USC Marshall
School of Business

DSO-560: Text Analytics and Natural Language Processing
Fall 2024

1.5 Units, The class will meet 3 hours per week
Monday/Wednesday 3:30 PM to 4:50 PM
October 14 – Dec 4, 2024
Classroom: BRI-202

Instructor:	Ash Pahwa, Ph.D.
Office:	Office location BRI-303E
Office Hours:	By appointment Monday/Wednesday 3:30 PM to 4:50 PM
Phone:	(949) 378-1229
Email:	ashpahwa@marshall.usc.edu

COURSE DESCRIPTION

All students enrolled in this course will need to be able to write, edit and run Python notebooks.

Natural Languages have evolved from thousands of years of human existence as they pass from generation to generation. The grammar of any natural language is complex and different from other languages. Moreover, it is evolutionary. This makes Natural Language Processing (NLP) a complex challenge.

The main goal of NLP is to understand the meaning of text. Only when computers understand the real meaning of the text, can they take decisive action which must be the intended action. Sentiment analysis of text is one of the important applications of NLP. The use case of sentiment analysis is for the purpose of analyzing customer feedback and tweets. Translation of text between languages is another significant NLP application.

We already use NLP capabilities in our daily lives. Personal assistant software like Siri (Apple's iPhone), Google Assistant (search engines), Amazon's Alexa, and robots use NLP (text and audio) to understand the user's intent and provide an accurate response. Spelling and grammar errors flagged by word processing software packages (like MS Word, Google Docs) use NLP to improve the written text. Mobile phones use NLP to understand and transcribe audio messages.

There are currently two different approaches to NLP. The first one is the analysis of words, sentences, and the semantics of text. There are various software packages that can provide these capabilities. These software packages are Natural Language Tool Kit (NLTK), TextBlob, and spaCy.

The other approach to NLP is using the Machine Learning strategy to analyze the text. Neural Network models are used to train a model by feeding it a lot of text data. Google Cloud Platform (GCP) provides a Machine

Learning (ML) API (Application Programming Interface) for the analysis of Natural Languages and provides translation service between languages. IBM Watson provides similar services for NLP.

This course explores both approaches to NLP.

For the first approach, the fundamental mathematical analysis of NLP will be covered. Students will write Python code to access NLTK, TextBlob and spaCy software packages. A major focus will be to analyze the polarity, subjectivity, and sentiments of text using the concepts of n-grams. Next, the concepts of Stemming and Lemmatization will be covered. Finally, entity recognition and similarity detection concepts will be explored.

For the second approach, students will explore the Machine Learning models for NLP and Transformers. The procedure to interface to HuggingFace portal to access Transformers will be covered and use them for business applications. The concepts of Large Language Models (LLM) (like ChatGPT) will be covered and discuss strategies to interface with LLM using Prompt Engineering.

Students will write Python code to access TensorFlow and Keras software packages running on the Google's Colab platform for text analysis.

COURSE OBJECTIVES

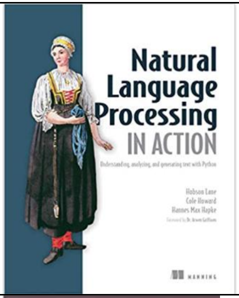
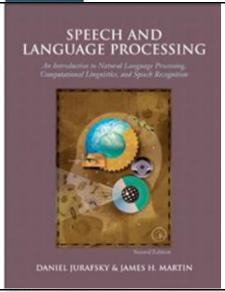
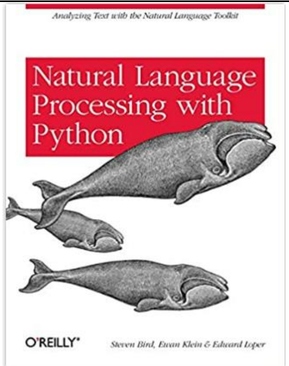
Upon successful completion of this course, students will be able to:

1. Describe how NLP is used to solve business problems.
2. Write functional code using Scikit-Learn, Keras, NLTK, TextBlob, and Pandas to solve business related queries and process data.
3. Describe the concept of Tokenization and Vectorization.
4. Develop word embeddings using several different approaches.
5. Classify text using several different approaches (sentiment analysis, intent etc.)
6. Analyze transformer architecture: Positional Encoding + Attention mechanism.
7. Extract transformers from HuggingFace portals and use them for business applications.
8. Communicate with Large Language Models (LLM) using prompt engineering.

