USC Marshall School of Business

DSO 599 – Text Analytics & Natural Language Processing

Syllabus – Spring 2019 (1.5 units, 2nd Half of the Semester)

Instructors: Yu Chen: ychen220@usc.edu (860-922-5954)
Office Hours: Mondays 7pm PST – 9pm PST (remote), Saturdays 9am – 12pm, by appointment
Weekly Hours: 3 hours per week (1.5 unit class)

Course Description

This course will provide students with a thorough introduction and overview of the core concepts and tools needed to acquire, analyze, visualize, and perform natural language processing (NLP) on text data. Students will learn the statistical methodology and develop computer code to detect and visualize patterns in text, extract useful knowledge, and make key business decisions.

There are many courses, both within formal higher education programs and also on distance and online learning platforms, that offer extremely high-quality technical training on natural language processing and basic text analysis. However, as we have observed within the industry, there is often a divide between the teams generating the insights and those who are making the final management decisions. This course serves to help students bridge the gap between management and business analytics- each week contains self-contained business use case modules that will introduce students to the full insight pipeline- from data text mining, data preprocessing, machine learning modelling, visualization, product/marketing strategy, and storytelling.

Learning Objectives

Upon successful completion of this course, students will be able to:

1. Describe how NLP is used to solve business problems
2. Work and write code with common NLP Python libraries: **scikit-learn, gensim, nltk** to solve business related problems
3. Understand and develop word embeddings using several different approaches
4. Classify text using several different approaches (sentiment analysis, intent etc.)
5. Pre-process and apply feature selection with text data
6. Extract themes from documents using several different approaches
7. Make business decisions based on NLP output

Course Materials

The course will utilize *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit* by Steven Bird, Ewan Klein, and Edward Loper and *Introduction to Algorithmic Marketing: Artificial Intelligence for Marketing Operations* by Ilya Katsov. The NLP textbook content is accessible at *http://www.nltk.org/book/*, and the Algorithmic Marketing textbook content is freely available online.

Tableau will be used for visualization of text data. Tableau offers a free 1-year subscription for currently enrolled students.

Students will be provided with **AWS credits** as part of cloud resources for NLP projects and exercises. The course will also make use of Databricks Community Edition, which is a freely available distributed computing platform based upon the open source Apache Spark project.

Prerequisites and/or Recommended Preparation:

Familiarity with one (or more) programing language(s) is strongly recommended. This can be achieved through a "Python for Business Analytics" (0.5 unit) course that is offered through USC during the first half of the Spring quarter **or** completing a "Python: Data Analysis" (2.5 hour) course offered through Lynda.com. The "Python: Data Analysis" course, taught by Michele Vallisneri, will cover two Python packages: Pandas and Numpy and is available for free for USC students.

Additional office hours will be available for students who require further support in accessing the technical content (programming and machine learning concepts) of this course. We want to emphasize that this course is bridging business with technical programming- the grading rubrics will emphasize holistic understanding of text analytics applications, versus how well a student can program a for loop in Python.

Familiarity with linear algebra is recommended, but not required. Many of the algorithms we will implement to analyze our business use cases will require us working with matrices.

Course Notes:

All course materials and announcements are posted on the Blackboard site. It is **your responsibility** to check that site and your email **regularly** to ensure class preparation.

GRADING DETAIL

Your final course grade, which will be curved, will be assessed as follows:

ASSIGNMENTS	Points	% of Grade
Midterm exam	20	20%
Homework assignments (5 assignments)	30	30%
Class participation	15	15%
Final Team project	35	35%
TOTAL	100	100%

Class Participation:
Each week during class, there will be a variety of exercises and mini-assessments. Students will be required to submit their work onto Blackboard within the provided time-frame in order to earn credit for that portion of class.

Please note that class participation submissions are graded for accuracy.

Homework: Each week's homework will consist of a small problem set of exercises that will serve to reinforce and extend that week's learnings. **Certain problems may involve self-contained programming/coding exercises. This code must be individually produced, as homework assignments are individual exercises.**

Each homework will represent a small, self-contained business use case and dataset. At the end of each homework, students are expected to present the final business use case recommendations for management, delivered in the form of an executive summary.

Homework is graded for accuracy.

Team Project: The Final Team Project will constitute 35% of the grade and must be completed in groups of no more than 5 students. No time during class will be devoted specifically to the final project, so students must coordinate amongst themselves to find times to meet. The project should require 15-20 hours of work (5-7 hours per student) if teams collaborate efficiently. The final deliverable will be in the form of a client-facing deck presentation (please convert and save as PDF prior to submission), as well as all code utilized and any workbooks for the visualizations.

The team project will be graded as follows:

- **Team Presentation (30%):** Week 8 final presentation
- **Business Recommendation (40%):** did your team's solutions clearly address a business problem?
- **Technical Implementation (30%):** was your team's technical solution clear, accurate, and scalable?

Exams: During the first half of Week 5, students will take a Midcourse Exam that will last no longer than 1.5 hour and will constitute 20% of the grade. The assessment will cover the first four weeks of class and will involve case study questions and some programming exercises in Python. There is no formal final exam for this course.

