# Special Topics in Natural Language Processing

## CS698O

Ashutosh Modi
CSE Department, IIT Kanpur


Lecture 1: Introduction and Logistics
Jan 3, 2020

# Communicating with Machines

- One of the holy grails for A.I.
  - Seamless communication with machines

# Communicating with Machines

- One of the holy grails for A.I.
  - Seamless communication with machines
- One of the main mode of communication with machines/computers has been via computer programs.

# Communicating with Machines

- One of the holy grails for A.I.
  - Seamless communication with machines
- One of the main mode of communication with machines/computers has been via computer programs.
- However, it unnatural for humans to communicate via computer programs.

Image: Source unknown (licensed under cc)

# Communicating with Machines

- One of the holy grails for A.I.
  - Seamless communication with machines
- One of the main mode of communication with machines/computers has been via computer programs.
- However, it unnatural for humans to communicate via computer programs.
- It would be desirable if humans could communicate with machines via the language that they speak/write i.e. ***Natural Language***.

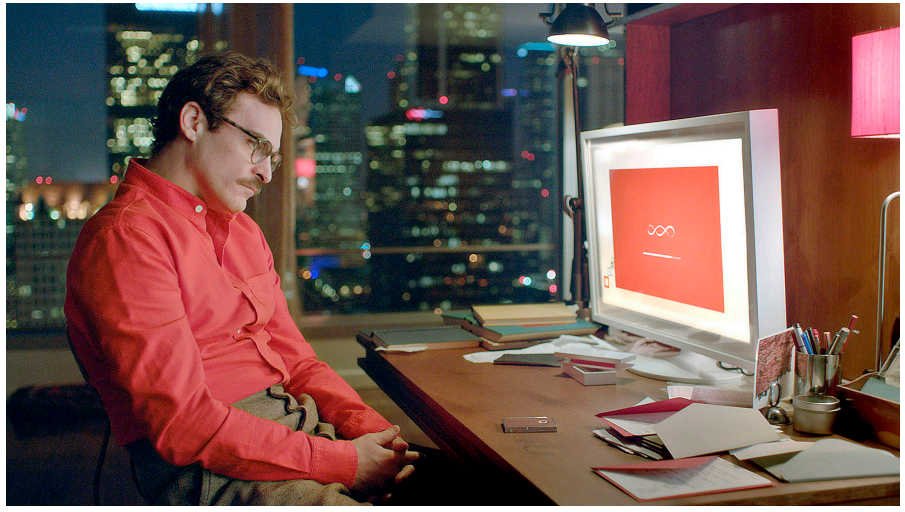


Image: Movie “Her”

# Natural Language Processing

- Natural language processing refers to computational techniques which aim to process, analyze and understand languages spoken/written by humans.

# Natural Language Processing

- Natural language processing refers to computational techniques which aim to process, analyze and understand languages spoken/written by humans.
- These computational techniques aim to make communication between humans and machines as natural as that between humans.
- Communication involves understanding as well as generation on the part of computers.

# Applications

- Personal Assistants: Siri, Alexa, etc.
- Search Engines: Google, Bing, DuckDuckGo, etc.
- Translation: Google translate, etc.
- Social Media Analysis: Facebook, Twitter, etc.
- Economics and Finance
- Automatic Text Generation for News Reports

:

:

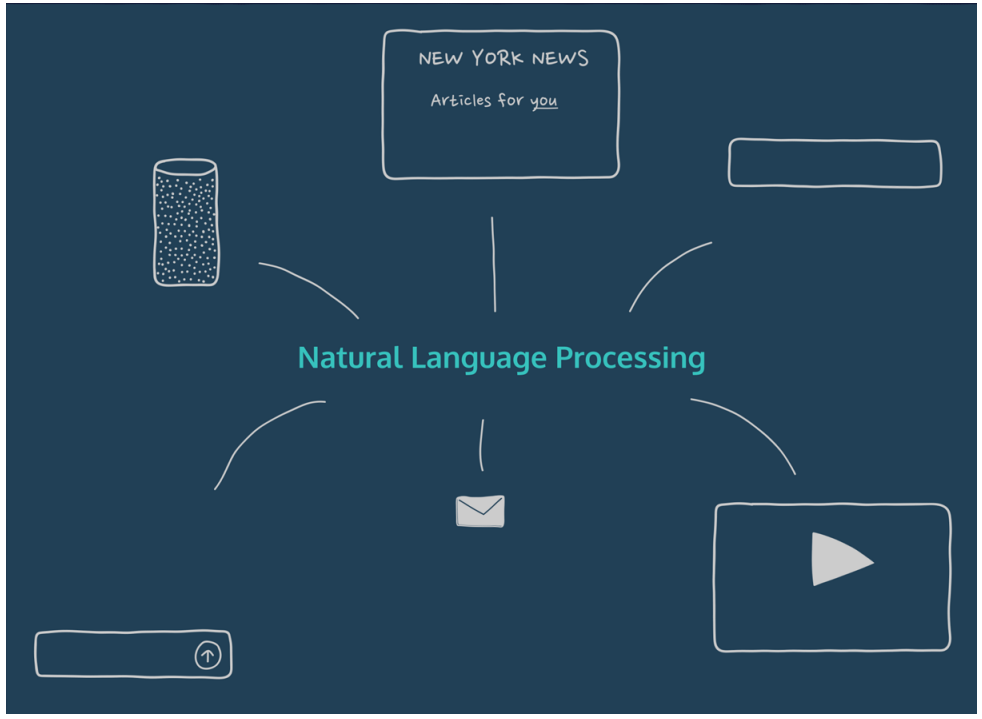



Image: https://towardsdatascience.com/a-gentle-introduction-to-natural-language-processing-e716ed3c0863

