DS 7337: Natural Language Processing

Overview

Natural language processing from an application-builder's perspective, using Python with NLTK and other packages. Emphasis on readying the student to become a working professional in real-world scenarios.

Course Description

This class introduces natural language processing (NLP) as applied to text mining and other tasks with unstructured big data. Students will receive a broad survey of the major tasks in natural language processing. Topics include document clustering and classification, automated tagging, topic modeling, and feature extraction to support a variety of applications. The focus is on best practices for choosing the right tool and method for an application, illustrated with real-world case studies. The class offers experience building solutions from real-world data sets, utilizing WordNet and the data interfaces of leading websites.

Learn: Feature extraction from unstructured text, text classification, document clustering, sentiment analysis, corpus analytics, syntax parsing, topic modeling

Apply: Python with NLTK, WordNet, scikit-learn, and several other Python packages

Core Resources

Python 2.7.X, available at https://www.python.org/download/releases/2.7/

NLTK (Natural Language Toolkit) available at http://www.nltk.org/install.html

Bird, Steven, Ewan Klein, and Edward Loper, *Natural Language Processing with Python* (NLTK, 2009, 2014). Available free online at http://www.nltk.org/book/

Sarkar, Dipanjan, *Text Analytics with Python* (New York: Apress, 2016). Available on Amazon or at https://www.apress.com/us/book/9781484223871

Code repositories for Sarkar available at https://github.com/dipanjanS/text-analytics-with-python

Weekly Schedule

Unit	Topic	Reading	Assignments Due
1	**Introduction:** What is NLP? Two sides of NLP Applications of NLP	Sarkar pp. 1–9 in Chapter 1 "Natural Language Basics" Bird-Klein Preface, Chapter 1 "Language Processing and Python"	Begin your glossary by creating entries for natural language, artificial language, NLP, NLU, NLG Homework 1, due the same day as the week 2

			live session Quiz 1, due the day after the week 1 live session
2	**Levels of Analysis in NLP** Lexical analysis Syntactic analysis Semantic analysis Discourse analysis	Sarkar Chapter 2 "Python Refresher" Bird-Klein Chapter 2 "Accessing Text Corpora and Lexical Resources"	Update your glossary with terms from this week Get a head start on Homework 2 Quiz 2, due the day after the week 2 live session
3	**Trade-Offs in NLP** Shallow vs. Deep Statistical vs. Symbolic Feature engineering vs. feature learning Top-down vs. bottom-up Transparent vs. opaque (AI vs. XAI)	Sarkar pp. 10-27 in Chapter 1 "Natural Language Basics" Bird-Klein Ch. 3 "Processing Raw Text"	Update your glossary with terms from this week Homework 2, due the same day as the week 4 live session Quiz 3, due the day after the week 3 live session
4	**Working in NLP** NLP and data science Job roles that utilize NLP Sectors that utilize NLP Organizations that relate to NLP	Sarkar pp. 28–50 in Chapter 1 "Natural Language Basics" Bird-Klein Chapter 4 "Writing Structured Programs"	Update your glossary with terms from this week Get a head start on Homework 3
5	**Low-Level Analysis** Text preprocessing Text normalization Low-level document feature extraction	Sarkar, "Text Tokenization" and "Text Normalization" pp. 107–131 in Chapter 3 Sarkar, "Text Normalization," "Feature Extraction," and "Keyphrase Extraction" in Chapter 5 pp. 223–233	Update your glossary with terms from this week Homework 3, due the same day as the week 6 live session Quiz 4, due the day after the week 5 live session

