OPER 655 - Text Mining

Course Syllabus - Fall 2019

TF 1500-1700 Bldg. 646 Rm. 214 (4.0 Credit Hours)

Welcome!

Text Mining is the organization, classification, labeling and extraction of information from structured and unstructured text data. With so much information readily available through the internet, analysts and decision makers often find themselves overloaded with data. Text mining describes a set of tools which can help analysts glean necessary information either for general understanding about a corpus of text documents, or for putting text into a form useful for the application of alternative analysis techniques.

This course introduces students to this rapidly growing field and equips them with some of its basic principles and tools as well as its general mindset. Students will learn concepts, techniques and tools they need to deal with text mining practice in a joint military context.

Course Learning Objectives

The objective of this course is to help students build an automated text analysis data pipeline that implements the techniques/processes of text mining for the purposes of making better decisions. More specifically, at the completion of the course, students should be able to:

- Understand when, where, and how to perform text mining to appropriate problems and data sets.
- Lead and work within a group of researchers analyzing a specific text mining problem.
- Use software to analyze text data and communicate the results of their analyses.
- Review code using relevant R and/or python packages in text mining.

In this year's course the pipeline will be constructed using the **1.4 million cell phone reviews: ratings and reviews for all brands of cell phones** dataset. The idea for will be to build a text analysis pipeline composed of modules that can be used to implement specific methods to a dataset in whatever sequence is required to achieve some desired result(s).

Pre-requisites & Co-requisites

Unless otherwise waived by the course instructor, students are required to have completed the following courses prior to enrolling in OPER 655 - Text Mining.

- None

Requests to waive these required courses must be made in accordance with AFIT/EN policy guidelines and must be approved by the course instructor. Waivers my be approved for students that have has successfully completed a similar courses or if special arrangements are made to meet AFIT requirements.

Instructor and Contact Information

- Maj. Jason Freels
- Email: **jason.freels@afit.edu**
- Email: **auburngrads@live.com**
- Phone: (937) 255-3636 ext. 4676
- Cell: (937) 430-6619
- Office: Bldg. 641, Rm. 245B

- Office hours: By appointment

Course Resources (Required & Optional)

There is no required text for this course however, there are a number of freely available resources that I will reference throughout the quarter. These resources are listed below (with links provided). Other references will be provided using the course **Github Page**

- **Daniel Jurafsky** Online textbook for Dan Jurafsky's NLP course at Stanford
- **Julia Silge and David Robinson**Text Mining with R, A Tidy Approach, Published with Bookdown, 09/2018
- **Daniel Jurafsky** YouTube videos to accompany the textbook for Dan Jurafsky's NLP course at Stanford
- **CRAN Task View: Natural language Processing**Large index of R packages related to NLP

Grading Policy & Course Deliverables

Your final course grade will be determined according to the following requirements and their respective weights. These deliverables are described at the bottom of this syllabus.

- 40% **Group Activities**
- 60% **Project**

In this course, grades will be assigned according to the values shown in the table below. The following sections detail the requirements of each graded deliverable.

Numerical grade range	Corresponding letter grade
(1.00 - 0.93]	A
(0.93 - 0.90]	A-
(0.90 - 0.87]	B+
(0.87 - 0.83]	B
(0.83 - 0.80]	B-
(0.80 - 0.77]	C

Github

Information about course requirements distributed and collected using GitHub. Thus, each student will first need to create an account on **GitHub**, if they don't already have one.

> Tip: Don't use your `awesome.student@afit.edu` email address to create your account. Use an email address that you'll have access to after you graduate. If desired, you can add your AFIT address as a secondary email.

After you've created your account on GitHub, click **HERE** to access the course "Roll Call". Read and follow the instructions to post your comment so I know that you've successfully created an account.

Software/Computer Programming

A key component of this course involves developing the skills and knowledge to create **reproducible & dynamic** data products to present results from your research. In previous offerings of this course, I allowed students to use any software package to complete their assignments. This became difficult, for the students to complete their work and for me to grade them. So, I've decided to require you to use the R programming language to complete and submit your assignments.

I realize that some of you may be new to coding or may have never coded before. Don't worry, you don't need an extensive background in R, LaTeX, or HTML to be successful in this course. Each week I'll provide you with LOTS of code examples that you can copy/paste to help you get more familiar with the software. Also, many created several demo presentations to get you up to speed and I'm always willing to help out when needed. The first link provided for Week 1 in the schedule below walks you through the process of getting the R/RStudio tool-chain installed and ready for the course.

