Course Syllabus

Jump to Today

COLUMBIA UNIVERSITY
School of Professional Studies

APAN PS5430: Applied Text and Natural Language Analytics

Schedule	Dates: May 26, 2020 – August 14, 2020 Time: Wednesday, 8:10 PM-10:00 PM Online
Credits	3
Contact Information	**Instructor: Javid Huseynov, Ph.D.** **Email** jbh2172@columbia.edu **Phone** (646) 926-7526 **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. **Olga Vilenskaia** **Email** **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.

May 2020

26	27	28	29	30	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

Assignments are weighted by group:

Group	Weight
Assignments	45%
Term Project	40%
Final Presentation	10%
Attendance & Participation	5%
Total	100%

Syllabus Content

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 enhance the scale and the quality of business analytics. But the enormous volume of domain text corpora makes the extraction of meaningful information possible only through the use of advanced natural language processing (NLP) and machine learning techniques. Jobs in the data analysis field increasingly require the use of extracting and analyzing information from diverse sources, structured as well as unstructured. This course will therefore train students in a technology that is seen as an essential part of a data analyst's toolkit.

This course will focus on advanced methods and systems that enable named entity recognition and disambiguation, topic modeling, sentiment analysis, word vector embeddings, abstractive summarization, meaning extraction, and deep learning for NLP. Weekly course lectures will offer a blend of theoretical material and hands-on class exercises, which will be put into practice through weekly assignments. Students who complete the course will be able to practice the gained knowledge as applied NLP data scientists in various business domains, including sales and marketing, financial modeling, credit risk analysis, legal trust and compliance, intellectual property and contracts management. Some examples include extraction of payment clauses from contracts, establishing customer needs from news, or forecasting industry trends from public announcements.

During the last four weeks of the semester, students will focus on the Term Project that may leverage both unstructured and structured data to discover knowledge about an entity of interest (e.g. company, person, geolocation, etc.). For example, the Term Project may focus on extracting evolving topics or key business events from news articles about a chosen public 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 taking this course for future academic research, for work in industry, or for an innovative startup idea, this course should help you to master the fundamentals of unstructured data analytics.

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

Prerequisites

- APAN PS5200: Applied Analytics Frameworks and Methods 1
- APAN PS5205: Applied Analytics Frameworks and Methods 2

Back to top

Learning Objectives

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

- L1: Derive structured data representation from an unstructured text
- L2: Extract entities and their contexts (relations, keywords, concepts, etc.) from unstructured text
- L3: Model meaning or natural language understanding from text corpora
- L4: Develop descriptive, predictive, and prescriptive analytics models based on text
- L5: Develop opinion / sentiment analysis model based on text corpora

Back to top

Required Readings

To purchase:

Bengfort, B. (2018). *Applied text analysis with Python: Enabling language-aware data products with machine learning*. O'Reilly Media, Inc. ISBN: 1491963042 (Hereafter referred to as "ATAP")

Hapke, H., Howard, C., & Lane, H. (2019). *Natural language processing in action*. Manning Publications. ISBN: 9781617294631 (Hereafter referred to as "NLPA")

Available online:

- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc. ISBN: 9780596550967. Also published as *Analyzing text with the natural language toolkit*. (Hereafter referred to as "NLTK")
- Joshi, P. (2018). An introduction to text summarization using the TextRank algorithm (with Python implementation). Analytics Vidhya. Nov. 1, 2018.
- Jurafsky, D., & Martin, J. (2018). *Speech and language processing*. Third Edition Draft. (Hereafter referred to as "SLP")

Back to top

Resources

Columbia University Information Technology

Columbia University Information Technology (CUIT) provides Columbia University students, faculty and staff with central computing and communications services. Students, faculty and staff may access University-provided and discounted software downloads.

Columbia University Library

Columbia's extensive library system ranks in 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 provides students with academic counseling and support services such as online tutoring and career coaching.

Back to top

Course Requirements (Assignments)

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

Term Project (40%) (Aligns with L1-L5)

The Term Project leverages unstructured news data and a blend of natural language processing (NLP) and text analytics methods covered in the course, to discover insights about an entity of interest (e.g. company, person, geographic location). Such insights include extracted and disambiguated named entities, evolving topics over varying time ranges, aggregated news summaries, and sentiment. More specific examples may include analyzing earnings statements of a company to predict stock price movements, extracting or categorizing contractual clauses, or tracking intellectual property infringements from the product announcements.

