Course Syllabus

Jump to Today

COLUMBIA UNIVERSITY
Fall 2020 - Spring 2021

APAN P5540: Applied Text and Natural Language Analytics

Course: Date: September 29, 2020 - December 14, 2020
Time: Monday, 6:00 PM-8:00 PM
Location: 407 Mudd, 500 W. 120th St. Mudd

Contact Information:

Instructor: Jamil Hasenane

Email:	jh4272@columbia.edu
Office:	Zoom TBA
Office Hours:	By appointment
Response Policy:	During the term, the easiest way to reach me is via email. You will usually get a response within 48 hours. If you have a question about an assignment, you are advised to email me several days before it is due. If your email arrives within 24 hours of the due date, you may not receive a timely response.

Associate: Arun Kumar

Email:	arun2019@columbia.edu
Zoom:	Meeting ID: 917 560 0394 Passcode: 299673
Office Hours:	Wednesday, 3pm - 7pm EST
Response Policy:	During the term, the easiest way to reach me is via email. You will usually get a response within 48 hours. If you have a question about an assignment, you are advised to email me several days before it is due. If your email arrives within 24 hours of the due date, you may not receive a timely response.

Associate: Shubham Sahoo

Email:	ss3230@columbia.edu
Zoom:	Meeting ID: 220 483 7679 Password: 123456
Office Hours:	Monday, 4:45pm - 5:45pm EST
Response Policy:	During the term, the easiest way to reach me is via email. You will usually get a response within 48 hours. If you have a question about an assignment, you are advised to email me several days before it is due. If your email arrives within 24 hours of the due date, you may not receive a timely response.

Syllabus Contents

Course Overview

With the growth of the internet in recent decades, there has been an exponential increase of unstructured textual data available from news and social media. This data is invaluable for extracting actionable insights that can inform business decisions. Text analytics techniques, including natural language processing, have made the extraction of meaningful information possible only through the use of advanced natural language processing (NLP) and machine learning techniques. Jobs in this data analysis field increasingly require the use of these advanced techniques to extract insights from text data. Given the ever-increasing volume of textual data, it is important for students to be comfortable having to work on an ensemble of analytic tools and techniques.

This course will focus on advanced methods and systems that enable named entity recognition and disambiguation, topic modeling, sentiment analysis, word vector embeddings, distributed representations, meaning extraction, and deep learning for NLP. Weekly course lectures will offer a blend of theoretical material and hands-on practice with the latest open-source software libraries and commercial cloud services. Students will be expected to write code in Python to perform the general knowledge or applied NLP tasks commonly in science, business domains, including sales and marketing, financial modeling, credit risk analysis, legal trust and compliance, and healthcare. Students will be exposed to business problems and datasets that have recently chosen from contexts, including customer reviews from news, or forecasting indicators towards from public announcements.

In the second to last weeks of the semester, students will focus on the Term Project that may leverage both structured and unstructured data to perform descriptive analytics about an entity of interest (e.g. company, person, geolocation, etc.). For example, the Term Project may focus on extracting marketing topics or key business events from news articles about a specific company in order to explain changes in its stock price.

The desired outcome of the course is the ability to put the obtained knowledge into practical use. Whether you are a student, a data scientist, a business analyst, or a software developer, the knowledge and skills gained in this course should help you to meet the fundamental challenges of text analytics.

Note: While extensive programming skills are not required, students are expected to learn, understand and make use of basic data structures and functions in Python. A refresher session on Python 3 on the afternoon during one of the weekly classes if necessary. Students have an option to code weekly assignments in either Jupyter Notebook or under Python.

Prerequisites:

- APAN P5200: Applied Analytics Frameworks and Methods 1
- APAN P5201: Applied Analytics Frameworks and Methods 2

Back to top

Learning Objectives

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

1. 1. Devise structured data representation from an unstructured text
2. 2. Apply machine learning and deep learning techniques to extract insights out of raw unstructured text
3. 3. Model meaning or latent language understanding from raw text
4. 4. Develop descriptive, predictive, and prescriptive analytics models based on text
5. 5. Apply basic and sentiment analysis for contextual modeling

Back to top

Required Readings

Textbook:

Bengfort, B. (2018). Applied text analysis with Python: Enabling language-aware data products with machine learning 1st. O'Reilly Media, Inc. ISBN: 1491962992 [Hereafter referred to as "AWP"]

Hapke, H., Howard, C., & Lane, H. (2019). Natural language processing in action 1st. Manning Publications. ISBN: 1617294136 [Hereafter referred to as "NLPIA"]

Available online:

- Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python 1st. O'Reilly Media, Inc. ISBN: 9780596516499 [Hereafter referred to as Analyzing text with the natural language toolkit or "Bleanlp" referred to as "NLTK"]
- Jurafsky, D. (2018). An introduction to text summarization using the TextRank algorithm (with Python implementation) 1st. Analytics Vidhya Nov. 1, 2018.
- Lane, H., Howard, C., & Hapke, H. (2019). Speech and language processing 1st. Pearson Education India.

