IUPUI

SCHOOL OF INFORMATICS AND COMPUTING

DEPARTMENT OF BIOHEALTH INFORMATICS

Indiana University–Purdue University Indianapolis

Natural Language Processing (NLP) and Text Mining for Biomedical Records and Reports– INFO I643

Spring

Course Info	3 Credit hours
Location	Classroom & Online
Prerequisites:	None

COURSE DESCRIPTION

This course familiarizes students with applications of the Natural Language Processing and text mining in health care. While the course provides a short introduction to commonly used algorithm, techniques, and software; it focuses on the existing health care applications including clinical records and narratives, biomedical literature and claims processing.

This course is designed to familiarize students with applications of Natural Language Processing (NLP) and text mining in health care. While the course provides a short introduction to commonly used algorithm, basic information theory, probabilistic and graphical modeling, link analysis, and semi-supervised learning techniques and software, the focus is on existing health care applications such as clinical records and reports, biomedical literature and health care claims processing. Attention also will be given to the development and uses of biomedical ontologies and application specific knowledge bases. Some of topics covered include text categorization, document summarization, sentiment analysis, word sense disambiguation, machine translating, and speech recognition.

Required Text(s):

Chen, H; Fuller, SS; Friedman, C; Hersh W. (eds) (2005); Medical Informatics: Knowledge Management and Data Mining in Biomedicine, Springer, New York, NY.

Meziane, F.; Metais, E. (eds) (2004): Natural Language Processing, Springer, New York, NY.

Course objectives

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.

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.
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 information from text.

EXPECTATIONS, GUIDELINES, AND POLICIES

This is a three-credit, graduate-level course. In accordance with IUPUI policies and expectations, a 3:1 workload is expected: On-average, in addition to 3 hours in-class, this course should take approximately 12 - 15 hours per week. This workload will increase dramatically before assignments are due. This translates to a significant commitment of time each week. A graduate course is the equivalent of a rigorous, part-time job (15+ hours per week). Plan accordingly, pace yourself, and frontload your workflow.

Attendance: Missing class course chat time will affect your grade. Students are allowed two (excused or unexcused) absences before their grade will be effected. In other words, whether you are sick or have personal problems or issues for missing class, it will amount to the same. Missing class means you do not show for the whole or majority of the session. The grade reduction policy works in this way.

a. On the third missed class time your final grade will drop 5 points (regardless of the reason).

b. On the fourth missed class your final grade will drop 10 points (regardless of the reason), and 5 additional points thereafter for each additional class missed.

Deliverables: Responsible for due dates and related materials: All weekly due assignments are the students' responsibility. If class is missed, the student is still responsible for the assignment, as well as to find out what was covered in class, e.g., any new assignments or variations to an existing assignment. ALL assignment deadlines are outlined in the syllabus or syllabus supplemental documents provided on CANVAS. The instructor will only give one reminder of these dates. In the end, each student is responsible for the deadline. Also, weekly assignment deadlines should be adhered to, to insure fairness to all students. For the purpose of maintaining an equal and fair evaluation of each student's work, no student will receive special treatment. As a result, the following rules will apply to this course:

a. All assignments must be ready to hand in or email at the designated time and place as stated on the assignment sheet, as communicated via email, or on the syllabus.

b. All assignments handed in late will be reduced 10 points for every day late (24 hrs. from the due date and time). For example, if the assignment is due at 6PM on the due date and it is post-marked 6:01PM, it will be reduced automatically by 10 points. If the class meets in the class room, students must be ready to hand the assignment in at the start of class time.

c. Incompletes will NOT be issued except under very extreme personal conditions that have been reviewed by the instructor and in some cases in consultation with the Dean's Office.

Topics Outline

- Natural language processing
  - History
  - Content intelligence systems

    - Concept specification language
    - Review of mathematical background
- Statistics of the English language
- Language modeling for application specific knowledge systems
    - information theory
    - A mathematical theory of communication
    - Ontology for biomedical knowledge representation
- Text Mining
    - Information retrieval and link analysis of clinical records and narratives
    - Biomedical Document summarization
    - Bringing order into clinical reports and narratives
    - Text categorization: Naive Bayes, logistic regression
    - Word sense disambiguation: unlabeled clinical reports as knowledge source
    - Information extraction with Conditional Random Fields
    - Parsing and context free grammars
    - Exploring clinical narratives using multiview semi-supervised learning
- NLP and Text Mining application in Health Care
    - Demonstration
    - Laboratory exercises

Grading Information:

Evaluation Forms: Students should review all grading forms that will be used by the instructor to grade projects, presentations, papers, and other assignments. Please see the course web site under the section called "Evaluation Forms." These documents will show you the checklist and criteria by which each class assignment will be evaluated.

If students want to see their grades at any time during the semester, they should contact the Instructor by phone or email.

Score: Criteria to Evaluate Threaded Discussions.

4
- Exceptional quality (not quantity)
- Evident that individual has completed all reading assignments
- Demonstrates applied level of understanding through personal reflections
- Answer is well-developed and logically reasoned
- Provides original insights or responses; extends comments of others
- Supports and leads others in discussion; respects others and their ideas.

3
- Superior quality (not quantity)
- Evident that individual has completed all reading assignments
- Demonstrates applied level of understanding through personal reflections
- Answer is provided; logic may not be clear
- Provides original insights or responses
- Makes connections to what others say; respects others and their ideas

2
- Satisfactory quality and quantity
- Evident that individual has completed all reading assignments
- Primarily consists of summary or paraphrasing of readings
- Answer is not fully developed; logic is not clear

	· Contribution is primarily a response to others; minimal originality · Is respectful of others and their ideas
1	· Does not meet expectations · Not clear that individual has completed reading assignments · Only consists of summary or paraphrasing of readings · Minimal effort put into answer · Is not respectful of others and their ideas
0	· Assignment not complete

Grading Scale:

A+	97 – 100	Outstanding achievement, given at the instructor’s discretion
A	93 – 100	Excellent achievement
A–	90 – 092.99	Very good work
B+	87 – 089.99	Good work
B	83 – 086.99	Marginal work
B–	80 – 082.99	Very marginal work
C+	77 – 079.99	Unacceptable work (Core course must be repeated)
C	73 – 076.99	Unacceptable work (Core course must be repeated)
C–	70 – 072.99	Unacceptable work (Elective or core course must be repeated)
D+	67 – 069.99	Unacceptable work (Elective or core course must be repeated)
D	63 – 066.99	Unacceptable work (Elective or core course must be repeated)
D–	60 – 062.99	Unacceptable work (Elective or core course must be repeated)
F	Below 60	Unacceptable work (Elective or core course must be repeated)