[DOWNLOAD A USC IDENTITY GUIDELINES COMPLIANT LOGOTYPE, THEN DELETE THIS TEXT AND DRAG/DROP THE FILE HERE]

Course ID and Title "Introduction to Information Extraction"

Units: 4.0

Term—Day—Time: Fall18—Tue/Thu—2:00-3:50pm

IMPORTANT:

The general formula for contact hours is as follows:
Courses must meet for a minimum of one 50-minute session per unit per week over a 15-week semester. Standard fall and spring sessions (001) require a final summative experience during the University scheduled final exam day and time.

(Please refer to the *Contact Hours Reference* guide.)

Location: TBD; (website will be up at http://www-bcf.usc.edu/~xiangren/cs699_IE_fall2018.html)

Instructor: Xiang Ren

Office: SAL 308
Office Hours: Tue 4-5pm (or by appointment)
Contact Info: xiangren@usc.edu, 213-821-4067, Timeline for replying to emails/calls: within 48 hours

Teaching Assistant:

Office: TBD
Office Hours: TBD
Contact Info: TBD

IT Help: N/A
Hours of Service:
Contact Info: Email, phone number (office, cell), Skype, etc.

Course Description

In today's computerized and information-based society, people are inundated with vast amounts of text data, ranging from news articles, social media posts, scientific publications, to a wide range of textual information from various vertical domains (e.g., corporate reports, advertisements, legal acts, medical reports). How to turn such massive and unstructured text data into structured, actionable knowledge, and how to enable effective and user-friendly access to such knowledge is a grand challenge to the research community. This course will introduce and discuss many of the sub-problems and methods of information extraction, including use of textual patterns, language and formatting features, generative and conditional models, rule-learning and deep learning techniques. We will discuss segmentation of text streams, classification of segments into fields, association of fields into records, and clustering and de-duplication of records.

Learning Objectives

At a high-level, through this course students will have a concrete idea of what information extraction is about, what the state-of-the-art is, and what the open problems are. Along the way, students will explore many of the mainstays of statistical modeling, including maximum likelihood, expectation maximization, maximum entropy methods, discriminative training, mixture models, and semi-supervised training methods, as well as deep learning models such as recurrent neural networks, convolutional neural networks, and neural sequence labeling. The hope is that by the end of this course students will have in-depth understanding about information extraction tasks as well as the related machine learning models, and can develop practical algorithms and implement systems for solving information extraction tasks.

Prerequisite(s): Familiarity with probability, natural language processing, and algorithms.

Co-Requisite(s): N/A

Concurrent Enrollment: N/A

Recommended Preparation: sufficient mathematical background; general background on machine learning and optimization; good programming skills

Course Notes

Lecture notes will be available online after each class.

Technological Proficiency and Hardware/Software Required

N/A

Required Readings and Supplementary Materials

- *Speech and Language Processing*, Daniel Jurafsky and James Martin, Prentice-Hall (second edition).
- Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press.
- Christopher M. Bishop. *Pattern Recognition and Machine Learning*.
- Trevor Hastie, Robert Tibshirani and Jerome Friedman. The Elements of Statistical Learning.
- Convex Optimization (Boyd)
- Deep Learning (Goodfellow & Bengio)
- Deep Neural Methods for NLP (Goldberg)

Description and Assessment of Assignments

What kind of work is to be done and how should it be completed, i.e. how the learning outcome will be assessed. Include any assessment and grading rubrics to be used.

Grading Breakdown

Including the above detailed assignments, how will students be graded overall? Participation should be no more than 15%, unless justified for a higher amount. All must total 100%.

Assigments	Points	% of Grade
HW1	100	10%
HW2	100	10%
HW3	100	10%
HW4	100	10%
Project survey paper	100	10%
Project	100	45%
Participation	15	5%

Grading Scale

A	95-100
A-	90-94
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
D+	67-69
D	63-66
D-	60-62
F	59 and below

Assignment Rubrics

N/A

Assignment Submission Policy

By email, before 11:59pm of the due date.

Grading Timeline

Assignments will be graded within one week after the due date.

Additional Policies

Late homework policy: you are given 4 late days for the assignments and project proposal/survey (no late days for the final project), to be used in integer amounts and distributed as you see fit. Additional late days will each result in a deduction of 10% of the grade of the corresponding assignment.

Course Schedule: A Weekly Breakdown

	Topics/Daily Activities	Readings and Homework	Deliverable/ Due Dates
Week 1	**Class Introduction and Outline.** - Self-introductions - IE overview slides **Various Information Extraction Settings** - source of data; amount of labeled data; closed/open-domain		
Week 2	**Named entity recognition I** - Different problem formulation for NER: pattern boostrapping, sequence labeling, etc. - Sequence labeling: HMM, MEC, CRF	Project team-up, prepare proposal/survey	
Week 3	**Named entity recognition II** - semi-supervised models - pattern-based methods - co-training, boostrapping - distant supervision	HW1	HW1 due by end of week 5
Week 4	**Entity typing** **Neural sequence models I** - Recurrent neural network for sequence tagging - Convolutional neural nets - Character-level models, subword methods		
Week 5	**Neural sequence models II** - subword methods **Parsing** - dependency parsing - semantic parsing		HW1 due
Week 6	**Entity linking** **Co-reference resolution**	HW2	HW1 due by end of week 8
Week 7	**Relation Extraction I** - rule-based methods - pattern bootstrapping - feature-based classifiers		
Week 8	**Relation Extraction II** - neural models - distantly-supervised methods		HW2 due
Week 9	**Ontology construction** - synonym/hypernym detection - ontology organization algorithms **Open-domain information extraction**	HW3	Project proposal due; HW3 due by end of Week 11
Week 10	**Event extractin** **Summarization I**		
Week 11	**Summarization II**		HW3 due

	IE from Semi-structured Data (guest lecture)		
Week 12	**Low-resources IE (guest lecture)** **Multi-media Information Extraction**		Project survey due
Week 13	**Knowledge graph and IE applications**	HW4	HW4 due by end of Week 15
Week 14	**Projection Presentation**		
Week 15	**Projection Presentation**		HW4 due
FINAL	**Projection Presentation and wrap up**		Project report due