Each of the following tools will be used this quarter:

- R Project for Statistical Computing
- RStudio IDE
- Python (via Anaconda)
- Git/Github
- LaTeX/Mathjax
- Pandoc Markdown
- HTML$_5$, CSS3, and JavaScript (don't need to know these - already built in!)

Course Schedule

The course schedule below is **TENTATIVE** and is therefore subject to change. Any changes made to this schedule will be clearly communicated by the instructor.

Week	Dates	Lesson Description	Learning Material	Presenter
1	09/30-10/04	Course intro/Software intro	DataCamp & Github	Maj Freels
2	10/07-10/11	Extracting data from documents		Maj Freels
3	10/14-10/18	Regular expressions & text parsing		Maj Freels
4	10/21-10/25	Methods for word relationship analysis		Maj Freels
5	10/27-11/01	Topic modeling		Maj Freels
6	11/04-11/08	Sentiment analysis	Student Led	Student Group
7	11/11-11/15	Document summarization	Student Led	Student Group
8	11/18-11/22	Named entity recognition	Student Led	Student Group
9	11/24-11/29	Parts of speech tagging		Maj Freels
10	12/02-12/06	In class project work	None	Maj Freels
11	12/09-12/12	Student project reports	None	Student Group

Other Noteworthy Dates

In addition to the course schedule, you should take note of the following dates.

- 14 Oct - Columbus Day (No Classes)
- 11 Nov - Veteran's Day (No Classes)
- 28 Nov - Thanksgiving Day (No Classes)
- 29 Nov - AETC Family Day (Classes will meet - TBD)
- 09 Dec - Fall Quarter Classes End
- 10 Dec - Final Exam Week Begins
- 13 Sep - Final Exam Week Ends

Important Policy Statements

Academic Integrity Policy Statement

All students must adhere to the highest standards of academic integrity. Students are prohibited from engaging in plagiarism, cheating, misrepresentation, or any other act constituting a lack of academic integrity. Failure on the part of any individual to practice academic integrity is not condoned and will not be tolerated. Individuals who violate this policy are subject to adverse administrative action including disenrollment from school and disciplinary action. Individuals subject to the Uniform Code of Military Justice may be prosecuted under the UCMJ. Violations by government civilian employees may result in administrative disciplinary action without regard to otherwise applicable criminal or civil sanctions for violations of related laws. (References: ***Student Handbook, ENOI 36-107, Academic Integrity***)

Attendance Policy Statement

Attendance at all class sessions and exams is mandatory for military and civilians assigned to AFIT as full-time students except for extenuating circumstances. Part-time students are expected to attend scheduled classes, and absences should be explained to the instructor. The student should provide advance notice, if possible. Scheduled classes and exams are defined by the instructor and they are documented in the course schedule. (References: ***Student Handbook, Graduate School Catalog***)

Academic Grievance Policy Statement

AFIT and the Graduate School of Engineering and Management affirm the right of each student to resolve grievances with the Institution. Students are guaranteed the right of fair hearing and appeal in all matters of judgment of academic performance. Procedures are detailed in ***ENOI 36-138, Student Academic Performance Appeals***.

Additional Notes

My Teaching Philosophy

As **AFIT graduates**, you should be expected to know how to approach and solve real-world problems AND present your results in a meaningful way so that decision makers can make defensible decisions.

As **AFIT instructors**, we do a disservice to our students by not teaching new and improved ways to produce and share your results. Further, we do a disservice by teaching you to solve problems using tools that you won't have access to after leaving AFIT. Therefore, I re-built this course using the R/Python/RStudio tool-chain to help you produce better results…faster.

Challenge your instructor

Challenging your instructor can often be a good thing. If you can't trip me up in this class from time to time, you're not trying. Discussion leads to a more interesting class, so questions are always good.

Course deliverables

In-class group activities

Many data science projects cannot be completed by one person - especially if the project has a deadline. Thus, it's important that every analyst be able to work with other analysts in a team - and even with other teams of analysts. In this course students will exercise their teamwork skills by completing two activities as described below. Teams members must work together and must develop their method for working successfully within a group.