The project builds on the content and sequence of weekly lectures and assignments. Depending on the number of students, the project may be an individual or a group assignment. The instructor will discuss the project outline and requirements in detail during a dedicated class session. Subsequently, students (individually or in groups) will have to define and develop their projects around well-articulated business problems. Some sample business cases include credit and risk assessment, marketing lead generation, stock market analysis, and sales recommendations. Students are expected to implement a solution and present their findings in a 10-minute presentation during the final class session, and submit their Python code and PowerPoint slides for instructor review and evaluation (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 (ca. 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 from the dataset
  - Train word vector models based on the dataset and generate topic classifications of articles using semantic similarity
  - Aggregate topics over time range to identify their evolution
- *Term Project Deliverable 2: Sentiment/Opinion Mining - Python Implementation*
  - Develop sentiment polarity and subjectivity models for the dataset articles using NLTK
  - Train a sentiment classifier using Naive Bayes
- *Term Project Deliverable 3: Final Project Summary*

The Final 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 source specification and procurement details
- Design choices and the rationale for the implemented methodologies
- Evaluation criteria and metrics
- Findings and Conclusions

Assignments (45%) (Aligns with L1-L4)

Weekly assignments give students opportunities to practice foundational skills and are

designed to prepare students for a successful completion of the Term Project.

- **Assignment 1 (Aligns with L1) :** Install PyCharm & Jupyter Notebook and submit the screenshots of installations. Implement an algorithm using Python dictionary, list, and set.
- **Assignment 2 (Aligns with L1):** Write a Python program that:
  - Implements Webhose.io API calls to obtain a JSON document dataset of web crawls about a chosen entity
  - Adds the dataset documents into a persistent storage for use in weekly assignments and the term project
- **Assignment 3 (Aligns with L1):** Write a Python program using the regular expression (re), SpaCy and/or NLTK packages to:
  - Tokenize words and sentences
  - Stem and lemmatize work token
  - Remove stop words
  - List and count *n*-grams for a given *n*
- **Assignment 4 (Aligns with L2):** Write a Python program to:
  - Train a simple named entity recognizer using spaCy library
  - Recognize and link entities mentioned in text to a knowledge base
- **Assignment 5 (Aligns 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 (Aligns with L3):** Write a Python or PySpark program to train a word or character embedding model using Word2Vec, Glove or FastText algorithms.
- **Assignment 7 (Aligns with L3):** Write a Python program that filters out exactly and/or semantically duplicate articles from the dataset, using SimHash and Word2Vec.
- **Assignment 8 (Aligns with L4):** Write a Python program to calculate document similarity using LSA: matrix decomposition and dimensionality rank reduction.
- **Assignment 9 (Aligns with L4):** Write a Python program that:
  - Implements LDA topic modeling on the dataset using a choice of Gensim, Scikit-learn or Spark MLLib libraries
  - Extracts topic cluster keywords
  - Assigns each article to the relevant cluster based on topics

Final Presentation (10%) (Aligns with L1-L5)

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 implemented solution(s)
- Evaluation criteria and results
- Findings and conclusions

You will be assessed on how clearly you present your Term Project; how well you facilitate class discussion and Q&A; and on the quality of your slide deck (i.e. does it present information clearly, effectively, and appealingly?). The instructor's frame of reference will be as if you were presenting this Project as an innovative idea to decision-makers within an organization - would the presentation inspire confidence that the activities in your Project are worth undertaking and investing organizational time and resources in?

Attendance and Participation (5%)

Your attendance and active participation are essential to succeed 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 engage with your classmates and instructor by answering questions, stating and defending your point of view, 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.

Final Reflection

Part of your participation grade will be earned through a short (ca. 500 word) reflection that you will write at the end of the course. This reflection will give you the opportunity to address the evolution 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 lingering questions to explore in the future. You will also be asked to describe how you hope to apply your learnings to your professional life.

Back to top

Evaluation / Grading

The final grade will be calculated as described below:

GRADE CALCULATION

ASSIGNMENT	WEIGHT
Team Project	40%
Assignments	45%
Final Reflection & Project Summary	10%
Attendance and Participation	5%

FINAL GRADING SCALE

GRADE	PERCENTAGE
A+	98–100 %
A	93–97.9 %
A-	90–92.9 %
B+	87–89.9%
B	83–86.9 %
B-	80–82.9 %
C+	77–79.9 %
C	73–76.9 %
C-	70–72.9 %
D	60–69.9 %

F	59.9 % and below

Back to top

Course Policies

Participation and Attendance

Your attendance and active participation are essential to succeed 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 engage with your classmates and instructor by answering questions, stating and defending your point of view, 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 noted 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 B+ to a B).