COURSE MATERIALS

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

Textbooks:

1	Natural Language Processing in Action By • Hobson Lane • Cole Howard • Hannes Max Hapke Publisher: Manning	
2	Speech and Language Processing Second Edition By • Daniel Jurafsky • James Matin Publisher: Pearson Prentice Hall	 SPEECH AND LANGUAGE PROCESSING DANIEL JURAFSKY & JAMES H. MARTIN
3	Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit By • Steven Bird • Ewan Klein • Edward Loper Publisher: O’Reilly	 Analyzing Text with the Natural Language Toolkit Natural Language Processing with Python O'REILLY Steven Bird, Ewan Klein & Edward Loper

Course materials: All course materials will be made available via the course site on BrightSpace when possible.

Technology requirements are different for each course. Marshall has site licenses for a variety of software that students can access free of charge. A list of available software is located here. You are responsible for ensuring that you have the necessary computer equipment and reliable internet access. You are invited to explore what lab or loaner options exist. Contact the Marshall HelpDesk (213-740-3000 or HelpDesk@marshall.usc.edu) if you need assistance.

If listing Brightspace (https://brightspace.usc.edu) as a resource, include “If you have any questions or need assistance with the Brightspace Course Pages, please contact the Marshall HelpDesk at 213-740-3000 (option 2) or HelpDesk@marshall.usc.edu.” Alternatively, (213) 740-5555 will get you the USC ITS Help Desk.

Instructor Information

Ash Pahwa, Ph.D., is an educator, author, entrepreneur, and technology visionary with three decades of industry and academic experience. He has founded several successful technology companies during his career, the latest of which is A+ Web Services.

Dr. Pahwa earned his doctorate in Computer Science from the Illinois Institute of Technology in Chicago. He is listed in *Who's Who in the Frontiers of Science and Technology*. He is also a Google Certified Analytics Consultant. His expertise includes search engine optimization, web analytics, web programming, digital image processing, database management, digital video, and data storage technologies.

In Industry, Dr. Pahwa has worked for General Electric, AT&T Bell Laboratories, Xerox Corporation, and Oracle. He founded CD-Gen, Inc. and DV Studio Technologies, LLC., which introduced successful products for CD-Recording (CDR) and MPEG encoding. His book, *CD-Recordable Bible* was published in English, Japanese, and German.

In Academia, Dr. Pahwa teaches courses at California Institute of Technology (Pasadena) and the University of California system. Since 2008, he taught many courses at UC Irvine, UCLA, and UC San Diego.

GRADING

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

ASSIGNMENTS	**% of Grade**
Final exam	25%
Homework assignments	60%
Participation	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.

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 in that session.

Final Exams: Students will take a Final Exam that will last no longer than 2 hours and will constitute 25% of the grade.

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.

Late Submission
All students are expected to submit homework assignments on or before the due date. I may post the answers to the homework questions after the due date.

Since the answers to the homework questions will be posted shortly after the due date, students submitting their homework assignments after the due date will lose points for that homework assignment, based on the following table:

Homework Submission Date	Maximum Points Earned
On Time	100%
1 Week Late	50%
2 Weeks Late	30%
3+ Weeks Late	0%

If you do not submit homework, you will get 0/10.

COURSE OUTLINE

DSO 560	**Text Analytics + Natural Language Processing**
L#	
1.0	**Introduction to NLP + Tools**
1.1	Overview of the Course
1.2	Introduction to NLP
1.3	NLP + Software Libraries
1.4	Colab & Codelabs: Platforms for Writing Python Code
1.5	Regular Expression-1
1.6	Regular Expression-2
2.0	**Tokenization + Text Analysis**
2.1	Tokenization
3.0	**Vectorization**
3.1	Vectorization + Cosine Similarity + BOW_TF
4.0	**Text Analysis + Zipf's Law + TF-IDF**
4.1	Analysis of Words
4.2	Analysis of Sentences + Stemming + Lemmatization
4.3	Zipf's Law
4.4	Vectorization: TF-IDF
5.0	**Semantic Analysis**
5.1	Eigen Vectors and Values
5.2	Singular Value Decomposition
5.3	Latent Semantic Analysis
5.4	Latent Derichlet Allocation
6.0	**Word Embeddings**
6.1	Introduction to Word2Vec
6.2	Creating Word2Vec Using Keras and TF
7.0	**Transformers**
7.1	RNN + LSTM
7.2	Transformer: Positional Encoding + Self Attention
7.3	HuggingFace Transformer Library
7.4	BERT and Glove Embeddings
8.0	**Large Language Models (LLM) + Custom GPT**

