DSO 560 – Text Analytics & Natural Language Processing

Department of Data Science and Operations, USC Marshall, School of Business

This syllabus is the contract between you and me. Please read it carefully.

Professor Information
Professor: Dr. Kanad Basu Email: kbasu@marshall.usc.edu

Office Hours
Zoom Meeting ID: 673 593 7364 & Passcode: 558271
Office Hours: By appointment and Saturday 9:30 am – 12:30 pm (remote)

Class Meetings
Lecture: Wed, 6:30- 9:30pm, 1.5 Units Location: JKP 102

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 modeling, visualization, product/marketing strategy, data ethics 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, semantic search etc.)
5. Pre-process and apply feature selection with text data
6. Make business decisions based on NLP output
7. Data Ethics & Privacy concerns reading text analytics

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.
- Mastering NLP from Foundations to LLMs: Apply advanced rule-based techniques to LLMs and solve real-world business problems using Python by Lior Gazit & Meysam Ghaffari

The NLP textbook content is accessible at http://www.nltk.org/book/, and the Algorithmic Marketing textbook content is freely available online.

Purchase of the textbooks is optional. All class material will be posted on LMS.

Software
It is recommended that students use Google's Colab platform to run Python code.
https://colab.research.google.com

Google's Colab platform is free for Google's customers. If you have a 'gmail' account, you can use Colab at no cost.

Anaconda's Jupyter Notebook. https://www.anaconda.com/code-in-the-cloud
The following Python libraries will be used.

- Python/Pandas
- Python/TextBlob
- Python/Scikit-Learn
- Keras/TensorFlow

We will use 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 Brightspacesite. 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:

Assignment Type	Percentage Contribution
Data Camp Assignment	30%
Homework Assignment	15%
Final Group Project & Proposal	40%
Participation/Forum Discussion	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 40% 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/Forum Discussion:** Active participation (5%) in class discussions is essential for developing a deeper understanding of the course material. This portion of your grade will be determined by the quality of your contributions during group discussions, your engagement with peers, and your demonstrated preparation for each session. By consistently participating and sharing insights, you will enrich both your own learning experience and that of your classmates.

  Several Discussion Board assignments (10%) posted to Brightspace will be used to engage students in social learning. These assignments provide opportunities for students to post thoughtful reflections on assigned topics or questions, as well as to consider and respond to classmates' posts on these topics or questions. Discussion Board as-

signments evidencing thoughtful reflections, fulfilling all of the stated requirements, and submitted on time will receive the full five points. A Discussion Board assignment fulfilling most, but not all, of the stated requirements and/or posted after the deadline but before the start of the next class session will receive a reduction of two points. A Discussion Board assignment not fulfilling most of the stated requirements and/or posted after the start of the next class session will receive no points.

- **Final Group Project** will be graded as follows:
  - Team Presentation (30%): Week 8 final presentation/deliverables
  - 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?

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 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:

- 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).
- The overall average percentage score within the class.
- Your ranking among all students in the class.
- 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.

The Use of AI

I expect you to use AI (e.g., ChatGPT and image generation tools) in this class. Learning to use AI is an emerging skill, and I welcome the opportunity to meet with you to provide guidance with these tools during office hours or after class. Keep in mind the following

- AI tools are permitted to help you brainstorm topics or revise work you have already

written.

- If you provide minimum-effort prompts, you will get low-quality results. You will need to refine your prompts to get good outcomes. This will take work.
- Proceed with caution when using AI tools and do not assume the information provided is accurate or trustworthy If it gives you a number or fact, assume it is incorrect unless you either know the correct answer or can verify its accuracy with another source. You will be responsible for any errors or omissions provided by the tool. It works best for topics you understand.
- AI is a tool, but one that you need to acknowledge using. Please include a paragraph at the end of any assignment that uses AI explaining how (and why) you used AI and indicate/specify the prompts you used to obtain the results what prompts you used to get the results. Failure to do so is a violation of academic integrity policies.
- Be thoughtful about when AI is useful. Consider its appropriateness for each assignment or circumstance. The use of AI tools requires attribution. You are expected to clearly attribute any material generated by the tool used.]

OPEN EXPRESSION AND RESPECT FOR ALL

An important goal of the educational experience at USC Marshall is to be exposed to and discuss diverse, thought-provoking, and sometimes controversial ideas that challenge one's beliefs. In this course we will support the values articulated in the USC Marshall Open Expression Statement.

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.

Academic Integrity

The University of Southern California is foremost a learning community committed to fostering successful scholars and researchers dedicated to the pursuit of knowledge and the transmission of ideas. Academic misconduct is in contrast to the university's mission to educate students through a broad array of first-rank academic, professional, and extracurricular programs and includes any act of dishonesty in the submission of academic work (either in draft or final form).

This course will follow the expectations for academic integrity as stated in the USC Student Handbook. All students are expected to submit assignments that are original work and prepared specifically for the course/section in this academic term. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s). Students suspected of engaging in academic misconduct will be reported to the Office of Academic Integrity.

