# Course Syllabus

Jump to Today

**LHS 712 – Natural Language Processing for Health**

**SYLLABUS**

**Class #: 32394**

**Instructor: V. G. Vinod Vydiswaran (vgvinodv@umich.edu)**

**Meeting schedule: Thursdays, 1:00 – 4:00pm, 2813/2817 Medical Sciences-II Building**

**Dates: Jan 5th, 2017 to April 13th, 2017**

**Office Hour: Thursdays, 4:00 – 4:55 pm, 1161F North Ingalls Building (starting Jan 05)**

Students in this course will learn advanced techniques to parse and collate information from text-rich health documents such as electronic health records, clinical notes, and peer-reviewed medical literature. In this elective, students will be able to delve deeper into challenges in recognizing medical entities in text documents, extracting clinical information, addressing ambiguity and polysemy, and building searchable interfaces to efficiently and effectively query and retrieve relevant patient data. Students will develop tools and techniques to analyze new genres of health information, and build resources to help in these tasks. Students will also participate in a semester-long project on addressing specific natural language processing challenges in real-life health data sets.

## A. Learning objectives

By completion of the course, students will be able to:

1. Describe the role of medical natural language processing in improving health and healthcare.
2. Identify the major natural language processing challenges in health data.
3. Develop skills to process and extract information from health-related free-text data.
4. Apply the state-of-the-art machine-learning techniques to extract information from medical text data.
5. Analyze and critique natural language processing tools currently available for medical text processing.
6. Explore the recent trends and open directions in the field of medical natural language processing.

## B. Course Format and Grading

This course will be taught using multiple methods, including instructor and student-led discussions, in-class exercises, programming and reading assignments, and a longitudinal project. The instructor will give tutorials and lead discussions to introduce the basic principles and tasks of natural language processing on health data. After the first two weeks, during every week for the rest of the semester, the instructor will start with a tutorial about the methods, followed by a student-led tutorial on applications of the methods. Depending on the background of the cohort, the instructor will decide whether to give more tutorials about the methodology of mining particular genres of data, or let students with the right background lead the tutorials.

### B.1. Grading

Grading will be based on:

- Reading and critiquing peer-reviewed papers (15%)
- Student-led presentation and discussion (10%)
- Mid-term examination (20%)
- Assignments based on NLP tools (25%)
- Semester-long course project (30%)

#### B.1.1. One-page summaries (15%):

Each week, starting week 2, students are expected to write a one-page summary based on the reading assignment for the previous week. The reading assignments will cover significant papers on the topics being discussed in class. The one-page summaries are expected to discuss the key concepts described in the paper, rather than merely stating what the paper is about. A summary of key contributions, potential limitations, suggested improvements, and ideas for future follow-up work based on the paper are encouraged.

All summaries will be read by the instructor, but not graded or returned to the students. Students are expected to hand-in at least ten summaries over the semester.

#### B.1.2. Student-led presentation and discussion (10%):

Students are expected to lead one tutorial on the applications of the methods discussed in class. The students will lead the discussion both during and after their presentation on the topics assigned to them. In preparation of their presentation, students will be required to survey the state-of-the-art techniques from major conferences and journals for recent developments and applications of these methods. The discussion will focus on how to apply the methods to solve particular problems and build various applications. Each student will be in charge of at least one topic, depending on the size of the cohort. Students who are not presenting or leading the discussion will be required to actively participate in discussion and write a short survey on the assigned topic.

#### B.1.3. Mid-term exam (20%):

The mid-term exam will be a take-home test and will be administered around Week 8 of the semester (early March 2017).

#### B.1.4. Assignments (25%):

There will be 3-4 assignments during the semester based on specific health text processing tasks. The tasks will be closely related to the course material, with real-world data and gold-standard judgments provided. This may include an in-situ data mining challenge using online competition services such as Kaggle-in-Class (http://inclass.kaggle.com). Students can submit and resubmit their results to the competition site and get instant feedback (evaluation metrics) from the system. The task will likely to be selected from one of the follows: severity identification, disease mention detection, forum post classification, etc.

#### B.1.5. Course Project (30%):

A course project is required. Individual projects are preferred. Small group projects are acceptable upon justification. The grading of group members will be adjusted according to their contribution to the project.