6	**Lexical Knowledge Bases** Lexical knowledge bases Resources for creating or extending lexical knowledge bases Applications of lexical knowledge bases	"What Is WordNet?" at https://wordnet.princeton.edu/ Sarkar "Exploring WordNet" in Chapter 7 pp. 321–329	Update your glossary with terms from this week Get a head start on Homework 4 Quiz 5, due the day after the week 6 live session
7	**Syntactic Analysis: POS-Tagging** POS-tagging Using POS tags	Bird-Klein Chapter 5 "Categorizing and Tagging Words" Sarkar "Part-of-Speech (POS) Tagging" in Chapter 3, pp. 135–142	Update your glossary with terms from this week Homework 4, due the same day as the week 8 live session
8	**Syntactic Analysis: Parsing** Shallow parsing Using chunks Full grammar parsing Uses for full parse trees	Bird-Klein Chapter 8 "Analyzing Sentence Structure" Sarkar "Understanding Text Syntax and Structure" in Chapter 3 pp. 132–166	Update your glossary with terms from this week Homework 5, due the same day as the week 9 live session Submit your glossary for check-in, due the same day as the week 9 live session Quiz 6, due the day after the week 8 live session
9	**Midterm Review** The midterm exam is available immediately after the week 9 live session and is due five days after the week 9 live session at 11:59 p.m. HST		Midterm
10	**Semantic Analysis: Semantic Relatedness** Word similarity Document similarity Applications of semantic similarity	Sarkar pp. 265–295 in Chapter 6 "Text Similarity and Clustering"	Update your glossary with terms from this week Homework 6, due the same day as the week 11 live session Quiz 7, due the day after the week 10 live session

11	**Semantic Analysis: Document Clustering** Methods of Clustering Working with Clusters	Sarkar pp. 296–318 in Chapter 6. “Text Similarity and Clustering”	Update your glossary with terms from this week Get a head start on Homework 7 Quiz 8, due the day after the week 11 live session
12	**Semantic Analysis: Text Classification** Document Classification SVMs for text classification Descriptor-based text classification	Sarkar Chapter 4 “Text Classification” Bird-Klein Chapter 6 “Learning to Classify Text”	Update your glossary with terms from this week Homework 7, due the same day as the week 13 live session
13	**Semantic Analysis: Topic Modeling** Topic Model Paradigms Organic Topic Modeling Canonical Topic Modeling Entity-Centric Topic Modeling	Sarkar “Topic Modeling” in Chapter 5 pp. 234–249	Update your glossary with terms from this week Get a head start on Homework 8 Quiz 9, due the day after the week 13 live session
14	**Semantic Analysis: Sentiment and Rhetoric** General Sentiment Analysis Advanced Sentiment Analysis	Sarkar “Sentiment Analysis” and “Sentiment Analysis of IMDB Movie Reviews” in Chapter 7 pp. 342–376	Update your glossary with terms from this week Homework 8, due the same day as the week 15 live session Quiz 10, due the day after the week 14 live session Final glossary submitted (turn in for grading), due the same day as the week 15 live session
15	**Review and Final** The final project will be introduced week 12		Final Due

	and due week 15.		

Assessments

All assessments will be graded on a standard scale (not a curve).

Quizzes

Quizzes are multiple choice and will be taken online. They may relate to material from the video and/or reading materials. Quizzes will be given most weeks that we are not reviewing for a midterm or final. Each quiz has the same weight for grading purposes as every other quiz regardless of the number of questions in each quiz.

Glossary and Homework Assignments

One of your most important assignments is to write your own glossary definitions of key terminology each week. This will be submitted one week before the midterm, to provide you with comments from the grader, to help guide you. The glossary will be turned in and graded one week before the final exam.

Other assignments may involve manually retrieving and analyzing textual data or writing some Python code to process text. You may be required to provide inputs, outputs, source code, and a written summary of your findings. Some of the assignments will build upon your work in previous assignments, so that sequentially, you will be building up a semester project.

All assignments (other than the glossary) have the same weight for determining the Homework Assignments portion of the final grade.

Midterm and Final Exams

These will comprise open-ended questions, requiring short answer or short essay responses. Most of the questions will involve application or problem-solving scenarios, requiring you to explain one or more problems, solutions, methods, apparatuses, and alternatives. Instructor will provide exam details directly to students.

Assignment Type	**Weight on Final Grade**
Quizzes (multiple choice, online)	20%
Glossary (keep up on a weekly basis)	10%
Homework Assignments (working with data and/or coding)	20%
Midterm (mainly essay questions)	25%
Final (mainly essay questions)	25%

This syllabus is only a guideline and is not a legal contract. The professor of record for the course has final say on any policies, due dates, etc.