8.1	Language Models
8.2	Custom GPT Using ChatGPT
8.3	Demo#1 (LangChain + OpenAI + HuggingFace)
8.4	Demo#2 (LlamaIndex + OPenAI + HuggingFace)
9	Review of the Course

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" (https://www.marshall.usc.edu/open-expression-statement).

ACADEMIC INTEGRITY

Academic Integrity:
The University of Southern California is a learning community committed to developing successful scholars and researchers dedicated to the pursuit of knowledge and the dissemination of ideas. Academic misconduct, which includes any act of dishonesty in the production or submission of academic work, compromises the integrity of the person who commits the act and can impugn the perceived integrity of the entire university community. It stands in opposition to the university's mission to research, educate, and contribute productively to our community and the world.

All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s).

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

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university. All incidences of academic misconduct will be reported to the Office of Academic Integrity and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

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 UNIVERSITY ACADEMIC AND 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 Lifeline) 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 is comprised 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) 740-5086 or (213) 821-8298
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.

AI Policy:

Permitted on specific assignments.

In this course, I encourage you to use artificial intelligence (AI)-powered programs to help you with assignments that indicate the permitted use of AI. You should also be aware that AI text generation tools may present incorrect information, biased responses, and incomplete analyses; thus they are not yet prepared to produce text that meets the standards of this course. To adhere to our university values, you must cite any AI-generated material (e.g., text, images, etc.) included or referenced in your work and provide the prompts used to generate the content. Using an AI tool to generate content without proper attribution will be treated as plagiarism and reported to the Office of Academic Integrity. Please review the instructions in each assignment for more details on how and when to use AI Generators for your submissions.

TECHNOLOGY REQUIREMENTS

All students coming into this course will need to be able to write, edit and run Python notebooks.

The lecture presentations, links to articles, assignments, quizzes, and rubrics are located on Blackboard. To participate in learning activities and complete assignments, you will need:

CLASS CONDUCT/NETIQUETTE

Professionalism will be expected at all times. Because the university classroom is a place designed for the free exchange of ideas, we must show respect for one another in all circumstances. We will show respect for one another by exhibiting patience and courtesy in our exchanges. Appropriate language and restraint from verbal attacks upon those whose perspectives differ from your own is a minimum requirement. Courtesy and kindness is the norm for those who participate in my class.

Our discussion board is a way for you to share your ideas and learning with your colleagues in this class. We do this as colleagues in learning, and the Discussion Board is meant to be a safe and respectful environment for us to conduct these discussions.

Some Netiquette Rules:

- Dress respectfully. Video conference business meetings are and will be the norm, so practice your professional telepresence.
- Your Virtual background should be professional
- Display both your first and last name during video conferencing and synchronous class meetings.
- Minimize the distractions of toggling muting and video when moving around
- Disagree respectfully
- Respectfully pay attention to classmates
- Do not use all CAPITAL LETTERS in emails or discussion board postings. This is considered "shouting" and is seen as impolite or aggressive.
- Do not use more than one punctuation mark, this is also considered aggressive!!!!
- Begin emails with a professional salutation (Examples: Dr. Name; Ms. Name; Hello Professor Name; Good afternoon Mr. Name). Starting an email without a salutation or a simple "Hey" is not appropriate.
- When sending an email, please include a detailed subject line. Additionally, make sure you reference the course number (Ex. BUAD306) in the message and sign the mail with your full name.
- Use proper grammar, spelling, punctuation, and capitalization. Text messaging language is not acceptable. You are practicing for your role as a business leader.
- Re-Read, think, and edit your message before you click "Send/Submit/Post.". As a check, consider whether you would be comfortable with your email or post or text being widely distributed on the Internet.

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.