The course project will take the format of either a software system that applies existing data mining techniques to a specific type of data, or a research experiment documented in the form of a research paper.

Examples of course projects include:

1. A de-identification tool for health records using conditional random fields
2. Retrieving information about relevant clinical trials for a given case
3. Comparing authorship networks and communities in different clinical specialties
4. Identifying high-quality consumer-centric resources

The grading for the course project will be split as follows:

1. Proposal (15%): A two-page proposal, describing the project topic, objectives, expected deliverable (software package, demo, and/or a technical report), and a list of team members and their expected contribution to the project.
   Tentative deadline: Around Week 5 (early-February)
2. Progress report (10%): A one-page summary of the progress, any hurdles towards timely completion of the stated objectives. If there are any significant changes to the submitted proposal, the students should describe them in detail in the progress report. Consider this as a checkpoint towards achieving the stated goals of the project. There are no penalties for changes to the proposal document, rather it may be more prudent to recalibrate or clarify the expected outcomes during this stage.
   Tentative deadline: Around Week 10 (mid-March)
3. Project Presentation (25%): Students will give a short presentation to showcase their project in class. The focus of this presentation is to demonstrate and describe what was done, report interesting observations, present key conclusions, and discuss potential limitations of the study. Students working in teams may choose to present as a group or elect one of the team members to present on their behalf. Students will not be penalized for choosing not to present individually, as long as the project itself is showcased.
   Tentative schedule: Last lecture of the course (Thursday, April 13, 2017)
4. Final project deliverable and report (50%): Students are expected to submit their project deliverable, along with a brief report. The report should include the key observations and conclusions based on the project and suggest potential follow-up studies. Teams working on the project together must also describe individual contributions of the team members.
   Tentative deadline: Thursday of Exam week (Thursday, April 20, 2017)

## C. Policies

### C.1. Late submission policy

Students have 72 hours of buffer grace period for the entire semester. If necessary, students may use it to submit any of the assignments, homework, or the course project reports late without any effect on the overall grade. The grace period, however, cannot be used to submit the exams or quizzes late. A student may use it all on one assignment or use a bit of it for any number of assignments. Once the buffer grace period is used up, late submissions will not be graded.

### C.2. Academic Conduct

#### C.2.1. Collaboration

The Department of Learning Health Sciences and the instructor strongly encourage collaboration while working on some assignments, such as homework problems and interpreting reading assignments as a general practice. Active learning is effective. Collaboration with other students in the course will be especially valuable in summarizing the reading materials and picking out the key concepts. You must, however, write your homework submission on your own, in your own words, before turning it in. If you worked with someone on the homework before writing it, you must list any and all collaborators on your written submission. Read the instructions carefully and request clarification about collaboration when in doubt. Collaboration is almost always forbidden for take-home and in class exams.

**C.2.2. Plagiarism**

All written submissions must be your own, original work. Original work for narrative questions is not mere paraphrasing of someone else's completed answer: you must not share written answers with each other at all. At most, you should be working from notes you took while participating in a study session. You may incorporate selected excerpts from publications by other authors, but they must be clearly marked as quotations and must be attributed. If you build on the ideas of prior authors, you must cite their work. You may obtain copy-editing assistance, and you may discuss your ideas with others, but all substantive writing and ideas must be your own, or be explicitly attributed to another. Please refer to the Rackham's Graduate School Academic for the definition of plagiarism, cheating, and other academic misconduct; the consequences for intentional or unintentional plagiarism; and resources to help you avoid it. The policy handbook is available here: http://www.rackham.umich.edu/current-students/policies/academic-policies (Links to an external site)

**C.3. Reasonable accommodations**

The university will provide reasonable accommodations to qualified individuals with disabilities upon request. If you think you need an accommodation for a disability, please let the instructor know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations. SSD (734-763-3000; http://www.umich.edu/sswd/) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. I will treat any information that you provide in as confidential a manner as possible. For more information, see https://ssd.umich.edu/article/americans-disabilities-act-ada (Links to an external site)

It is also the University's policy that every reasonable effort be made to help students avoid negative academic consequences when their religious obligations conflict with academic requirements. Students who expect to miss classes, examinations, or other assignments as a consequence of their religious observance are requested to contact the instructor by the drop/add deadline. For more information see https://www.provost.umich.edu/calendar/religious_holidays.html (Links to an external site)

**D. Tentative Schedule**

Please check this page periodically for a more detailed, accurate, and up-to-date schedule and reading list.