Citation & 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 this material is limited to instructors and students currently registered for this 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 materials. 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 University expects its 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 personal conduct of any sort that disrupts the life of the University or denigrates or endangers members of the University community is unacceptable and will be dealt with severely. It is essential to the academic integrity and vitality of this community that individuals do their own work and properly acknowledge the circumstances, ideas, sources, and assistance upon which that 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 and Community Standards. You are required to read these standards within the first few days 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's disability status and reasonable accommodations are individually determined based upon 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.

If the recording is posted, 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	Objectives
1	Introduction & Course Overview	NLTK, Ch 1: Language Processing & Python (29) ATAP, Ch 1: Language & Computation (18)	Due by 11:59 8 days after class: Assignment 1: Install PyCharm & Jupyter Notebook and submit the screenshots of installations. Implement an algorithm using Python dictionary, list, and set.	• **1.1**: Install and configure Python development tools, including PyCharm IDE, Anaconda & Jupyter Notebook • **1.2:** Install Python libraries using pip, conda • **1.3:** Identify basic Python data structures • **1.4:** Implement a simple algorithm in Python using data structures, library functions
2	Data Sources & Crawling	NLTK, Ch 3: Processing Raw Text (49) ATAP, Ch 2: Building a Custom	Due by 11:59 8 days after class: **Assignment 2:** Write a Python program that: • Implements Webhose.io API calls to obtain a JSON document dataset of web	• **2.1:** Build a text corpus • **2.2:** Recognize different packages that deal with specific document formats (Word, JSON, PDF, HTML, etc.), encodings (UTF-8, ASCII, etc.) and different types of documents (emails,

		Corpus (17)	dataset of web crawls about a chosen entity • Adds the dataset documents into a persistent storage for use in the weekly assignments	tweets, web pages, SEC filings, etc.) • **2.3:** Extract documents from the web and from document repositories • 2.4: Extract metadata
3	Basic Text Processing	NLPA, Ch 2: Build your vocabulary (word tokenization) (58) ATAP, Ch 3: Corpus Preprocessing & Wrangling (17)	Due by 11:59 8 days after class: **Assignment 3:** Write a Python program using the regular expression (re), SpaCy and/or NLTK packages to: • Tokenize words and sentences • Stem and lemmatize work token • Remove stop words • List and count *n*-grams for a given *n*	• **3.1:** Explore regular expression patterns and functionality • **3.2:** Strip HTML tags, images, code scripts • **3.3:** Tokenize words and sentences • **3.4:** Lemmatize stem word tokens • **3.5:** Assign Part-of-Speech (POS) tags

4	Information Extraction 1:
	Named Entity Recognition & Linking
		NLPA, Ch 11: Information Extraction (named entity extraction) (35)
		NLTK, Ch 7: Extracting information from Text (29)
			Due by 11:59 8 days after class:
			Assignment 4:
			Write a Python program to
			• Train a simple named entity recognizer using spaCy library
			• Recognize and link entities mentioned in text to a knowledge base
				• **4.1:** Take preconditioned text and apply transformations for:
				  ◦ Chunking/shallow parsing
				  ◦ Tagging named entities
				  ◦ Entity Recognition
				  ◦ Entity Disambiguation
				  ◦ Coreference Resolution

5	Information Extraction 2: Key Phrase Extraction & Text Summarization	NLPA, Ch. 3: Math with words (33) Joshi, P. (2018), Introduction to text summarization using the TextRank algorithm (20)	Due by 11:59 8 days after class: **Assignment 5:** Write a Python program to: • Apply TextRank for ranking and selecting key phrases in a document • Apply LexRank for extractive sentence summarization	• **5.1:** Extract key phrases from text using unsupervised (statistical or graph-based) and supervised methods • **5.2:** Distinguish Abstractive from Extractive Summarization • **5.3:** Apply TextRank, LexRank and other algorithms for key phrase and sentence ranking
6	Vector Space Modeling using Neural Networks	NLPA, Ch 6: Reasoning with word vectors (word2vec) (56) SLP, Section 6.8: Vector Semantics: Word2Vec (5)	Due by 11:59 8 days after class: **Assignment 6:** Write a Python or PySpark program to train a word or character embedding model using Word2Vec, Glove or FastText algorithms	• **6.1:** Recognize word vectors / embeddings and different architectures • **6.2:** Train models using the popular word vector algorithms • **6.3:** Evaluate contextual or semantic similarity and find synonyms
7	Locality Sensitive Hashing (LSH) & Text Deduplication	NLPA, Appendix F: Locality Sensitive Hashing (11)	Due by 11:59 8 days after class: **Assignment 7:** Write a Python program that filters out exactly and/or semantically duplicate articles from the dataset, using SimHash and Word2Vec.	• **7.1:** Deduplicate texts using LSH methods • **7.2:** Experiment with distance measures • **7.3:** Use SimHash & Word Vectors for semantic deduplication
8	Topic Modeling 1: Algebraic methods LSA and SVD	NLPA, Ch 4: Sections 4.1-4.3(44) ATAP, Ch 6: Clustering for Text Similarity (27)	Due by 11:59 8 days after class: **Assignment 8:** Write a Python program to calculate document similarity using LSA: matrix decomposition and dimensionality rank reduction	• **8.1:** Apply Latent Semantic Analysis and Singular Value Decomposition to measure the document similarity • **8.2:** Construct a basic tagging mechanism using LSA
			Due by 11:59 8 days after class: **Assignment 9:** Write a Python	• **9.1**: Build an LDA topic model to develop topic clusters out of a

