Predictive Modeling of Spacial, Textual and Network Data  
(Statistics 5931)

Zack W. Almquist  
Fall Semester, 2017

Class Schedule

Lecture:  Th  2:00 – 4:00 pm   Fordhall Statistics Library
       URL:   http://moodle.umn.edu
Note: Requires UMN login and registration in class to access.
       URL:   http://www.github.com
Note: Requires github login.

Professor

Name:   Zack W. Almquist  
Office: 372 Ford Hall  
Office Hours: By Appointment Only  
Email: almquist@umn.edu  
Telephone: 612-624-4300 (not recommended)

Course Objectives

This course will focus on learning and developing applied skills in text analysis and machine learning classification techniques. Students will be exposed to key software tools in R and python to engage in applied network and text analysis research. Students will learn to work in teams and engage in scholarly research through developing, analyzing and writing up a 6-10 page conference proceeding. All students will be expected to submit (as a team) their final project to an academic conference.

Prerequisites

This course assumes exposure to the R statistical environment and statistics on par with the Statistics sequence 3011 and 3022. More exposure to statistics or programming is
helpful, but not required.

Course Requirements

This course meets once a week on Wednesdays from 2:00 to 4:00pm and students are expected to attend regularly, do the readings, exercises and be engaged during each seminar session. At the end of the semester each student team is expected to submit a paper to a conference based on the research project completed over the semester.

Computers

It is not required that students bring their computers/laptops to lecture and lab (if one is owned), but it is highly recommended since both lecture and lab will make extensive use of the computer software R. Computer labs are available on campus, please consult with the TA if you have trouble finding the various locations of campus computer labs.

Readings

Weekly readings assignments can be found on the course syllabus. All readings are assumed to be completed before each lecture/seminar. You are expected to read over the class notes each week and make sure you are familiar with the material as the course progresses. Questions are encouraged.

Homework

Homework assignments will normally be administered on a bi-weekly basis and will be due on every other Wednesday. Homework assignments are meant to achieve three results: (1) provide practice with the statistical concepts discussed in class, and (2) provide practice with the computational and statistical programing language R and (3) provide a chance to demonstrate your mastery of material and highlight areas where more work is needed. You may work in a group, but all write-ups must be done independently. All collaborators should be appropriately cited in your write up and any detailed R code should also be provided.

Final Project

All students will be expected to work in teams of 2-3 individuals to develop, writing and analyze a project focused on employing methods and models for text and network analysis. This final project will take the form of a 8-10 page conference proceedings. Students are expected to submit their project to an academic conference.
Participation

Individuals are expected to attend every course, to have completed every reading, and to participate with questions and discussion on each topic as presented. If you plan on missing any class period you are responsible for all material and for contacting the instructor in a timely manner.

Grading

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation</td>
<td>10%</td>
</tr>
<tr>
<td>Homework</td>
<td>20%</td>
</tr>
<tr>
<td>Final Project</td>
<td>70%</td>
</tr>
</tbody>
</table>

Lectures, readings, labs, and review sessions are provided for each student’s benefit. It is the responsibility of the student to take advantage of these opportunities to acquire and demonstrate mastery of course material, so as to achieve his or her desired grade.

Letter grade assignment

A    93%+
A-   90-92.99%
B+   87-89.99%
B    83-86.99%
B-   80-82.99%
C+   77-79.99%
C    73-76.99%
C-   70-72.99%
D    60-69.99%
F    <59.99%

Required Texts

Natural Language Processing

- Jurafsky, Daniel and James Martin. 2008. Speech and Language Processing. Prentice Hall.

Machine Learning


Computer Languages


Readings

Be prepared to discuss all readings assigned at anytime in lecture/seminar.

Required Software

We will be using the R statistical programming language. R can be downloaded at http://www.r-project.org/.

RStudio IDE (Integrated Development Environment) is a software application which facilitates interaction with the R statistical programming language. It is often preferred to the GUI (Graphic User Interface) made available through CRAN. You can download it at http://www.rstudio.com/.

Latex is a word processor and a document markup language. It can be downloaded and installed on Windows (http://miktex.org/), OSX (https://tug.org/mactex/) or Linux (use the package manager of your choice).