Other violations of academic misconduct include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

Academic dishonesty has a far-reaching impact and is considered a serious offense against the university. Violations will result in a grade penalty, such as a failing grade on the assignment or in the course, and disciplinary action from the university itself, such as suspension or even expulsion.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment or what information requires citation and/or attribution.

Statement on Academic and Support Systems

Students and Disability Accommodations:
USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Student Financial Aid and Satisfactory Academic Progress:
To be eligible for certain kinds of financial aid, students are required to maintain Satisfactory Academic Progress (SAP) toward their degree objectives. Visit the Financial Aid Office webpage for undergraduate- and graduate-level SAP eligibility requirements and the appeals process.

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

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Life-

line) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline consists of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services (though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
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.

Reporting Incidents of Bias or Harassment - (213) 740-2500
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
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
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) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.

Note

It is possible that, during the semester, there may need to be updates to how the course is run. In such a situation this document will be updated. I will keep you apprised of any changes to the syllabus, but it your responsibility to make sure that read and understand such updates and/or changes.

DSO 560: Text Analytics & NLP (Tentative Schedule)		
Date	**Topics**	**Deliverables and Due Dates**
	Week 1	
01/15/25	**Overview on NLP** • Course Expectation & Explanation • Introduction to Text Analytics • Business Use Cases of NLP in HR, Fraud Detection, Market Research • Text Embeddings, Classification & Semantic Search • **Guest Lecture**: Mr. Biswajit Pal, Director at Kenvue	**Reading:** Social Media & News Listening, and Digital Outreach **HomeWork:** DataCamp:(Introduction to Natural Language Processing in Python) • Regular expressions & word tokenization
	Week 2	
01/22/25	**Tokenization & Vectorization** • Working with Text: Google's Colab (Python + Pandas + Jupyter Notebook + Text blob) • Regular Expression • Vectorization: Similarity & Distance measure • Collocation & N-gram • Zipf's Law • Lemmatization & Stemming • Word vectors: TF + TF-IDF • **Dataset excercise: Spam / Ham SMS**	**Reading:** Semantic Intelligence Use Case **Discussion:** "Social Listening: The Good, The Bad, and The Costly" **HomeWork:** DataCamp:(Introduction to Natural Language Processing in Python) • Simple topic identification

	Week 3	
01/29/25	**Text Classification** • Probability Distribution: Naïve Bayes • Bayes Model for Predicted Analytics • Bayes Model for NLP – Text Classification • Understanding our dataset • Data Preprocessing • Word Embeddings: Bag of Words(BoW) • Naive Bayes implementation using scikit-learn • Evaluating our model	**Reading:** Algorithmic Marketing pages 179 - 184, 193 - 201 (Search) **Discussion:** "How Semantic Intelligence Transforms Raw Customer Data into Strategic Gold" **HomeWork:** DataCamp:(Introduction to Natural Language Processing in Python) • Named-entity recognition
	Week 4	
02/05/25	**Introduction to Neural Network** • Deep Learning Application: Word Embeddings • Final Project Guideline Discussion • **Guest Lecture**: Ms. Annie Flippo, CDO at Urgently	**Reading:** Algorithmic Marketing pages 218- 222, 224- 231 **Discussion Assignment** **HomeWork:** DataCamp:(Introduction to Natural Language Processing in Python) • Building a "fake news" classifier
	Week 5	
02/12/25	**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	**Project Checkpoint**: Initial research summary and hypothesis

	Week 6	
02/19/25	**Language Models: Transformers:** • **Guest Lecture**: Mr. Lior Gazit, Machine Learning Group Manager at S&P Dow Jones Indices • Large Language Model • Self-Attention Architecture & BERT for NLP tasks • Positional Encoding • GPT-3, ChatGPT	**Discussion Assignment HomeWork:** DataCamp:(Feature Engineering for NLP in Python) • Basic features and readability scores • Text preprocessing, POS tagging and NER
	Week 7	
02/26/25	**Data Privacy & Responsible AI:** • Risks of GenAI • **Guest Lecture**: Dr. Subho Majumdar, Co-Founder & Head of AI at Vijil	**Discussion Assignment HomeWork:** DataCamp:(Feature Engineering for NLP in Python) • N-Gram models

	Week 8	
03/05/25	**Project Presentation** **Course Highlights:** • Mastered fundamental NLP concepts and techniques • Applied text analytics tools to real-world business problems • Developed practical skills in data preprocessing and analysis • Collaborated on innovative group projects • Explored cutting-edge developments in AI and language processing **Next Steps:** • Final projects are due by [date] • Course evaluations will be available through [date] • Final grades will be posted by [date] • Resources and materials will remain accessible through Blackboard until [date] **Future Opportunities:** Consider advanced courses in data science and analytics Apply your NLP skills to internships and research projects Stay connected with classmates for future collaborations Keep exploring new developments in this rapidly evolving field	**Final Project is due 3/05** **HomeWork:** DataCamp:(Feature Engineering for NLP in Python) • TF-IDF and similarity scores