Back to top

Resources

Columbia University Information Technology:

Columbia University Information Technology's (CUIT) provides Columbia University students, faculty, and staff with a variety of computing resources and technical assistance services. Students, faculty and staff may access University-provided software through Courseware downloads.

Columbia University Library:

Columbia's extensive library system is ranks the top five academic libraries in the nation, with many of its services and resources available online.

SPS Academic Resources:

The Office of Student Affairs's provides students with academic counseling and support services such as online tutoring and crate coaching.

Back to top

Course Requirements (Assignments)

Detailed descriptions and assessment criteria will be provided for all assignments in the Canvas/Courseware site.

Term Project (40pts with L1 L2)

The term project builds on the content and sequence of weekly lectures and assignments. Its purpose is to develop an end-to-end data science solution to a business problem using NLP and ML techniques. Students will work on this individual assignment. Each group can develop own project definition, subject to the instructor's approval.

The proposed data science solution must leverage, at least three NLP and text mining techniques studied in the course, e.g. entity recognition, topic modeling, semantic analysis using word embeddings, locality sensitive hashing, etc. Sample problem definitions may include analyzing earnings statements of a company to predict stock price movements, summarizing or categorizing contracted clauses, training a chatbot to answer questions about a product, predicting customer satisfaction or assessment as well as sales and market intelligence based on news media.

Groups are expected to implement their proposed end to end solution in Python and document their work in a Jupyter Notebook. The instructor will meet with each group during the last two classes. (See "Final Presentation" below for more details).

The Term Project comprises the following deliverables:

- Term Project Proposal

In Week 10, you will submit a short (i.e. 1 page) proposal in which you select a business use case for the Term Project and define the business impact and value of the project.

- Term Project Deliverable 1: Topic Modeling - Python Implementation
  - Using topic modeling techniques studied in class, develop a topic taxonomy for the text data
  - Train word vector models based on the dataset and generate topic classifications of articles using semantic similarity
  - Evaluate the topic models using coherence and human evaluation
- Term Project Deliverable 2: Sentiment Classifier Modeling - Python Implementation
  - Develop sentiment polarity and subjectivity models for the dataset articles using NLTK
  - Train a sentiment classifier using Maxent Bayes
  - Evaluate the classifier using cross validation

The Term Project Summary shall be presented in the form of either a 3-page written document or 10- to 12-page slide deck and cover the following key elements of the project:

- Background information and objective of the project
- Data sources, data preparation and preprocessing steps
- Design choices and the techniques for the implemented methodologies
- Evaluation and model explainability criteria and metrics
- Findings and conclusions

Assignments (40pts with L1 L4)

Weekly assignments give students opportunities to practice foundational skills and are intended to reinforce the concepts and methodologies covered in class. The

- **Assignment 1 (40pts with L1):** Install Python 3, Jupyter Notebook and select the ecosystem of installations. Implement an algorithm using Python dictionary, list, and set.
- **Assignment 2 (40pts with L1):** Write a Python program that:
  - Reads a text file and computes a word count or a bigram/trigram dataset of each unique word in a corpus vocabulary
  - Adds the dataset documents into a persistent storage for use in weekly assignments and the term project
- **Assignment 3 (40pts with L1):** Write a Python program using the regular expression (re), SpaCy and/or NLTK packages to:
  - Tokenize words and sentences
  - Stem and lemmatize word tokens
  - Generate n-grams
  - List part-of-speech or governs for a given n
- **Assignment 4 (40pts with L2):** Write a Python program to:
  - Read in a text file and compute document term frequencies
  - Recognize and label entities mentioned in the text by a knowledge base
- **Assignment 5 (40pts with L2):** Write a Python program to:
  - Apply TextRank for ranking and selecting key phrases in a document
  - Apply LexRank for extractive sentence summarization
- **Assignment 6 (40pts with L4):** Write a Python program to train a word vector model using Word2Vec, Glove, or FastText word embedding algorithms.
- **Assignment 7 (40pts with L4):** Write a Python program that filters out exactly and/or semantically duplicate articles from the dataset, using SimHash and Locality Sensitive Hashing.
- **Assignment 8 (40pts with L4):** Write a Python program to calculate document similarity using LSH, matrix decomposition and dimensionality reduction.
- **Assignment 9 (40pts with L4):** Write a Python program that:
  - Implements LDA topic modeling or the document using a choice of Gensim, Scikit-learn or Spark MLLib libraries
  - Extracts topic clusters keywords
  - Assigns each article to the relevant cluster based on topics

