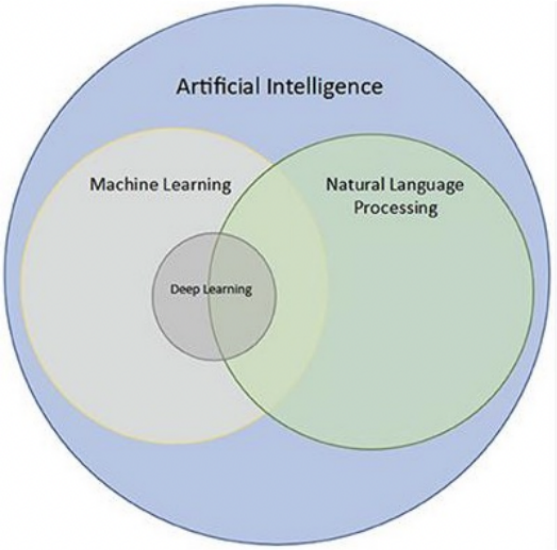
CAS CS 505
Lec 1

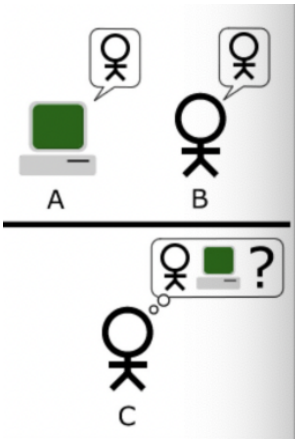
What is Natural Language Processing (NLP)?

1. NLP and Artificial Intelligence
    a. Subfield of AI that focuses on the interaction between computers and humans through natural language
    b. The ultimate objective of NLP is to enable computers to understand, interpret, and generate human language in a way that is both meaningful and useful
    c. NLP has always been a core component of AI and in the last decade has made a huge leap due to advances in Machine Learning, and especially, Deep Learning (Artificial Neural Networks)
    d. 

    e. But since language is such a core component of all human activity, NLP has relationships with a large number of other subfields of mathematics and computer science
2. NLP in AI
    a. In fact, the very first effort to define the notion of AI, and to provide a test for when an algorithm can be considered “intelligent” was provided by Alan Turing, with the famous “Turing Test”
    b. There are two "entities" A and B behind a wall, one a computer and one a person; the human interrogator C asks questions (by typing text) of each, not knowing which is the computer. If after a reasonable time, C can not figure out which is the human, then the machine may be considered intelligent.

3. Turing Test
    a. Turing gave several examples of the kind of the conversations that might take place

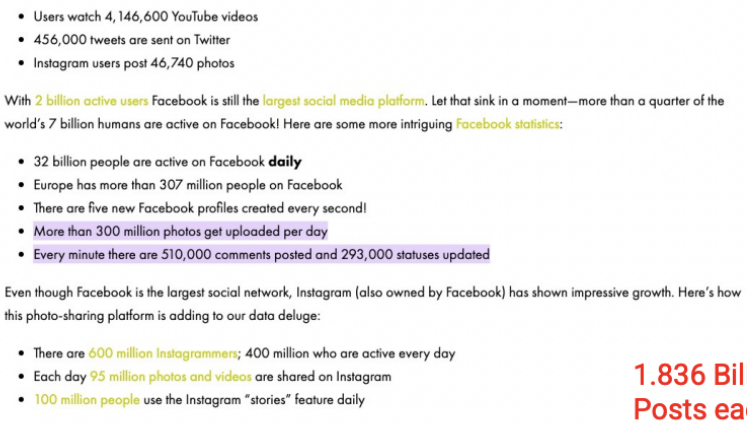
Q : Please write me a sonnet on the subject of the Forth Bridge.
A : Count me out on this one. I never could write poetry.
Q : Add 34957 to 70764
A : (Pause about 30 seconds and then give as answer) 105621.
Q : Do you play chess?
A : Yes.
Q : I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A : (After a pause of 15 seconds) R-R8 mate.

    b.
    c. 

4. Chatbots and the Turing Test
    a. Throughout the history of AI, the Turing Test has been a semi-serious benchmark to aim for, starting with Joseph Weizenbaum's Eliza
    b. Do you think ChatGPT passes the Turing Test?
        i. Not good at developing logic
5. History of NLP
    a. 1940s –1950s: Foundations
        i. Development of formal language theory (Chomsky, Backus, Naur, Kleene)
        ii. Information theory (Shannon)
    b. 1957 – 1970s:
        i. Use of formal grammars as basis for natural language processing (Chomsky, Kaplan)
        ii. Use of logic and logic-based programming (Minsky, Winograd, Colmerauer, Kay)