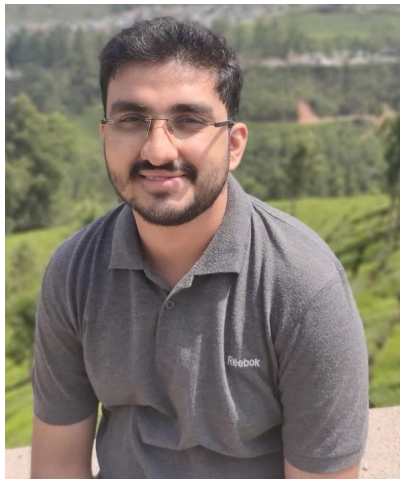
# Welcome to CS698O

- Introduction to Natural Language?
  - Why is NLP hard?
  - Linguistics Fundamentals
- Different levels of meaning in Language? (Lexical, Syntactic, Semantics)
  - Language Models, Parsing
- Classical Structured Prediction Problems in NLP
  - POS tagging, NER, Coreference
- Distributional Semantics
  - Distributional Hypothesis, Vector Space Models
- Deep Learning
  - Word Embeddings, Sequence Models, Transformers

# Pre-Requisites

- Intro to Machine Learning (CS771)
- Fluency in Python Programming
- Linear Algebra and Probability

# Teaching Assistants for the Course



**Avideep**

**Samik**

**Munender**

**Naman**

**Nitin**

# Course Logistics

- Fundamentals of NLP ( ~ 40 Lectures )
  - Classical NLP
  - Deep Learning Based NLP

- 4-5 Assignments : Mix of Implementation + Theory
  - Theory part solutions should be only Latex Typeset. No Handwritten Assignments!!
  - Implementation part solutions -> assignment specific.

- There will be in class Quizzes (Graded and Ungraded)

- Project Focused Course

# Course Logistics

- Join Piazza: www.piazza.com/indian_institute_of_technology_kanpur/winter2020/cs698o
- Access Code: cs698o$
- Discussions and Announcements on Piazza
- Couse Email: nlp.course.iitk@gmail.com

- **Absolutely NO PERSONAL EMAIL !!**

- Office Hours on Piazza

# Libraries, Platforms

- PyTorch: https://pytorch.org/
- Google CoLab : *https://colab.research.google.com*
- Spacy: https://spacy.io/
- Stanford CoreNLP: https://stanfordnlp.github.io/CoreNLP/
- Writing code for NLP research: https://tinyurl.com/rw9c2n5

# Project

- Research Project Focused Course
- Project in a group of 3 but each member must work on separate part/component
- Topic decided by you after consultation with the Instructor
- Preferably in NLP domain! (Can possibly combine with other modalities)

# Project

- Take the project seriously! It is a worthy investment!
- Project must address a research problem/question not solved before
- It is expected that the quality of the project is such that it can be published at the very least in a workshop in a top-tier conference.
- If the project work is good, we will try our best to get it published!

# Project

- Decide the topic within 1 week and submit a project description (max 1 page)
- Project description should describe:
  - What problem do you want to solve?
  - How will you do it?
  - Goals and timeline

# Project

- Project Spread-Sheet Link : https://tinyurl.com/vspkmhv

- SS will be populated with some prospective topics

# Project Ideas

- For some ideas: Check out SemEval Shared task challenges: https://tinyurl.com/uy34csw
- Project could also be about improving an existing published work
- Check out shared tasks in top-tier conferences
- Research problem addressed by the project can be theoretical or applied

# Project Evaluation

- Creativity/Novelty
- Approach
- Technical Soundness
- Project Paper and Presentation
- There will be a competition among the projects! Best projects will have chance to continue it further for a research paper or possibility for external internships!

# Scribe Cheat-sheets

- Idea is to create cheat sheets for each topic (e.g. https://github.com/ml874/Data-Science-Cheatsheet)
- Each group will be assigned a topic (covered in one or more lectures). Each member must contribute.
- Use Latex Typesetting: https://github.com/ml874/Data-Science-Cheatsheet
- At the end of the course, you will have good repo of quick notes

# Cheating and Plagiarism

- Very Strict about Plagiarism
- It is ok to use/modify ideas/code/implementation from others **but Please cite them**
- Cite any source that you use
- Failure to cite can have very serious consequences! **No Compromises!!**
- Cheating will be taken very seriously. **No Compromises!!**

# Cheating and Plagiarism

- Idea is to learn and not just get marks!
- If you learn it well, marks will come automatically.
- In the future, in the real world, you will not have the opportunity to cheat!

# Rough Distribution

- 10% for scribe notes
- 10% for quizzes
- 20% for assignments
- 60% for the project

# Quiz

https://tinyurl.com/u23o847

- Next week:
  - Why is NLP hard?
  - NLP tools of trade
  - Language Models


Image: flickr.com