COSC 6336 Natural Language Processing

Spring 2018 Syllabus

Instructor: Thamar Solorio, tsolorio@uh.edu
Office hours: F 5:00-6:00pm, or by appointment, in PGH 584.

TA: Gustavo Aguilar, gustavoaguilar91@gmail.com
Office hours: TTH- 3:00-4:00pm in PGH 550A.

Lecture sections: W 4:00-7:00pm, Room M 106

Class web site: **piazza.com/uh/spring2018/cosc6336/home**

Course Description:

This is a graduate level introductory course to natural language processing (NLP). The course is intended to develop foundations in NLP and text mining. The broader goal is to understand how NLP tasks are carried out in the real world (e.g., Web) and how to build tools for solving practical language processing problems. Throughout the course, large emphasis will be placed on tying NLP techniques to specific real-world applications through hands-on experience. The course is standalone and covers required topics of machine learning and mathematical foundations.

PREREQUISITES:

1. Algorithms and Data Structure (COSC 3320) or equivalent

2. Sufficient programming experience (in C++/Java/Python, etc.) for building projects.

Textbooks:
The official book is the 3rd Edition Book from Jurafsky and Martin:
http://web.stanford.edu/~jurafsky/slp3/

The missing chapters will be based on the previous edition:
SPEECH and LANGUAGE PROCESSING, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition, by Daniel Jurafsky and James H. Martin, Prentice Hall, 2008.

Natural Language Processing in Python, NLTK. This book has open access and will help you get started on your programming assignments. I refer to this book as NLTK in all course materials.

Policies and Other Information
Students are required to check the email provided for contact and visit the course website on piazza on a daily basis. Official announcements for the class will be made through these channels.

Grading:
Final grades will be based on a combination of practical assignments, quizzes, paper presentations and in class participation. The approximate percentages are as follows:

- 40% - Assignments (mini projects, 3-4 total)
- 40% - Exams (Tentative dates: 1st exam: 3/7/18, second exam: 5/9/18)
- 20% - In-class participation, exercises, and quizzes

Class participation will include engagement in class discussions and on piazza.

Additionally, any one of the following will result on a final grade of F, even if the overall average is greater than 60%.

- Obtaining an average of less than 60% on the projects
- Obtaining an average of less than 60% on the exams
- Missing more than four lectures

The nominal percentage-score-to-letter-grade conversion is as follows:

- 90% or higher is an A
- 80-89% is a B
- 70-79% is a C
- 60-69% is a D
- below 60% is an F

We reserve the right to adjust these criteria downward, e.g., so that 88% or higher represents an A, based on overall class performance. The criteria will not be adjusted upward, however.

Late assignments:
All assignments up to one day late will receive up to 80% percent of full credit, and more than one day late will receive no credit. All class assignments are due at the beginning of the class period.

Collaboration:
Discussion of homework among students is encouraged, but your answers and your code should be written and tested by you alone. **Do not exchange programs or let someone look at your code or solutions, even if "just so they can see how you did it."** If you need help, consult the instructor or the TA.

Standards of Conduct and Academic Dishonesty:
You are expected to conduct yourself in a professional and courteous manner, as prescribed by the UH Student Code of Conduct. Academic dishonesty includes but is not limited to abetting, cheating, plagiarism, fabrication and misrepresentation. *Abetting* involves collaborating with another person to commit an academically dishonest act, for instance allowing another student to

copy your homework or present your work as their own. *Cheating* may involve copying from another student, or possessing unauthorized materials during a test. *Plagiarism* occurs when someone represents the work or ideas of another person as his/her own. *Fabrication* is the act of presenting falsified data as genuine. Examples of *misrepresentation* include falsifying data (for example program outputs) in laboratory reports or projects.

Any violation of the UH Student Code of Conduct will result in a grade of 0 for the given assignment and possible failure of the course and a report to the Dean of Students.

For the dedicated students

I love it when students read the complete syllabus to understand the standards and grading criteria of the course. If you have made it this far, congratulations! Email me a nice landscape picture you like (it doesn't have to be somewhere you have been, but it's ok if you have been there too) by January 24th and receive two extra credit points added to you final grade. Please do not tell any current student about this, let them discover it on their own. That way this continues to be a reward for those who do the work of reading the syllabus.

Disabilities: If you feel that you may have a disability that requires accommodation, contact the Center for students with disabilities at (713)- 743-5400, or email: uhcsd@central.uh.edu

Major Topics Covered:

- Linguistics Background & Text Processing
- Language models
- Vector Semantics
- Hidden Markov Models
- Sequence Labelling and POS tagging
- Syntactic Parsing
- Higher Level NLP tasks: Information Extraction, Question Answering, Dialogue Systems