9	Topic Modeling 2: Probabilistic methods, LDA	NLPA, Ch 4: Sections 4.5, 4.6 (17) ATAP, Ch 6: Clustering for Text Similarity (27)	program that: • Implements LDA topic modeling on the dataset using a choice of Gensim, Scikit-learn or Spark MLLib libraries • Extracts topic cluster keywords • Assigns each article to the relevant cluster based on topics **Term Project Proposal**	dataset using Python packages (sklearn, gensim and Spark MLLib) • **9.2:** Determine the optimal number of topics on a given dataset • **9.3:** Apply LDA model to assign each article in a dataset to the appropriate topic cluster
10	Evolving Topic Classification	No readings	Due by 11:59 8 days after class: **Term Project Deliverable 1: Topic Modeling - Python Implementation:** • Using topic modeling techniques studied in class, develop a topic taxonomy from the dataset • Train word vector models based on the dataset and generate topic classifications of articles using semantic similarity • Aggregate topics over time range to identify their evolution	• **10.1:** Distinguish taxonomy from ontology • **10.2:** Construct evolving topic taxonomy • **10.3:** Classify articles into semantically similar taxonomy categories using word vector models
11	Sentiment Analysis	No readings	Due by 11:59 8 days after class: **Term Project Deliverable 2:** **Sentiment/Opinion Mining - Python Implementation:** • Develop sentiment polarity and subjectivity models for the dataset articles using NLTK • Train a sentiment classifier using Naive Bayes	• **11.1**: Distinguish between sentiment polarity, subjectivity, and emotion • **11.2:** Train a basic sentiment analyzer using a labeled dataset • **11.3:** Apply sentiment analysis in conjunction with topic classification to correlate sentiment with evolving topics
12	Final Term Project Presentations	No readings	Due in class: **Final Presentation** Due by 11:59 5 days after class: **Term Project Deliverable 3: Final Project Summary** **Final Reflection**	• **12.1:** Design and deliver an impactful presentation that explains analytical outputs to a professional audience • **12.2**: Synthesize and leverage all your class learnings in your Term Project

Back to top

Course Summary:

Date	Details	
Wed May 27, 2020	1. Session	due by 8:10pm
	Discussion: Getting Acquainted	to do: 11:59pm
Wed Jun 3, 2020	Assignment 1	due by 7pm
	2. Session	due by 8:10pm
Wed Jun 10, 2020	Assignment 2	due by 7pm
	3. Session	due by 8:10pm
Wed Jun 17, 2020	Assignment 3	due by 7pm
	4. Session	due by 8:10pm
Wed Jun 24, 2020	Assignment 4	due by 7pm
	5. Session	due by 8:10pm
Wed Jul 1, 2020	Assignment 5	due by 7pm
	6. Session	due by 8:10pm
Wed Jul 8, 2020	Assignment 6	due by 7pm
	7. Session	due by 8:10pm
Wed Jul 15, 2020	Assignment 7	due by 7pm

Wed Jul 15, 2020	8. Session	due by 8:10pm
Wed Jul 22, 2020	Assignment 8	due by 7pm
	9. Session	due by 8:10pm
Sun Jul 26, 2020	Term Project Proposal	due by 11:59pm
Wed Jul 29, 2020	Assignment 9	due by 7pm
	10. Session	due by 8:10pm
Sun Aug 2, 2020	Term Project Deliverable 2: Python Implementation	due by 11:59pm
Wed Aug 5, 2020	12. Session	due by 8:10pm
Sun Aug 9, 2020	Term Project Deliverable 2: Sentiment/Opinion Mining - Python Implementation	due by 11:59pm
Wed Aug 12, 2020	13. Session	due by 8:10pm
	Term Project Deliverable 1: Live Class Presentation	due by 8:10pm
Fri Aug 14, 2020	Final Reflection	due by 11:59pm
Sun Aug 16, 2020	Term Project Deliverable 3: Presentation Slide Deck or Summary	due by 11:59pm