CS 406-01: Unstructured Information Processing
Monsoon '19-'20

Class Hours: M, W – 11:50 to 13:20 (AC 02, TR 004)

Instruction:	Dr. Ravi Kothari (ravi.kothari@ashoka.edu.in)
Office Hours:	M, W - 15:00 to 16:00 (AC-03, Room 316)
TA:	Bhavesh Neekhra (bhavesh.neekhra_ga@ashoka.edu.in)
Office Hours:	Tu, Th - 18:30 – 19:30, Fr - 18:30 - 20:30

1 Introduction

There is broad agreement that an overwhelming majority of data (> 80%) is unstructured in nature. Such data does not follow an underlying data model and/or is not organized in a pre-defined manner making it difficult to interpret. Text, images, and videos are examples of unstructured data. This course focuses to a large extent on the analysis of textual data and to a slightly lesser extent on the analysis of images. It is anticipated that you will be able to analyze and put to use a significant portion of the (> 80%) data on completion of this course.

2 Prerequisites

Ability to program in a high level language e.g. Java, Python, etc.; Introductory Probability and Calculus.

3 Required Reading

- Class notes and handouts. A copy of the notes will be on the course web page. *Do not use notes from previous years as I have revised them considerably.* The new notes are dated June 1, 2019 or later.

4 Suggested Reading

- D. Jurafsky, and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2017.
- C. D. Manning, and D. Schutze, *Foundations of Statistical Natural Language Processsing*, MIT Press, 1999.

5 Topics and Schedule

Date	Topic	Sub-Topic	Deadline/Remarks
Aug. 26	Natural Language Processing		
Aug. 26		Introduction	
Aug. 28		Text Understanding Using RE	HW # 1 given
Sep. 4	Language Models		
Sep. 4		n-grams	
Sep. 9		Word Embedding	HW # 2 given
Sep. 11	Statistical Inference		
Sep. 11		MLE	
Sep. 16		Naive Bayes	
Sep. 18		Logistic Regression	HW # 3 given
Sep. 23		Hidden Markov Models	
Sep. 25		Hidden Markov Models (contd.)	
Sep. 30	Test # 1		Up to HMMs
Oct. 14	Parts-of-Speech Tagging		
Oct. 16	Distribution Models		
Oct. 16		PMI, TF-IDF	
Oct. 21		TF-IDF (contd.), Cosine Similarity	HW # 4 given
Oct. 23	Word Senses		
Oct. 23		Disambiguation	
Oct. 30	Information Extraction		
Oct. 30		Named Entity Recognition	
Nov. 4		Named Entity Recognition (contd.)	HW # 5 given

Nov. 6	Project List Given		Select by Nov. 10
Nov. 6	Test # 2		Up to NER
Nov. 11	Image Processing		
Nov. 13		Low and Mid-Level Image Descriptors (SIFT)	
Nov. 18		Low and Mid-Level Image Descriptors (HOG)	
Nov. 20	Identifying Objects in an Image		
Nov. 25	Caption Generation		
Nov. 27	Content Based Image Retrieval		
Dec. 1	Final Project Reports Due		No extensions

6 Grading

Percentage in parentheses indicate the contribution to the final score used to determine grade in the class.

- **Home-Work (30%):** Home-work will be assigned as indicated in the previous section and is due by midnight (IST) on the day it is due. Late home-work carries a penalty of 50%/day. Home-work may involve building a system, constructing proofs, thought experiments, reading/presenting (in class)/critiquing a paper, and other such activities

- **Test 1 (25%):** Date given in the previous section

- **Test 2 (25%):** Date given in the previous section

- **Project (20%):** A set of candidate topics will be provided (feel free to propose and discuss any specific ideas you wish to pursue) and you will work in groups to develop and implement the project you choose. Project reports are due by midnight (IST) on the day it is due. No extensions.

- No makeup examinations unless it is truly an exceptional circumstance that is supported by documentary evidence

- Use of any unfair means or copying will result in an F for the course for everyone involved (the individual(s) who copied and the individual(s) who allowed the copying to occur). Please do not do it.