IUPUI

SCHOOL OF INFORMATICS AND COMPUTING

DEPARTMENT OF BIOHEALTH INFORMATICS

Indiana University–Purdue University Indianapolis

INFO-B 643 Natural Language Processing and Text Mining for Biomedical Records and Reports

Credit Hours	3
Instructor	Jiaping Zheng
Email	jizhen@iu.edu
Prerequisite	INFO-I 501 & INFO-B 518

Description

This course familiarizes students with the basic analysis and applications of Natural Language Processing and text mining. This course introduces commonly used processes, algorithms, techniques, and software. The assignments and projects will provide hands-on experience with a variety of text mining applications.

Objectives

- Students will describe the principles, formal methods, and components used in the design and analysis of language processing algorithms and text mining techniques for health care applications.
- Students will deploy major NLP algorithms from existing libraries to analyze text, including lexical, morphological, syntactic, and semantic analysis, with the primary focus on parsing algorithms and their analysis.
- Students will design an actual text mining application in which NLP modules contribute to extracting information from text and present their findings to the class orally and in a written report that conforms to academic writing standards.

Prerequisites

INFO-I 501 and INFO-B 518, and programming experiences.

Textbook

The following book is recommended as a reference but not required. Additional reading materials will be posted on Canvas.

- Dipanjan Sarkar, *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*, 2nd ed., Apress, 2019.

Learning Objectives:

Course Objectives	AMIA Functional Domains	Proposed competency driven objectives	Miller's Pyramid(map)	Class activities	Assessment
1. Develop an understanding of the principles and formal methods used in the design and analysis of language processing algorithms and text mining techniques for health care applications.	F4	Students will describe the principles, formal methods, and components used in the design and analysis of language processing algorithms and text mining techniques for health care applications.	Knows How	Lectures	Assignment
2. Provide an in-depth presentation of the major algorithms used in NLP, including lexical, morphological, syntactic, and semantic analysis, with the primary focus on parsing algorithms and their analysis.	F4	Students will deploy major NLP algorithms from existing libraries to analyze text, including lexical, morphological, syntactic, and semantic analysis, with the primary focus on parsing algorithms and their analysis.	Shows How	Lectures, Labs	Assignment, Project Work
3. At the technical level, the course offers hands-on training and experience in building an actual text mining application in which NLP modules contribute to extracting	F4, F8, F9	Students will design an actual text mining application in which NLP modules contribute to extracting information from text and present their findings to the class orally and	Does	Class project	Project Work

information from text.		in a written report that conforms to academic writing standards.			

Grading and Course Evaluation

Grade breakdown

- Homework assignments (40%)
  10 assignments, 4 points each
- Quizzes (20%)
  2 in class quizzes, 10 points each
- Final project (40%)
  Includes a proposal, a report, and a short in-class presentation. Detailed information and grading rubrics are available in the *Project Guidelines* document.

Scale

A+	97 – 100	Outstanding achievement, given at the instructor's discretionA
	93 – 100	Excellent achievement
A-	90 – 92	Very good performance and quality of work
B+	87 – 89	Good performance and quality of work
B	83 – 86	Modestly acceptable performance and quality of work
B-	80 – 82	Marginal acceptable performance and quality of work
C+	77 – 79	Unacceptable work (Core course must be repeated for credit)
C	73 – 79	Unacceptable work (Core course must be repeated for credit)
C-	70 – 72	Unacceptable work (Course must be repeated for credit)
D+	67 – 69	Unacceptable work (Course must be repeated for credit)
D	63 – 66	Unacceptable work (Course must be repeated for credit)
D-	60 – 62	Unacceptable work (Course must be repeated for credit)
F	0 – 60	Unacceptable work (Course must be repeated for credit)

Course Logistics

Canvas, the Indiana University online teaching resource, will be used for this course. Students will be given instructions to use Canvas after they enroll. All communication will be initiated from Canvas, so correctly specifying your email address, and setting up the mail forward feature is critical for communication. Use the Canvas messaging feature to communicate with the instructor so your emails do not get lost. Under rare circumstances, if you need to email the instructor directly, please *include 'B643' in the subject line*.

Course Policy

Homework assignment

All homework assignments will be available shortly after class on Canvas and are due before the next class *electronically in PDF format only* on Canvas. No paper copies or email copies will be accepted, unless there is a Canvas system malfunction or prior arrangements with the instructor are made.

The assignments are to be done by each student individually. Discussions of the questions are allowed, but the discussion may not include specific answers to any of the questions.

Each student is given three grace days in the entire term for late assignments. These grace days can only be used for homework assignments, not for quizzes or the project. It is up to the student to decide whether and when to use these days. However, they can only be used in increments of one day. Therefore, a 5-minute late submission constitutes one grace day usage. If all grace days are used, late homework assignments will not be accepted. Exceptions can only be granted in extreme cases.

Project

The project is a critical component of the course and requires significantly more time commitment than for regular homework assignments. Students can work alone or as a team of no more than two people.

Course Schedule

	Topics	Lab
1	Course introduction and organization; Overview, what and why text mining	Python introduction
2	Applications of NLP (general, clinical, bio surveillance and public health); Terminologies, Coding, and Natural Language Processing	NLTK introduction
3	Linguistic essentials, Morphology, stemming, tokenization	Stemming and tokenization using NLTK
4	Regular expressions and rule-based NLP methods	Regex tutorial
5	Part of speech tagging; Introduction to graphical models (CRF and HMM)	POS tagging using NLTK
6	Named entity recognition	Quiz Q1; NER using NLTK

7	Parsing I	Parsing in NLTK
8	Parsing II	Parsing in NLTK
9	Semantic role labeling	
10	Reference standard; Text annotation; Evaluation metrics	
11	Word sense disambiguation; Negation detection; Coreference resolution	
12	Document clustering; Text classification	Quiz Q2
13	Case Study: Biomedical text analysis	Weka tutorial
14	Case study: EHR text analysis	Project Q&A
15	Project presentations	Project Q&A