For both of the assigned activities students will be separated into teams - the number of students in each team will depend on the number of students registered for the course. Each team will address one of the following text mining goals:

- Sentiment analysis
- Document summarization
- Named entity recognition

Student teams will be assigned to lead class discussions for both class meetings during a one week of the quarter - as specified in the course syllabus. What the teams should do during these class meetings is described below. My goal is to balance the number of students in each team - which means that some students may not be assigned to their preferred topic.

Activity #1 Overview of Current Methods (20 pts):

During the first class meeting of their assigned week, teams will present a detailed overview of the current methods used to accomplish their assigned goal. Teams should present the conceptual and mathematical theory behind several methods used to accomplish their chosen text mining goal.

This activity is worth 20% of your overall course grade. Student grades for this activity will be graded based on (1) completion of the assignment, and (2) the degree to which each student contributed to the team in completing this activity. I will assign this value based a survey each student will fill out to describe the level of participation for each member of the team. The survey will be used to distribute the grade appropriately as demonstrated by your effort towards assisting the group accomplish the assignment. Survey scores will be based on a scale from 1 to 5, where a ‘5’ means that the student was very involved in the effort and made important contributions a ‘1’ means that a student was not involved and made few (if any) contributions.

Activity #2 R/Python Programming Tutorial (20 pts):

During the second class meeting of their assigned week, teams will give a tutorial on how to use the R and/or Python programming languages to implement the assigned text mining method on the cell phone dataset. There are a number of excellent resources to help with this project.

- **Stack Overflow**
- **CRAN Task View: Natural language Processing**
- **Github**
- **Kaggle Kernels**

Students will cite all sources from which they obtained any examples used in their tutorial and will clearly indicate which packages they used obtain their results.

This activity is worth 20% of your overall course grade. Student grades for this activity will be graded based on (1) completion of the assignment, and (2) the degree to which each student contributed to the team in completing this activity. I will assign this value based a survey each student will fill out to describe the level of participation for each member of the team. The survey will be used to distribute the grade appropriately as demonstrated by your effort towards assisting the group accomplish the assignment. Survey scores will be based on a scale from 1 to 5, with a ‘5’ meaning that the student made important contributions to the effort and a ‘1’ meaning that a student was not involved and made few (if any) contributions.

The group assignments are as shown in the table below

Named Entity Recognition	Sentiment Analysis	Document Summarization

Named Entity Recognition	Sentiment Analysis	Document Summarization
tyler-spangler	bjhufstetler	clarence0083
ACGidds	anson25cheng	treypujat
maxwell2818	hrdCory	mnschro6

Final project

This course requires completion of an individual project that applies the data pipeline developed in the class. The intent of the project is to apply the concepts and methods learned in the class to an actual data set in order to perform knowledge discovery. **The dataset must be approved by Major Freels, NLT 21 October 2019.** Students may use an application within their thesis project for work within this class. The data set used for this project can be either pre-formatted text or something in textual form. At a bare minimum, the text should include more than 5,000 words and should be in English. While these requirements are minimums, the intent is that the problem is analyzed according to the concepts presented in this course and that the results are articulated in a clear manner.

Deliverables

This project is worth **60% of your total grade** for this class. You will provide two deliverables.

1. **Project IPR - Due Monday, 3 November (10 pts) -** This should be a focused 4-5 slide presentation (no more than 5 slides). The goal of this presentation is to describe what task you are doing and where you are in your analysis. You will not present this presentation; make sure the slides can speak for themselves. The slides should include the following:

- Title Slide
- Purpose
  - What are you doing?
  - Why?
- Background
  - Context into your analysis
  - Insight into your topic with rudimentary analysis
- Interim Analysis
  - Work not used in any other class previously
- Discuss Next Steps

2. **Project Report - Due Friday, 6 December (50 pts) -** For this part of the project assignment, white a short description of your text analysis, including how it informs on the problem laid out in your project proposal. Focus on insights, trends and analysis. This is where you are allowed to put the work you've done, but remember this document should flow. Your document should include any charts/graphs relevant to the discussion imbedded and referenced in the body of the document. Bulky or large figures or charts and examples of your code should be put into the appendix(cies). The format of the report should have at least the following headings:

- Abstract
- Problem Background (ending with a clear statement of the problem)
- Methodology (describe the data you used and your analysis)
- Findings and Conclusions (to include any future work)

Think of this report as something that could be read by a research sponsor. All reports will be made available to the rest of the class (unless there are circumstances that might prevent this).