USC Marshall
School of Business

DSO 560 – Text Analytics & Natural Language Processing

Syllabus – Spring 2022 (2nd half of the semester)– Tu, 6:30-9:30pm, 1.5 Units
Instructor: Yu Chen
Teaching Assistants:
Siyuan Ni (siyuanni@usc.edu)
Mengqi Tan (mengqi.tan.2021@marshall.usc.edu)
Office: N/A
Email: ychen220@marshall.usc.edu
Office Hours: By appointment and Saturday 9am – 12pm (remote)

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 utilize core Python data science and machine learning packages 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. Write functional code using scikit-learn, gensim, nltk, pandas, and spacy to solve business-related queries and process data
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 the following texts and resources:

- *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit* by Steven Bird, Ewan Klein, and Edward Loper
- *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.

Software

We will use Keras (https://keras.rstudio.com) and its Python/Tensorflow interface to implement code examples in this course. Keras is a high-level neural networks application-programing interface (API) developed with a focus on enabling fast experimentation. Keras has the following key features:

- Allows the same code to run on CPU or on GPU, seamlessly.
- User-friendly API which makes it easy to quickly prototype deep learning models.
- Built-in support for convolutional networks (for computer vision), recurrent networks (for sequence processing), and any combination of both.
- Supports arbitrary network architectures: multi-input or multi-output models, layer sharing, model sharing, etc. This means that Keras is appropriate for building essentially any deep learning model, from a memory network to a neural Turing machine.
- Is capable of running on top of multiple back-ends including TensorFlow, CNTK, and Theano.

Prerequisites and/or Recommended Preparation:

Prerequisite: DSO 545, Statistical Computing and Data Visualization.
Corequisite: DSO 530, Applied Modern Statistical Learning Methods.

Students are also expected to be familiar with Python and linear algebra. Many of the algorithms we will implement to analyze our business use cases will require us working with matrices.

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.

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	% of Grade
Final exam	20%
Homework assignments	35%
Final Group project	30%
Classwork	15%
TOTAL	100%

- **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, and most will be completed using Python Jupyter Notebooks. 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.
- **Final Group Project:** The **Final Group Project** will constitute 30% 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.
- **Participation:** Participation will be assessed each week via short, in-class assignments. These assignments will be submitted near the end of the lecture and will be based upon the classwork and lecture concepts introduced that particular session.

- **Final Group Project** will be graded as follows:
    - Business Recommendation (50%): did your team's solutions clearly address a business problem?
    - Technical Implementation (50%): was your team's technical solution clear, accurate, and scalable?

  Each group member will also be expected to complete a 360 peer evaluation where each team member's contributions to the final deliverables are outlined and an assessment of percentage contribution (which must sum to 100%) is provided as a subjective point of view. This evaluation will be used to ensure that group members that contribute significantly more or less to the final group output are given grades that reflect their individual contributions.

  Final Exams: Students will take a Final Exam that will last no longer than 2 hours and will constitute 20% of the grade. The assessment will cover the first 6 weeks of class and will involve case study questions and other mathematical concepts from lecture.

Final grades represent how you perform in the class relative to other students. Your grade will not be based on a mandated target, but on your performance. Three items are considered when assigning final grades:

1. Your average weighted score as a percentage of the available points for all assignments (the points you receive divided by the number of points possible).
2. The overall average percentage score within the class.
3. Your ranking among all students in the class.
4. Observable effort to improve, ask questions, or come to office hours when struggling / stuck with a homework assignment or concept

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:
Students are expected to make themselves aware of and abide by the University community's standards of behavior as articulated in the Student Conduct Code. 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" 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 at http://policy.usc.edu/scientific-misconduct.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
studenthealth.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention and Services (RSVP) - (213) 740-9355(WELL), press "0" after hours – 24/7 on call
studenthealth.usc.edu/sexual-assault
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED)- (213) 740-5086 | Title IX – (213) 821-8298
equity.usc.edu, titleix.usc.edu
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants. The university prohibits discrimination or harassment based on the following *protected characteristics*: race, color, national origin, ancestry, religion, sex, gender, gender identity, gender expression, sexual orientation, age, physical disability, medical condition, mental disability, marital status, pregnancy, veteran status, genetic information, and any other characteristic which may be specified in applicable laws and governmental regulations. The university also prohibits sexual assault, non-consensual sexual contact, sexual misconduct, intimate partner violence, stalking, malicious dissuasion, retaliation, and violation of interim measures.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
usc-advocate.symplicity.com/care_report
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Support and Advocacy - (213) 821-4710
uscsa.usc.edu
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101

diversity.usc.edu
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
dps.usc.edu
Non-emergency assistance or information.

