1. Update: difference between IST565 Data Mining and IST736 Text Mining

A number of students have asked the question “what is the main difference between the two courses: data mining and text mining?” Here is a brief answer.

The two classes share the theoretical foundation in machine learning. Therefore the fundamental concepts in machine learning, such as classification and clustering, are covered in both classes. However, these two classes differ in the following aspects:

Content wise, the data mining class focuses on structured data, meaning the data sets we play in the class are usually in .csv format. Text mining focuses on unstructured, text data, which come in words. How to convert text to numbers that still bear the meaning of text is an important topic in text mining. In text mining we will have to deal with some problems that do not exist in mining structured data, such as the subjectivity in decisions. For example, how to determine if a tweet is positive, negative, or neutral? Different people might give different assessment.

Technology wise, the data mining class uses GUI-based tool like Weka to analyze data, and thus does not require nor practice programming skills. The Text Mining class uses a python-based command line tool call Sci-kit learn, which requires students to learn/practice python programming on linux platform. Students without programming background can still take the class but are expected to spend a little more time on programming.

2. Course description and objectives

The main goal of this course is to increase student awareness of the power of large amount of text data and computational methods to find patterns in large text corpora. This course is designed as a general introductory level course for all students who are interested in text mining. Programming skill is preferred but not required in this class.
This course will introduce the concepts and methods of text mining technologies rooted from machine learning, natural language processing, and statistics. This course will also showcase the applications of text mining technologies in (1) information organization and access, (2) business intelligence, (3) social behavior analysis, and (4) digital humanities.

Upon completion of this course, students are expected to be able to:

- Describe basic concepts and methods in text mining, for example document representation, information extraction, text classification and clustering, and topic modeling;
- Use benchmark corpora, commercial and open-source text analysis and visualization tools to explore interesting patterns;
- Understand conceptually the mechanism of advanced text mining algorithms for information extraction, text classification and clustering, opinion mining, and their applications in real-world problems;
- Choose appropriate technologies for specific text analysis tasks, and evaluate the benefit and challenges of the chosen technical solution.

3. Course material

Textbooks


Software

4. Assessment

Class participation 20%
Assignments 40%
Group project 40%
**Class participation**

Class participation is important for this class. Students are required to attend classes and actively participate in class discussions and exercises. Therefore attendance will be taken at the beginning of each session. Student should also actively participate in discussions in face-to-face class as well as online discussions held in the BlackBoard discussion board.

**Assignment**

Eight assignments will be given throughout the semester, 5 points for each assignment.

<table>
<thead>
<tr>
<th>Assignment topic</th>
<th>Released in week</th>
<th>Due in week</th>
</tr>
</thead>
<tbody>
<tr>
<td>An evaluation of online sentiment analysis demos</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Does stemming or stopword removal affect sentiment classification?</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>A comparison of Benoulli and Multinomial NB models</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Lie detection</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Comparing the performance of naïve Bayes and SVM</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>A simple ensemble learning experiment</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Document clustering</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>LDA for topic clustering</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

**Sample Homework Assignments in previous semester**

**Sample Homework 1: An evaluation of online sentiment analysis demos**

Assume that you are a manager in a PR firm, and a client of your firm, e.g. Target, would like you to run an investigation about recent public opinions on Target after its data breach. You are going to license an automated sentiment analysis tool to help you with the analysis. Now you task is to evaluate the following three candidates and choose the best one for use.

http://nlp.stanford.edu:8080/sentiment/rntnDemo.html
http://www.etcml.com/
http://sentistrength.wlv.ac.uk/

To conduct the evaluation, you will need to collect a sample data that represents public opinions, such as news articles, forum posts, tweets, etc. Gather 30 text documents, and describe where and how you chose these documents as your test sample. You will also need to manually annotate these text documents in terms of their sentiment toward Target. This will be your ground truth for evaluation. Run these examples through the three demos, and compare their performance. Decide which one to choose, and explain your decision process.

**Sample Homework 2: Lie detection**

Some people claimed that machine learning algorithms can figure out whether a person is lying or not. Do you believe that? To test this claim, we have collected a collection of customer reviews, some are true some are fake, and you are going to test how good multinomial NB can be for fake review detection.
This data set also has sentiment label for each review. You will also test multinomial NB’s performance in sentiment classification.

**Sample Homework 3: Document clustering**

Fangzhouzi is a famous science writer and blogger in China. His blogs are usually about two topics, science blogs (e.g. whether GMO products are a safety threat to public health) and debate about misconduct in scholarly work such as plagiarism. In this homework, you are given a data set consisting of 31 blogs written by Fangzhouzi. Can you use document clustering algorithms like k-means and EM to separate science blogs and debate blogs (labeled as “science” and “social” respectively in the data)? Actually this data set includes three debates, each on one scholarly misconduct case. Can you use clustering algorithms to further separate these debates?

