Course	BUAN6342.501
Course Title:	Natural Language Processing
Professor	Gasan Elkhodari
Term	Spring 2025
Meetings	Wednesday 7:00 PM – 9:45 PM – Room JSOM 12.206

Professor Contact Information

Email Address:	Gasan.Elkhodari@utdallas.edu
Mobile:	(214) 883-3670
Office Location:	JSOM 3.428
Office Hours:	By appointment only
TA Information	shailender.joseph@utdallas.edu

General Course Information

**The descriptions and timelines contained in this syllabus are subject to change at the discretion of the Professor.*

Pre-requisites, Co-requisites, & other restrictions	BUAN 6341: There are no exceptions It is expected that the learners will be able to perform the following tasks independently: • Python Programming using machine learning packages e.g. Scikit-learn and data packages e.g. Pandas, Numpy • Ability to interpret fundamental quantitative concepts in Statistics, basic Linear Algebra (for matrix operations), machine learning as well as basic English language grammar (for NLP operations) • Perform basic management of toolkits e.g. debugging, software and package installation skills.

Course Description	This course focuses on natural language processing and the utility of textual data to gain meaningful quantitative and actionable insights about the language (mainly English) using rule-based and statistical methods and to extract the information for real-world applications. Our eventual goal will be to create machine-learning programs that analyze and interpret human language using classical text, social media and business text/unstructured data.
What we will cover	You will learn some very practical, state-of-the-art, relevant applications in a hands- on manner: • Unstructured Data Sourcing and Complication • Text Pre-processing in a realistic setting • Regular Expressions • Text Visualization • Topic Modeling • Text Summarization • Text Similarity • Content Search • Sentiment Analysis • Deep Learning applications (Text Generation) • Deep Learning -Text Based Classification and Clustering • Text Processing via KNIME analytical tool (zero code platform) The course uses Python's NLTK (Natural Language Toolkit) and spaCy as the primary vehicle. While this is an applied course, we will also focus on the key concepts behind linguistics, text processing and the overall theory of NLP so that the learners will appreciate the field of NLP. Topics include: • Various elements of linguistic understanding e.g. lexical, syntactic, and semantic • Understanding the statistical/dynamic properties of language such as n- grams, minimum edit distance, TFIDF etc. • Rule-based and data-driven approaches to building language models. • Machine learning for understanding language. • Application of NLP to real-world problems The theoretical understanding will be supplemented with hands-on assignments and practice using Python Jupyter notebooks with additional

	Python packages such as NLTK, scikit-learn, pandas, numpy, maptlotlib among others. PyTorch: PyTorch is a powerful open-source machine learning library that provides a wide range of functionality for building and training neural networks.
Learning Outcomes	By the end of this course, students will be able to: 1. Understand the use of regular expressions to detect patterns in text. 2. Understand text preprocessing using general techniques such as tokenization, stemming, lemmatization, stop-word removal, spell correction and word- frequency analysis. 3. Understand NLP specific techniques such as parsing, entity extraction, POS tagging and NER. 4. Apply machine learning, NLP and text processing techniques to build analytical applications to process unstructured data-sets. 5. Use multiple NLP packages e.g. NLTK, VADER, TextBlob, Gensim 6. Build your own Corpora for NLP applications 7. Apply an array of NLP operations to large and realistic datasets
Required Texts & Materials	For this course, we will be using the below required textbook. 1. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, by Bird, Klein and Loper (2nd edition). This book is freely available online. This book is available free. We will be referring to the Python 3.X version. 2. Srinivasa-Desikan, Bhargav, *Natural Language Processing and Computational Linguistics*, Pakt 2018

Week #	Week Start Date	Topic	Description	Lab #
1	1/22/2025	Class Introduction	Overview of NLP and its applications; Basic concepts and techniques in NLP; Introduction to spaCy and NLTK	1
2	1/29/2025	Text Preprocessing	Tokenization and stemming; Stopword removal and lemmatization; Text cleaning and normalization	2
	2/5/2025	Test Processing (Continue)		
4	2/12/2025	Topic Modeling	Advanced training tips Exploring documents Topic coherence and evaluating topic models Visualizing topic models	3
5	2/19/2025	Feature Engineering	One-hot encoding; Count vectorization; Term frequency-inverse document frequency (TF-IDF)	4
6	2/26/2025	Sentiment Analysis	Understanding the problem; Implementing a simple sentiment analysis model; Evaluating the model	5
7	3/5/2025	Word Embeddings	Understanding word embeddings; Training word embeddings; Using pre-trained word embeddings	6
8	3/12/2025	Assessment – 01		Details to be provided
9	3/19/2025	**Spring Break**		
10	3/26/2025	Text Classification	Understanding the problem; Implementing a simple text classification model using PyTorch	7
11	4/2/2025	Deep Leaning (Keras) Activation Functions and Text Generation	Introduction to sequence-to-sequence models and attention mechanisms; Implementing a machine translation	8
12	4/9/2025	Advanced NLP Applications Text Processing with Generative-AI	Text Processing with Generative-AI	9
13	4/16/2025	Exam One – At the testing center. (no class)		

14	4/23/2025	KNIME Analytics	Applications' practices	
15	4/30/2025	Assessment – 02 Individual project / Or an assessment.		
16	5/7/2025	Group Project Presentation		

Note: Any request for a makeup exam without a legitimate excuse will result in a 20% deduction.

Course Policies

- The Homework-Labs are to be completed in class and are due by the end of class, unless otherwise stated in eLearning.
- Makeup Exam: There is no makeup exams. In case of medical emergency, a medical report is required including physician information.
- Missing exam: Any missing exam without medical report will be graded as Zero.
- Assignments must be submitted through eLearning. Emailed submissions are not accepted.
- Late Assignments: Subject to 10% penalty, 20% penalty after the third day.
- Class Attendance: Students who fail to attend class regularly are inviting scholastic difficulty. Absences may lower a student's grade where class attendance and class participation are deemed essential by the instructor.
- UTD Syllabus Policies and Procedures: Please visit https://go.utdallas.edu/syllabus-policies
- Cheating will not be tolerated. When I find evidence of cheating, the documentation is turned over to the Office of Community Standards and Conduct. (https://www.utdallas.edu/conduct/dishonesty/)

Grading scale

Top 25% Students	**A**
Next 25% Students	**A-**
Next 25% Students	**B+**
Remaining 25% Students	**B and below**

Calculated Grade Weights**

- Labs / Assignments (15%)
- Exam I – (25%)
- Exam 2 – (25%)
- Group Project (25%)
- Class attendance (10%)

***The calculated grade weights are subject to change at the discretion of the Professor.*

Classroom citizenship

- Cell phone use is not allowed during class or exam.
- eLearning will be used for class content.
- Slides and other class materials will be posted after class is held.
- Class announcements (e.g., change in assignment dates) will be posted in the eLearning announcements. It is the students' responsibility to regularly check the announcements (typically by having the announcement automatically forwarded to their email accounts).