COURSE CALENDAR (tentative)

Week	Date	Topics	Deliverables and Due Dates
1	3/8	**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 ● **Dataset exercise:** Tale of Two Cities Word Count ● **Dataset exercise:** Regex practice for parsing text ● Regular expressions	**Homework (Introduction to Text Processing in Python) due when class starts 3/22**: ● Basic Python operations ● Unicode encoding and decoding exercise ● Regular expression exercises
2	3/22	**Data Preprocessing & Linear Algebra Review** ● Similarity / distance measures ● Collocations and n-grams ● Tokenization Lemmatization/Stemming ● Word vectors: TF-IDF, One-Hot, Count ● **Dataset exercise:** BBC Sports news articles ● **Dataset exercise:** Spam / Ham SMS **Dimensionality Reduction:** ● PCA, t-SNE, t-SVD ● Visualizing high-dimensional data	**Homework (Data Preprocessing) due 3/29**: ● TF-IDF summarization of Yelp McDonalds restaurant reviews ● Search

3	3/29	**Feature Selection/ Text Classification for Sentiment Analysis** ● Train/test split, K-Fold cross validation ● Hyperparameter tuning **Model Evaluation** ● Accuracy ● Precision, Recall, Confusion Matrix ● AUROC, F1 Scores **Word Embeddings** ● Word2Vec, GloVe ● FastText **Emoji mapping and internationalization** **NLP Recommendation Systems: Product Search**	**Homework (due 4/6)**: Analysis of social media data **Reading**: ● *Algorithmic Marketing* pages 222 - 249 **Reading:** ● Chapter 1&2: LSA/LDA ● Part 1&2: LDA/HDP
4	4/5	**Deep Learning for NLP:** ● Feedforward, RNN, CNNs for NLP tasks ● Hyperparameter tuning and search for NLP tasks / transfer learning **Dataset exercise:** Keras model for processing Amazon product reviews to classify sentiment	**Homework (due 4/12):** Word embeddings, fuzzy matching, and advanced regex exercises.
5	4/12	**Deep Learning for NLP:** ● LSTM models / encoder / decoder architectures for machine translation ● Parts of Speech Tagging, ● Named Entity Recognition ● Hidden Markov Models **Dataset Exercise:** labelling NER and POS on BBC news reports for text summarization	**Homework (due 4/19)** **Project Checkpoint #1:** Team Assignments and initial data exploration Initial research summary and hypothesis
6	4/19	**Deep Learning for NLP (Part II):** ● LSTM models / encoder / decoder architectures for machine translation ● GRUs, RNNs	**Homework (due 4/26)**
7	4/26	**Deep Learning for NLP:** ● Self-Attention, Transformers, & BERT for NLP tasks ● GPT-2 / GPT-3 **Dataset exercise:** Keras deep learning models to implement self-attention / LSTM for English to French translation.	**Study for Final Exam**

8	5/10	**Final Exam (90 minutes)**	**Final Project is due by 5/11, 11:59pm PST**

Appendix I. MARSHALL GRADUATE PROGRAMS LEARNING GOALS

How DSO 560 Contributes to Marshall Graduate Program Learning Goals

Marshall Graduate Program Learning Goals	**DSO 560 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.	5	Homework,

		Presentation, and Project
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.		
2.2 Acquire advanced skills to understand and analyze significant business opportunities, which can be complex, uncertain and dynamic.	2, 3, 4	Homework, Presentation, 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.	1-4	Homework, Presentation, 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.	3-6	Homework, Presentation, 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.	3-6	Homework, Presentation, and Project