This dataset include the original blogs in Chinese and the processed feature vectors with function words as features. If you don’t understand Chinese, please invite a Chinese student to your group.

If the clustering algorithms managed to separate the two topics, can you use Weka visualization functions to examine which features contribute to the separation, and whether it makes sense or not?

**Group or individual projects**

Students are encouraged to form a small group of 2-3 persons to conduct the course project. However, individual project is also allowed if there is difficulty finding partners.

Sample project I: Recruit Amazon Mechanical Turk workers to help you annotate the sentiment (positive, negative, or neutral) of tweets. [https://www.mturk.com/mturk/welcome](https://www.mturk.com/mturk/welcome)

Sample project II: Participate in the Kaggle Sentiment Analysis Competition [https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews](https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews)

### 3 Tentative Class Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Readings before class</th>
<th>Items assigned</th>
<th>Items due</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01/14</td>
<td>Introduction</td>
<td></td>
<td>HW1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>01/21</td>
<td>Converting text to numbers</td>
<td></td>
<td>HW2</td>
<td>HW1</td>
</tr>
<tr>
<td>3</td>
<td>01/28</td>
<td>Naïve Bayes</td>
<td></td>
<td>HW3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>02/04</td>
<td>Model evaluation</td>
<td></td>
<td>HW4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>02/11</td>
<td>Support Vector Machines</td>
<td></td>
<td>HW5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>02/18</td>
<td>Faceted text classification</td>
<td></td>
<td>HW6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>02/25</td>
<td>Document clustering</td>
<td></td>
<td>HW7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>03/04</td>
<td>Topic modeling</td>
<td></td>
<td>HW8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>03/11</td>
<td>Spring Break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>03/18</td>
<td>Crowdsourcing for text annotation</td>
<td></td>
<td>Project I</td>
<td>HW8</td>
</tr>
<tr>
<td></td>
<td>03/25</td>
<td>Sklearn</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4 Course policies

Prerequisite
Curiosity about language and meaning;
Willingness to take on new challenges;
Programming skills preferred but not required.

Communications
This course will use the SU BlackBoard System as the main communication platform in and out of class time. Students are required to check their BlackBoard accounts on a regular basis. Important announcements will be posted to the Announcements board. Failure to read the class announcements will not be considered a suitable excuse for not being informed. The BlackBoard can be accessed at http://blackboard.syr.edu. Questions regarding the BlackBoard should be directed to ilms@syr.edu or Peggy Brown at 315-443-9370.

All emails to the instructor should be sent to byu.teaching@gmail.com with subject line starting with the course number. This email account is dedicated to teaching matters. All your emails will be archived for easy retrieval afterwards.

Academic integrity
The academic community of Syracuse University and of the School of Information Studies requires the highest standards of professional ethics and personal integrity from all members of the community. Violations of these standards are violations of a mutual obligation characterized by trust, honesty, and personal honor. As a community, we commit ourselves to standards of academic conduct, impose sanctions against those who violate these standards, and keep appropriate records of violations. The academic integrity policy can be found at: http://supolicies.syr.edu/ethics/acad_integrity.htm

Respect Intellectual Property Rights and cite all sources in your work. Any valid citation style may be used. The following link may be used for further information regarding appropriate citation styles: http://researchguides.library.syr.edu/citation

Student with disabilities
If you believe that you need accommodations for a disability, please contact the Office of Disability Services (ODS), http://disabilityservices.syr.edu, located in Room 309 of 804 University Avenue, or call (315) 443-4498 for an appointment to discuss your needs and the process for requesting accommodations. ODS is responsible for coordinating disability-related accommodations and will issue students with documented disabilities Accommodation Authorization Letters, as appropriate. Since accommodations may require early planning and generally are not provided retroactively, please contact ODS as soon as possible. The non-discrimination, students with disabilities policy can be found at http://supolicies.syr.edu/studs/nonD_stud_disability.htm
Ownership of student work

This course may use course participation and documents created by students for educational purposes. In compliance with the Federal Family Educational Rights and Privacy Act, works in all media produced by students as part of their course participation at Syracuse University may be used for educational purposes, provided that the course syllabus makes clear that such use may occur. It is understood that registration for and continued enrollment in a course where such use of student works is announced constitutes permission by the student. After such a course has been completed, any further use of student works will meet one of the following conditions: (1) the work will be rendered anonymous through the removal of all personal identification of the work’s creator/originator(s); or (2) the creator/originator(s)’ written permission will be secured. As generally accepted practice, honors theses, graduate theses, graduate research projects, dissertations, or other exit projects submitted in partial fulfillment of degree requirements are placed in the library, University Archives, or academic departments for public reference.