# Text Mining and Natural Language Processing [IAL 620]

Course Instructor: Aaron Beveridge

Office: 500 Forest Building, Office 113

Email: [akbeveri@uncg.edu](mailto:akbeveri@uncg.edu)

Office Hour: By appointment

## Overview

As a core course in the Cultural Analytics concentration for the MS of Informatics and Analytics at UNCG, IAL 620 explores and complicates a simple question: how do we extract the most important or meaningful information from a collection of documents (corpus)? Obviously, answers to questions of “importance” and “meaning” are framed by the researcher asking the questions and by the inherent limitations in the data explored. In this case, the primary form of data will be text–words and phrases, how they are interrelated, and the documents in which they are contained. For this course students create a unique corpus using web scraping methods, and then students describe and analyze their corpus using text mining and natural language processing.

### Student Learning Objectives

* Recognize and apply text mining and natural language processing methods
* Recognize relevant issues pertaining to individual privacy rights and intellectual property for web scraping
* Identify and discriminate among various methodologies for cleaning, organizing, describing, and comparing natural language data

## Course Content and Grading

### Workshops – 30%

For the workshops, students will complete tutorials and show weekly progress toward the completion of their course project. For most of the tutorials, students will work through code examples and learn how to apply the methods to their own corpus/project. In other words, it is not enough to simply work through a tutorial and copy the example code as shown, but students should modify code examples and provide clear documentation on how they are adapting various methods in completing their own individual project.

### Course Project Proposal – 20%

For the course project, students produce a “proof of concept” project utilizing a corpus of their choosing. These projects will act as a preliminary test for the types of projects students may develop more thoroughly during their final capstone for the MSIA program. The project proposal will provide (1) a short abstract for the project, (2) a description of the corpus and the possible insights to be drawn from it, (3) a list of the types of analytics (descriptions, data visualizations) the student will attempt to produce in completing the project, (4) and potential industry/commercial/institutional applications for the work proposed. The goal is for students to begin building a portfolio of work and to get a sense of the types of projects they might undertake during their capstone.

### Final Presentation (final report on course project) – 20%

The final presentation serves as the “final” for the course. Drawing from the project proposal and work completed in the final weeks of the course, students will present a short presentation of their semester long project. The goal for the final presentation is to provide an honest accounting of the work completed, and to explain any successes and failures that resulted from the work. Because the course projects are intended to act as a test run for projects students may undertake during their capstone for MSIA, the final presentation is not meant to “sell” so much as “describe.” In other words, a failed project does not equate to a failed grade on the final presentation. The goal is to encourage risk-taking, innovation, and an expansion of student competencies through the completion of projects.

### Ethics Response – 10%

The ethics response is an informal, personal essay where students consider the ethical implications of text mining and natural language processing more generally, and the essay also provides students an opportunity to consider the ethical implications for their own individual projects. Students will respond to writing prompts provided by the instructor and no formal or additional research is required. The goal is reflection: demonstrating an effective engagement with the videos and readings that will inform our ethics discussion this semester.

### Test – 10%

The test for this course will be an open book, open internet test. The test will cover the readings, key concepts, and terminologies from the first part of the semester.

### Industry/Commercial/Institutional Presentation 10%

Each student will conduct research regarding industrial, commercial, or institutional applications of text mining and natural language processing. The specific areas chosen by students should be closely related to their course projects. These will be informal presentations conducted during class times, allowing students an opportunity to learn collaboratively during the semester.

### Grading Scale

* A 94-100
* A- 90-93
* B+ 87-89
* B 83-86
* B- 80-82
* C+ 77-79
* C 73-76
* C- 70-72
* D+ 67-69
* D 63-66
* D- 60-62
* F Below 60

### Tools and Technologies Required for Course

* R programming language
* GitHub
* Atom text editor or R Studio

### Texts

* *Where Wizards Stay Up Late* (used copies widely available)
* *Text Mining with R* (free, online)
* Articles and book excerpts made available in Canvas

## Course Guidelines

### Attendance

Attendance and class participation are mandatory. Each student can miss one class, and each additional class missed will result in a 5% reduction in the final grade for the course.

### Email

The instructor will respond to all emails during a regular “work” schedule from 8am to 4pm, Monday-Friday. This means that any emails received after 4pm will be responded to on the following day, and any emails received after 4pm on Friday will be responded to on the following Monday.

### Assignment Submission

All assignments are to be submitted through Canvas or GitHub. Emailed assignments will not be accepted.

### Late/Missed Assignments

Late assignments will not be accepted. If for some reason you are unable to complete an assignment, then you should negotiate an extension ahead of time. This means emailing the instructor at least one day prior to the due date and asking for help or permission. Failure to do so will result in a zero if an assignment is turned in late.

## Schedule

All courses for the Cultural Analytics concentration maintain a 12 week schedule, and this allows the final 3 weeks (weeks 13-15) to be dedicated to ethics discussions and workshops for the final projects in the course.

### [Week 1–8/19] Course Introduction

* [Lecture] “What is Cultural Analytics?”
* [Readings for 8/26] NLP Overview

### [Week 2–8/26] Natural Language Processing

* [Lecture] “Key Terms and Concepts in NLP”
* [Readings for 9/2] From “Prologue” to chapter 5 of *Where Wizards Stay Up Late*
* [Workshop] Install R, Tidyverse, Rvest, Atom/R Studio

### [Week 3–9/2] Text as Data: Choosing/Creating a Corpus for Analysis

* [Lecture] “Text as Data: Types, Terminologies, and Choosing a Corpus”
* [Assignment for 9/9] Rough draft of Project Proposal
* [Readings for 9/9] From chapter 6 to “Epilogue” of *Where Wizards Stay Up Late*
* [Workshop] Rvest tutorial

### [Week 4–9/9] Wrangling, Scrubbing, and Transforming Text Data

* [Lecture] “Wrangling, Scrubbing, and Transforming Text Data”
* [Workshop] R tutorials

### [Week 5–9/16] No Class

* [Test] Covering readings, key concepts, and terminologies from first 3 weeks
* [Optional] Individual conferences

### [Week 6–9/23] Word Frequency

* [Lecture] “Word Frequency and Zipf’s Law”
* [Workshop] Tutorials TBA
* [Industry/Commercial/Institutional Presentation]

### [Week 7–9/30] N-Grams and Collocations

* [Lecture] “N-Grams and Collocations”
* [Workshop] Tutorials TBA
* [Industry/Commercial/Institutional Presentation]

### [Week 8–10/7] Sentiment Analysis

* [Lecture] “Sentiment Analysis”
* [Workshop] Tutorials TBA
* [Industry/Commercial/Institutional Presentation]

### [Week 9–10/14] Topic Models

* [Lecture] “Topic Models”
* [Workshop] Tutorials TBA
* [Industry/Commercial/Institutional Presentation]

### [Week 10–10/21] Word Embeddings

* [Workshop] “Word Embeddings”
* [Workshop] Tutorials TBA
* [Industry/Commercial/Institutional Presentation]

### [Week 11–10/28] Project Proposals

* [Assignment] Project Proposal Due
* [Workshop] Show continued progress on project

### [Weeks 12–11/4] Ethics

* [Workshop] Show continued progress on project
* [Readings/Videos] TBA

### [Week 13–11/11] Ethics

* [Workshop] Show continued progress on project
* [Readings/Videos] TBA

### [Week 14–11/18] Final Presentation Preparation

* [Assignment] Ethics Response
* [Workshop] Show continued progress on project

### [Week 15–12/2]

* [Assignment] Final Presentation