**Week 1: Introduction (Jan 05)**

- Why Medical Natural Language Processing?
- Challenges of Big Data in Health

A. NLP Essentials

**Week 2: Dealing with words (Jan 12)**

- tokenization, normalization
- word sense disambiguation
- ngrams
- statistical NLP
- Tools: NLTK

**Week 3: Processing sentences and corpora (Jan 19)**

- sentence boundary, syntax
- part of speech tagging
- negation detection and hedging
- Regular expression
- Tools: NegEx

B. NLP Tasks and Techniques

**Week 4: Text classification (Jan 26)**

- Decision trees
- Support Vector Machines
- Naïve Bayes
- Tools: Weka

**Week 5: Information extraction (Feb 02)**

- Hidden Markov Models
- Conditional Random Fields
- Tools: cTAKES, …

**Week 6: De-identification (Feb 09)**

- Named entity recognition
- Protected health information
- De-identification
- Tools: MetaMap, MIST

**Week 7: Information retrieval (Feb 16)**

- Vector space models
- Probabilistic models
- term weighting (tf-idf)
- index construction
- ranking retrieved results
- Tools: EMERSE, TREC Clinical Decision Support Task

**Week 8: Question answering (Feb 23)**

- question classification
- query construction
- passage retrieval
- answer extraction and ranking
- Tools: Watson Health

**Week 9: Advanced topics (Mar 09)**

- Summarization
- Sentiment analysis
- Challenges due to acronyms (polysemy, synonymy)
- Deep learning

C. Medical NLP Resources

**Week 10: Medical ontologies (Mar 16)**

- UMLS
- ICD code
- SnoMed

**Week 11: Medical NLP systems (Mar 23)**

- MetaMap
- MedLEE
- cTAKES

**Week 12: Datasets and Shared Challenges (Mar 30)**

- i2b2 and MIMIC
- NHANES and NAMCS
- Healthdata.gov
- TREC-Clinical Decision Support

**Week 13: Research directions in Medical NLP and Biomedical informatics (Apr 06)**

**Week 14: Project presentations (Apr 13)**

**E. Suggested Readings**

The readings of this course will be selected from the recent literature in major journals and conference proceedings in the field of medical informatics. They include, but are not limited to, the Journal of American Medical Informatics Association (JAMIA), the Journal of Biomedical Informatics (JBI), the Journal of Medical Internet Research (JMIR), Bioinformatics, and conferences such as the Annual Meeting of the American Medical Informatics Association (AMIA). Some relevant papers published in the Computer Science venues that describe relevant methodologies for natural language tasks will also be selected. Such venues include the Association of Computational Linguistics (ACL), Empirical Methods of Natural Language Processing (EMNLP), and the

Association for the Advancement of Artificial Intelligence (AAAI). It is also encouraged that students review and suggest relevant literature to add to the reading list.

### E.1. Optional Textbook

The following is an optional textbook that could be used for supplemental reading.

1. Kevin B Cohen and Dina Demner-Fushman. Biomedical Natural Language Processing.
   This book has a good introduction to various biomedical natural language processing tasks for those with a working knowledge of NLP.

### F. Administrative notes

1. Regular class schedule: Thursdays, 1:00pm to 4:00pm, starting Jan 05
   Note: The classes will run on Michigan time (start 10 minutes past the scheduled time).
2. Classroom: 2813/2817 Medical Sciences-II Building
   For directions, see https://campusinfo.umich.edu/campusmap/102(Links to an external site.)
3. Office hours: Thursdays, 4pm to 4:55pm, 1161F North Ingalls Building, starting Jan 05.
4. Course website: We'll be using Canvas for this course. https://ctools.umich.edu(Links to an external site)
5. Instructor:
   V.G.Vinod Vydiswaran, Ph.D.
   Assistant Professor, Department of Learning Health Sciences, Medical School, University of Michigan
   Assistant Professor (courtesy), School of Information, University of Michigan
   1161F - NIB, 300 N. Ingalls Street, Ann Arbor, MI 48109
   (734) 763 - 0080
   vgvinodv@umich.edu(preferred mode to reach the instructor). Note: Please begin the subject line with [LHS 712].