Assignment Submission Policy:
Assignments must be turned in on the due date/time. Any assignment turned in late will receive a 10% grade deduction per day.

Evaluation of Your Work:
You may regard each of your submissions as an "exam" in which you apply what you've learned according to the assignment. I will do my best to make my expectations for the various assignments clear and to evaluate them as fairly and objectively as I can. If you feel that an error has occurred in the grading of any assignment, you may, within one week of the date the assignment is returned to you, write me a memo in which you request that I re-evaluate the assignment. Attach the original assignment to the memo and explain fully and carefully why you think the assignment should be re-graded. Be aware that the re-evaluation process can result in three types of grade adjustments: positive, none, or negative.

ADDITIONAL INFORMATION

Add/Drop Process

Most Marshall classes are open enrollment (R-clearance) through the Add deadline. If there is an open seat, students can add the class using Web Registration. If the class is full, students will need to continue checking the *Schedule of Classes* (classes.usc.edu) to see if a space becomes available. Students who do not attend the first two class sessions (for classes that meet twice per week) or the first class meeting (for classes that meet once per week) may be dropped from the course if they do not notify the instructor prior to their absence.

If a graduate class is full students should sign up on the wait list.
www.marshall.usc.edu/registrationpolicies

Retention of Graded Coursework

Exam and all other graded work which affected the course grade will be retained for one year after the end of the course ***if*** the graded work has not been returned to the student. If I returned a graded paper to you, it is your responsibility to file it.

USC Statements on Academic Conduct and Support Systems

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Part B, Section 11, "Behavior Violating University Standards" https://policy.usc.edu/scampus-part-b/. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

Support Systems

Student Counseling Services (SCS) - (213) 740-7711 – 24/7 on call

Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention. https://engemannshc.usc.edu/counseling/

National Suicide Prevention Lifeline - 1-800-273-8255

Provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week. http://www.suicidepreventionlifeline.org

Relationship & Sexual Violence Prevention Services (RSVP) - (213) 740-4900 - 24/7 on call

Free and confidential therapy services, workshops, and training for situations related to gender-based harm. https://engemannshc.usc.edu/rsvp/

Sexual Assault Resource Center

For more information about how to get help or help a survivor, rights, reporting options, and additional resources, visit the website: http://sarc.usc.edu/

Office of Equity and Diversity (OED)/Title IX compliance – (213) 740-5086

Works with faculty, staff, visitors, applicants, and students around issues of protected class.
https://equity.usc.edu/

Bias Assessment Response and Support
Incidents of bias, hate crimes and microaggressions need to be reported allowing for appropriate investigation and response. https://studentaffairs.usc.edu/bias-assessment-response-support/

Student Support & Advocacy – (213) 821-4710
Assists students and families in resolving complex issues adversely affecting their success as a student EX: personal, financial, and academic. https://studentaffairs.usc.edu/ssa/

Diversity at USC – https://diversity.usc.edu/
Tabs for Events, Programs and Training, Task Force (including representatives for each school), Chronology, Participate, Resources for Students

USC Emergency Information
Provides safety and other updates, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible. emergency.usc.edu

USC Department of Public Safety – UPC: (213) 740-4321 – HSC: (323) 442-1000 – 24-hour emergency or to report a crime.
Provides overall safety to USC community. dps.usc.edu

Students with Disabilities

USC is committed to making reasonable accommodations to assist individuals with disabilities in reaching their academic potential. If you have a disability which may impact your performance, attendance, or grades in this course and require accommodations, you must first register with the Office of Disability Services and Programs (www.usc.edu/disability). DSP provides certification for students with disabilities and helps arrange the relevant accommodations. Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me (or to your TA) as early in the semester as possible. DSP is located in GFS (Grace Ford Salvatori Hall) 120 and is open 8:30 a.m.–5:00 p.m., Monday through Friday. The phone number for DSP is (213) 740-0776. Email: ability@usc.edu.

Emergency Preparedness/Course Continuity

In case of a declared emergency if travel to campus is not feasible, the *USC Emergency Information* web site (http://emergency.usc.edu/) will provide safety and other information, including electronic means by which instructors will conduct class using a combination of USC's Blackboard learning management system (blackboard.usc.edu), teleconferencing, and other technologies.