A github account will be required of all students. One can register for a github account at https://github.com/. You can find information about how github works with Rstudio at http://z.umn.edu/rstudiogit, and github maintains a quite good help-system at https://help.github.com/.
Course Policies

Missing Class, etc.
It is expected that each member of the class will attend every lecture/discussion. If there is an appropriate reason to miss class it is expected that the individual will email or discuss in person with the instructor at least one week in advance. For any medical issues please see the UMN website for university policies.

Cheating, etc.
All work is assumed to be your own and all individuals are expected to follow the university policy on cheating and misconduct. If you have any questions please consult the UMN website for university policies.

Assignments

Homework Assignments
Homework will be assigned on a biweekly basis starting on the second Wednesday of the Semester and will be due two weeks later at 5:00pm. There will be a total of six homework assignments. Homework assignments will be graded on a 100 point basis. Each assignment must be turned in through github, no late assignments will be accepted. Homework must be turned in using github, knitr/latex and must include all R code. Your lowest score will be dropped in the final calculation of grades.

Final Team Project
The final project is Due by the end of semester and it is expected that all students will submit the team project to an academic conference.

Reading Assignments
Week 1 (09/07): Text as Data
• Readings:
• Homework/Lab:
- R Code and Questions Provided Via Github.

**Week 2 (09/14): Statistics and Computing Refresher**

- **Readings:**

- **Homework/Lab:**
  - R Code and Questions Provided Via Github.

**Week 3 (09/21): Acquiring and Manipulating Text Data**

- **Readings:**
  - NLP Stemming TBD

- **Homework/Lab:**
  - R Code and Questions Provided Via Github.

**Week 4 (09/28): Dictionary Methods: Measuring Weighted Word Usage**

- **Readings:**

- **Homework/Lab:**
  - R Code and Questions Provided Via Github.
Week 5 (10/05): Methods for Finding Discriminating Words and Applications

- **Readings:**

- **Homework/Lab:**
  - R Code and Questions Provided Via Github.

Week 6 (10/12): The Vector Space Model and Geometry of Text

- **Readings:**
  - Linear Algebra Handout

- **Homework/Lab:**
  - R Code and Questions Provided Via Github.

Week 7 (10/19): PCA, MDS and Text

- **Readings:**
  - Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning Springer.

- **Homework/Lab:**
  - R Code and Questions Provided Via Github.

Week 8 (10/26): Categorical and Dirichlet Distributions and Other Distributions on the Simplex

- **Readings:**
  - Chapter 2 Bishop, Christopher. 2006. Pattern Recognition and Machine Learning (Sections 2.1, 2.2 especially).

• Homework/Lab:
  - R Code and Questions Provided Via Github.

**Week 9 (11/02): Clustering Methods 1**

• Readings:
  - Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning Springer.
  - Chp 9. Bishop, Christopher. 2006. Pattern Recognition and Machine Learning (Sections 2.1, 2.2 especially)

• Homework/Lab:
  - R Code and Questions Provided Via Github.

**Week 10 (11/09): Clustering Methods 2**

• Readings:

• Homework/Lab:
  - R Code and Questions Provided Via Github.

**Week 11 (11/16): Topic Models 1: LDA**

• Readings:

• Homework/Lab:
  - R Code and Questions Provided Via Github.
Week 12 (11/23): Topic Models 2: sLDA

[Thanksgiving]

- **Readings:**

- **Homework/Lab:**
  - R Code and Questions Provided Via Github.

Week 13 (11/30): Classification: Naive Bayes, SVM, Ensemble Classifiers

- **Readings:**
  - D orazio et al. Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines. Political Analysis 22, 2 224-242.,
  - 7.10. Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning Springer.

- **Homework/Lab:**
  - R Code and Questions Provided Via Github.
Week 14 (12/07): Model Fit, Complexity and Cross Validation

- **Readings:**

- **Homework/Lab:**
  - R Code and Questions Provided Via Github.

Week 15 (12/14): Word Scores and Item Response Theory

- **Readings:**

- **Homework/Lab:**
  - R Code and Questions Provided Via Github.