Final Presentation (40pts with L1 L2)

At the end of the course, you (or your group) will prepare and deliver a 10-minute presentation on your Term Project. In your presentation, you should address:

- Data source specification & the choice of entity
- Design choices and implementation
- Evaluation criteria and results
- Business impact

You will be assigned to Zoom breakouts to present your Term Project. Your well-you, facilitate class discussion and Q&A, and on the quality of your slide deck (i.e. does it contain the necessary elements?). Your presentation will be graded on the quality of adherence with (or, in the case of presenting the Project as a conceptual idea or decision makers within an organization - would the presentation inspire confidence that the potential to your Project will be worth undertaking and investing organizational time and resources to?

Attendance and Participation

Your attendance and active participation are essential to success in this course. You are expected to attend all class sessions and come to class prepared, having completed all assigned readings and assignments. During our class meetings, you are expected to ask questions, participate in discussions, and contribute to a collaborative and respectful atmosphere. Your ability to ask questions, offer your own ideas, and challenging the points of view of others. If you need to miss a class for any reason, please discuss the absence with me in advance.

Peer Review:

Part of your participation grade will be earned through a short (i.e. 500 word) peer review of a classmate's term project. The purpose of the peer review is to give the opportunity to address the evaluation of your own thinking and learning in this course. You will be asked to identify areas where you were challenged and where you still have questions or feel uncertain about your work. You will also be asked to describe how you have to apply core learnings to your professional life.

Back to top

Evaluation / Grading

The final grade will be calculated as described below:

GRADE CALCULATION	
ASSESSMENT	WEIGHT
Team Project, including:	50%
• Proposal	5%
• Python Implementation	20%
• Final Presentation	25%
Assignments	40%
Final Reflection	5%
Attendance and Participation	5%

FINAL GRADING SCALE

GRADE	PERCENTILE
A+	98 - 100 %
A	95.97 %
A-	90-94.9 %
B+	87-89.9%
B	84-86.9 %
B-	80-83.9 %
C+	77-79.9 %
C	74-76.9 %
C-	70-73.9 %
F	69-69.9 %
F	57.9 % and below

Back to top

Course Policies

Participation and Attendance:

Your attendance and active participation are essential to success in the course. You are expected to attend all class sessions and come to class prepared, having completed all assigned readings and assignments. During our class meetings, you are expected to engage with your peers and participate in discussions and contribute to a collaborative and respectful atmosphere. Your ability to ask questions, offer your own ideas, and challenging the points of view of others. If you need to miss a class for any reason, please discuss the absence with me in advance.

Late work

Work that is not submitted on the due date related in the course syllabus without advance notice and permission from the instructor will be graded down 1/3 of a grade for every day it is late (e.g., from A+ to A- to B+).

Grades & Submission

All written assignments must use APA format, cite sources, and be submitted to the course website (not via email).

Back to top

School Policies

Copyright Policy

Please note - Due to copyright restrictions, online access to the material is limited to instructors and students currently registered for the course. Please be advised that by clicking the link to the electronic materials in this course, you have read and accept the following:

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use," that user may be liable for copyright infringement.

Academic Integrity:

Columbia expects students to be students to act with honesty and propriety at all times and to respect the rights of others. It is fundamental University policy that academic dishonesty in any guise or pretense, conduct of any sort that disrupts the life of the University or denigrates or endangers members of the University community, or violations of University rules and regulations, will not be tolerated. The integrity and reliability of this community that individuals do their own work, that they acknowledge the circumstances, ideas, sources, and assistance upon which their work is based. Academic honesty in class assignments and exams is expected of all students at all times.

SPS holds each member of its community responsible for understanding and abiding by the SPS Academic Integrity Policy and Honor Code. Violations of these policies will be treated as a matter of the gravest nature of class. Ignorance of the School's policy concerning academic dishonesty shall not be a defense in any disciplinary proceedings.