COURSE CALENDAR (tentative)

Week	Date	Topics	Deliverables and Due Dates
1	3/5	**NLP Overview** • The three segments of NLP (speech recognition, natural language understanding, and natural language generation) • Probability distributions, Naïve Bayes • Working with text in Pandas dataframes, streams, and bytes • **Dataset exercise:** Amazon Toy Product Reviews	**Reading**: None **Homework (Introduction to Text Processing in Python)**: • Probability distributions and Naïve Bayes • Basic Python operations
2	3/19	**Data Preprocessing & Linear Algebra Review** • Similarity / distance measures • Collocations and n-grams • Tokenization Lemmatization/Stemming • Regular expressions • Word vectors: TF-IDF, One-Hot, Count • **Dataset exercise:** BBC Sports news articles • **Dataset exercise:** Spam / Ham SMS • **Dataset exercise:** Databricks / AWS Machine Learning for processing Amazon product reviews • Visualization of text topics with Tableau	**Reading:** • *Algorithmic Marketing* pages 179 - 184, 193 - 201 (Search) • "What is Natural Language Processing? The Business Use Case Explained" (CIO.com) **Homework (Data Preprocessing)**: • Exercises from NLPP • TF-IDF summarization of Reddit posts
3	3/26	**Dimensionality Reduction:** • PCA, t-SNE, t-SVD • Visualizing high-dimensional data **Feature Selection/ Text Classification for Sentiment Analysis** • Train/test split, K-Fold cross validation • Hyperparameter tuning • Logistic regression, Random Forest **Model Evaluation** • Accuracy • Precision, Recall, Confusion Matrix • AUROC, F1 Scores	**Readings:** • *Algorithmic Marketing* pages 218- 222, 224-231 **Homework**: Analysis of social media data **Project Checkpoint #1:** Team Assignments and initial data exploration
4	4/2	**Word Embeddings** • Word2Vec • FastText **Emoji mapping and internationalization**	**Reading**: • *Algorithmic Marketing* pages 222 - 249 **Homework (Due** : Choose one of the methods and create a word

		NLP Recommendation Systems: Product Search • **Databricks Dataset**	embedding for one of the datasets provided here. Prove that semantic meaning and relationships are captured within the embedded vectors.
5	4/9	**First Half of Class:** • Midcourse Quiz (70 minutes) • Midcourse Quiz review of answers and deep-dive **Second Half of Class: Parts of Speech Tagging, Named Entity Recognition** • Hidden Markov Models **Dataset Exercise:** labelling NER and POS on BBC news reports for text summarization	**Homework (Practice Classifying Text and Generating Business Recommendations):** • Exercises based off of *Algorithmic Marketing* pages 303 - 342 **Project Checkpoint #2:** Initial research summary and hypothesis
6	4/16	**Deep Learning for NLP:** • Feedforward, RNN, CNNs for NLP tasks • Sentiment analysis for classification • Sequence to sequence models • **Dataset exercise:** Databricks / AWS Machine Learning for processing Amazon product reviews	**Reading:** • Chapter 1&2: LSA/LDA • Part 1&2: LDA/HDP **Homework:** Databricks notebook for AWS S3 hosted product reviews
7	4/23	**Visualization, Storytelling, and Beyond:** • Tableau dashboards and worksheets **Natural Language Generation:** capturing business logic within automated insights **NLP in the Big Data era:** - Cloud computing frameworks - Spark notebooks and use cases	**Project Checkpoint #4:** All deliverables
8	4/30	**AAC Project Client Presentations (first half of class)** - Team presentations and Q&A **Building a Natural Language Product** - **Pricing** - **Product versus Service** - **Business context and landscape for machine learning projects** **Data Exercise:** hosting a recommendation system as a product feature with Python and Spark	**Reading:** Algorithmic Marketing Chapter (Pricing, Demand and Supply)

Appendix I. MARSHALL GRADUATE PROGRAMS LEARNING GOALS
How DSO 599 Contributes to Marshall Graduate Program Learning Goals

Marshall Graduate Program Learning Goals	DSO 599 Objectives that support this goal	Assessment Method*
Learning Goal #1: Develop Personal Strengths. **Our graduates will develop a global and entrepreneurial mindset, lead with integrity, purpose and ethical perspective, and draw value from diversity and inclusion.**		
1.1 Possess personal integrity and a commitment to an organization's purpose and core values.		
1.2 Expand awareness with a global and entrepreneurial mindset, drawing value from diversity and inclusion.		
1.3 Exhibit awareness of ethical dimensions and professional standards in decision making.		
Learning Goal #2: Gain Knowledge and Skills. **Our graduates will develop a deep understanding of the key functions of business enterprises and will be able to identify and take advantage of opportunities in a complex, uncertain and dynamic business environment using critical and analytical thinking skills.**		
2.1 Gain knowledge of the key functions of business enterprises.	1	Homework, and Project

2.2 Acquire advanced skills to understand and analyze significant business opportunities, which can be complex, uncertain and dynamic.	1-7	Homework, and Project
2.3 Use critical and analytical thinking to identify viable options that can create short-term and long-term value for organizations and their stakeholders.	1-7	Homework, and Project
Learning Goal #3: Motivate and Build High Performing Teams. **Our graduates will achieve results by fostering collaboration, communication and adaptability on individual, team, and organization levels.**		
3.1 Motivate and work with colleagues, partners, and other stakeholders to achieve organizational purposes.	1-7	Homework, and Project
3.2 Help build and sustain high-performing teams by infusing teams with a variety of perspectives, talents, and skills and aligning individual success with team success and with overall organizational success.	1-7	Homework, and Project
3.3 Foster collaboration, communication and adaptability in helping organizations excel in a changing business landscape.	1-7	Homework, and Project