Statistics 7240/4240 **Fall, 2022**

Text Analytics

Instructor

Sharath Bennur

Teaching Assistant

Office hours

Overview

This course is an introduction to the analysis of unstructured data, focusing on statistical models for text with specific emphasis on techniques having evident business applications.

Statistical methods for the analysis of text have come of age. Techniques that allow you to mine text for underlying sentiments, scan for discriminatory language, or create predictive models are commonly available in modern software. These methods do more than simply count, though counting through millions of words is impressive in itself. Beyond counting, algorithms developed in natural language processing (NLP) provide rich grammatical and syntactic analysis, such as identifying parts of speech, sentence parsing (*e.g.*, subject and predicate), and named entity recognition (people and places). Modern software tools allow the routine use of these methods by non-specialists – at least those who have taken this course!

Other information typically accompanies textual data. Modern data streams routinely combine text with numerical data commonly that can be used, for instance, in predictive models such as those related to regression analysis. For example, real estate listings typically show the listed price of a property and other features such as the number or square feet, type of heating, and a count of rooms. Most listings also come with a written description, often composed in an idiosyncratic vernacular. Advances in text analytics allow us to convert this accompanying text into numerical features suitable for other statistical models. Unsupervised techniques are available to create features directly from text, requiring minimal user input. Because these constructions are unsupervised, the resulting features play the role of familiar explanatory variables. Techniques for building these variables range from naïve to subtle. One can simply use raw counts of words, form principal components from these counts, or build features from counts of grammatical attributes.

Several running examples that span the course will illustrate the surprising success of text analytics with an emphasis on real world applications in various industries such as finance, healthcare and ecommerce.

Prerequisites

Students should be familiar with regression models at the level of Stat 613 or Stat 102, and the Python language at the level of Stat 477 or Stat 777. Familiarity with the Jupyter notebook development environment is presumed, as well as common Python packages such as pandas, NLTK and SpaCy. Those with more knowledge of Statistics, such as from Stat 722/422, or computing skills will benefit. The predominant software used in the course is Jupyter notebooks that use a Python interpreter. Familiarity with basic probability models is helpful but not presumed.

Grading

Grades for this course are determined by performance on weekly homework assignments and a final project.

Weekly assignments	75%	(equally weighted)
Project	25%	

Assignments are due weekly; the course outline below indicates due dates. Assignments are due prior to the lecture on Tuesday of the week following the lectures that cover the relevant material. That means you have the rest of the week and weekend to work on the assignment. The deliverables will be in the form of Jupyter notebooks. A substantial portion of the class on Tuesday will discuss and build upon the assignment that was just submitted. Assignments that have not been submitted by the time of the lecture receive **no** credit.

Note: *Grading questions must be resolved within one week of posting grades*.

Collaboration

Assignments are to be completed individually. I expect you will discuss with each other what's going on in the course and help each other understand the concepts. Each assignment, however, is your own responsibility. Google can help you, but not your classmates. You won't learn the material unless you do these assignments yourself.

Classroom Expectations

There is no formal participation component to the final grade but questions are strongly encouraged. Phones, laptops and other electronic devices are not to be used in class except for Python related activities.

Materials

Lecture notes

Distributed via Canvas (in the Files section). Notes will be in the form of iPython-notebooks distinguished by the suffix ".ipynb" in the file name. These files combine text, Python commands, and output from the Python commands. These form an outline for each class.

Software

Anaconda. Download the most recent, appropriate version (*i.e.*, Windows,

Mac or Linux) of this free software from the website
https://www.anaconda.com/products/individual

Textbooks

No required textbooks for the class, though the following textbooks could be helpful:

VanderPlas, J. (2017). *Python Data Science Handbook*. O'Reilly.

McKinney, W (2017) *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython 2nd Edition.* O'Reilly

Planned schedule of lectures and assignments

Assignments are due at 12 midnight of the day **prior** to the date of the indicated lecture. The due date for the course project will be scheduled near the date of the regularly scheduled final exam. Relevant readings (aside from obvious textbook chapters) are noted.

Lec.	Topic	Assign
1	Overview and introduction to natural language processing (NLP)	
2	Converting text from files into statistical data	
3	Counting, types and tokens, n-grams	Assign 1
4	Tagging and parsing text	
5	Sentiment analysis, lexicons, regular expressions	Assign 2
6	Text analytics for E-commerce	
7	Classifiers: naïve Bayes, logistic models; Latent semantic analysis	Assign 3
8	Text analytics for Finance	
9	Topic models, latent Dirichlet analysis, Word embeddings	Assign 4
10	Text Analytics in Social media	
11	NLP applications - chatbots, voice assistants, translation, analytics	Assign 5
12	Text Analytics in Healthcare	
13	State of the art language models, deep learning and commoditization	
14	Project discussion	

Annotated Bibliography

There are many great resources online for text analytics, natural language processing and natural language understanding. These resources range from the highly computational and theoretical to the application oriented, making it difficult to recommend a single source. I would suggest refining your search for the appropriate resources based on whether the theoretical underpinning, or practical applications, or computational techniques are of interest.