevent frauds and swiftly process claims.
- Content enrichment: While it's true that working with text content still requires a bit of human effort, text analytics techniques make a significant difference when it comes to being able to more effectively manage large volumes of information. Text mining techniques enrich content, providing a scalable layer to tag, organize and summarize the available content that makes it suitable for a variety of purposes.
- Question Answering is the problem that was given a subject, such as a document of text, to answer a specific question about the subject.

Publicly available data:

Read this article to find data from Twitter: https://lionbridge.ai/datasets/top-20-twitter-datasets-for-natural-language-processing-and-machine-learning/

Datasets for spams, recommender system, language modeling, etc. https://blog.cambridgespark.com/50-free-machine-learning-datasets-natural-language-processing-d88fb9c5c8da

Yelp, Jeopardy questions, Stanford QA dataset, Amazon products review dataset, etc. https://analyticsindiamag.com/10-nlp-open-source-datasets-to-start-your-first-nlp-project/