Accessibility:

Columbia is committed to providing equal access to qualified students with documented disabilities. A student requesting academic accommodations must provide current documentation of a qualifying disability-documentation and related information gathered through the intake process. For more information regarding this service, please visit the University's Health Services website.

Class Recordings:

All or portions of the class may be recorded at the discretion of the instructor to support your learning. At any point, the instructor has the right to discontinue the recording if it is deemed to be obstructive to the learning process.

Video or audio recording is prohibited, it is considered confidential and it is not acceptable to share the recording outside the purview of the faculty member and registered class.

Back to top

Course Schedule

Week	Topic(s)	Readings	Activities / Assignments
1	Introduction & Course Overview	NLTK, Ch 1: Language Processing & Python AWP, Ch 1: Language & Computation (1ff)	• Resources • Session • Assignment 1
2	Data Sources & Crawling	NLTK, Ch 3: Processing Raw Text AWP, Ch 2: Building a Corpus (Corpus 1ff)	• Resources • Session • Assignment 2
3	Basic Text Processing	NLPIA, Ch 2: Build your vocabulary (word tokenization) (33) AWP, Ch 3: Corpus Preprocessing & Wrangling (Wrangle 1ff)	• Resources • Session • Assignment 3
4	Information Extraction & Named Entity Recognition & Linking	NLPIA, Ch 11: Information Extraction (named entity extraction) (311) NLTK, Ch 7: Extracting information from Text (279)	• Resources • Session • Assignment 4
5	Information Extraction & Text Classification & Summarization	NLPIA, Ch. 3: Math with words (53) Jurafsky & Martin: Introduction to text summarization using the TextRank algorithm (20)	• Resources • Session • Assignment 5
6	Vector Space Modeling, Word Embeddings, Neural Networks	NLPIA, Ch 6, Reasoning with word vectors (word2vec) (186) SLP, Section 6.8: Vector Semantics: Word2Vec (6)	• Resources • Session • Assignment 6
7	Locality Sensitive Hashing (LSH) & Text Deduplication	NLPIA, Appendix F: Locality Sensitive Hashing (LSH)	• Resources • Session • Assignment 7
8	Topic Modeling	NLPIA, Ch 4: Sections 4.1-4.3(44) AWP, Ch 6: Clustering for Text Similarity (17)	• Resources • Session • Assignment 8
9	Evolving Topic Classification	No readings	• Resources • Session • Assignment 9
10	Term Project Proposal Presentations	No readings	• Session • Term Project Proposal
11	Deep Learning in NLP	Towards Data Science Blog Post: How attention works in deep learning BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (presentation slides by Jacob Devlin, Google AI)	• Resources • Session
12	Sentiment Analysis	No readings	• Session
13	Final Term Project Presentations	NLPIA, Ch 5: Sections 4.1-4.3(44) AWP, Ch 6: Clustering for Text Similarity (17)	• Session • Term Project Deliverable 1: Live Code Presentation • Term Project Deliverable 2: Live Code Presentation • Term Project Summary Presentation Slide Deck or Summary Final Reflection

Back to top

Course Summary:

Date	Details
Sun Sep 13, 2020	Discussion: Setting Accepted
Mon Sep 14, 2020	1: Session
Mon Sep 21, 2020	2: Session
	Assignment 1
Mon Sep 28, 2020	3: Session
Mon Sep 28, 2020	Assignment 2
Mon Oct 5, 2020	4: Session
Mon Oct 5, 2020	Assignment 3
Mon Oct 12, 2020	5: Session
Mon Oct 14, 2020	Assignment 4
Mon Oct 19, 2020	6: Session
Mon Oct 19, 2020	Assignment 5
Mon Oct 26, 2020	7: Session
Tue Oct 27, 2020	Assignment 6
Mon Nov 9, 2020	8: Session
Mon Nov 9, 2020	Assignment 7
Mon Nov 16, 2020	9: Session
Tue Nov 17, 2020	Assignment 8
Mon Nov 23, 2020	Assignment 9
Mon Nov 23, 2020	10: Session
Sun Nov 29, 2020	Term Project Proposal
Mon Nov 30, 2020	11: Session
Mon Dec 7, 2020	12: Session
Mon Dec 14, 2020	13: Session
	Term Project Deliverable 1: Live Code Presentation
Fri Dec 18, 2020	Final Reflection
	Term Project Deliverable 2: Presentation
	Term Project Deliverable 3: Presentation