    c. 1970s – 1983:
        i. Probabilistic methods for early speech recognition (Jelinek, Mercer)
        ii. Discourse modeling (Grosz, Sidner, Hobbs)
    d. 1983 – 1993:
        i. Finite state models (morphology) (Kaplan, Kay)
    e. 1993 – present:
        i. Strong integration of different techniques, different areas.
6. Most important Tasks in NLP
    a. Text Preprocessing
    b. Spelling Correction
        i. Finding most likely re-spelling of a word not in the dictionary
    c. Normalization → take the words and make them shorter
        i. Tokenization – taking the sentence and choosing the best words
        ii. Stemming – taking the end out of it
        iii. Lemmatization – taking word and reducing to its stem (is, are, was → reduce to infinitive)
    d. Part Of Speech (POS) tagging
        i. tagging a word in a text with its part of speech. A part of speech is a category of words with similar grammatical properties, such as noun, verb, adjective, adverb, pronoun, preposition, conjunction, etc.
    e. Word Sense Disambiguation
        i. associating words in context with their most suitable entry in a pre-defined sense inventory (typically WordNet).
    f. Grammatical Error Correction
        i. correcting different kinds of errors in text such as spelling, punctuation, grammatical, and word choice errors.
7. Most Important Tasks in NLP
    a. Classification
    b. Text Classification: assigning a category to a sentence or document
    c. Methods:
        i. Rule-based: human-designed rules on keywords and phrases
        ii. Machine learning: Naive Bayesian, Logistic regression, Support Vector Machines
        iii. Deep learning
        iv. Hybrid Methods: Add rules downstream from ML approach to deal with exceptions

d. Applications:
    i. Spam detection
    ii. Topic Labeling (where to store this data for later retrieval)
    iii. Customer Feedback
    iv. Urgency Detection (how important is this email)
    v. Intent Detection (what is the reason behind this customer feedback)
    vi. Language Detection (e.g., before input to machine translation system)
    vii. Deep fake detection ("I am not a robot")
    viii. Sentiment Analysis
e. Sentiment Analysis: identifying the position of a piece of text in some scale of sentiment
f. Position may be categorical (2 out of 5 stars) or continuous in some range
g. Types of sentiment
    i. Positive - Negative
    ii. Aspect or point of view or bias (e.g. political)
    iii. Intent detection
    iv. Emotion Detection
        1. Happiness
        2. Excited/enthusiastic
        3. Frustration or Anger
    v. Friendship, affection, love or sexual attraction
    vi. Humorous
    vii. Irony
    viii. Hate speech and fake news detection
h. Fake news Detection: detecting and filtering out texts containing false and misleading information
i. Stance Detection: determining an individual's reaction to a primary actor's claim. It is a core part of a set of approaches to fake news assessment
j. Hate Speech Detection: detecting if a piece of text contains hate speech



- Users watch 4,146,600 YouTube videos
- 456,000 tweets are sent on Twitter
- Instagram users post 46,740 photos

With 2 billion active users Facebook is still the largest social media platform. Let that sink in a moment—more than a quarter of the world's 7 billion humans are active on Facebook! Here are some more intriguing Facebook statistics:

- 32 billion people are active on Facebook **daily**
- Europe has more than 307 million people on Facebook
- There are five new Facebook profiles created every second!
- More than 300 million photos get uploaded per day
- Every minute there are 510,000 comments posted and 293,000 statuses updated

Even though Facebook is the largest social network, Instagram (also owned by Facebook) has shown impressive growth. Here's how this photo-sharing platform is adding to our data deluge:

- There are 600 million Instagrammers; 400 million who are active every day
- Each day 95 million photos and videos are shared on Instagram
- 100 million people use the Instagram "stories" feature daily

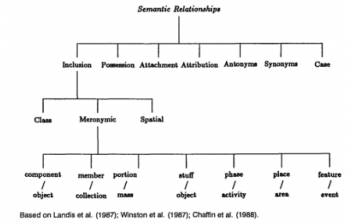
1.836 Billion Facebook Posts each day....

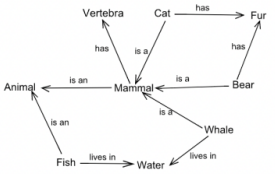
    i.

k. Information Retrieval (An old subject, even before google made it the most popular text-processing task)
    i. Resource Retrieval from text queries/questions
        1. Resource could be
            a. Highly structured (relational database, code)
            b. Semi-structured (Markup Languages (XML), labeled documents)
            c. Unstructured (documents)
        2. Database search from keywords
        3. Google search
        4. Backend to Speech to Text systems (siri)
        5. Question Answering
    ii. Sentence/document similarity: determining how "similar" two texts are
        1. Notion of "similar" is variable (similar topic, similar sentiment, ...)
        2. Relationship to IR:
        3. How similar is text query to a document?
        4. "Retrieve more documents similar to this one"
        5. Create a map/graph of documents similar to given sentence/document • Plagiarism/copyright infringement
    iii. Document Ranking: Rank documents as to some criterion (e.g., PageRank)
        1. How well does this document satisfy my query?
        2. How important/authoritative is this document?
l. Entities, Relations, and Knowledge Graphs

Concept Map about Electricity:

    i.





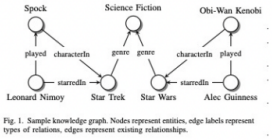


Fig. 1. Sample knowledge graph. Nodes represent entities, edge labels represent types of relations, edges represent existing relationships.

ii.

iii. Named Entity Recognition: tagging entities in text with their corresponding type, typically in BIO notation.)

iv. Coreference Resolution: clustering mentions in text that refer to the same underlying real-world entities.

v. Relation extraction: extracting semantic relationships from a text, e.g.,
    1. Is-A
    2. Has-A Son-Of
    3. Part-Of
    4. Size-of

vi. Build a graph structure:
    1. Knowledge Graph
    2. Concept Map
    3. Mind Map

vii. Graphs can be used to enhance other NLP tasks: search, similarity, question answering, etc

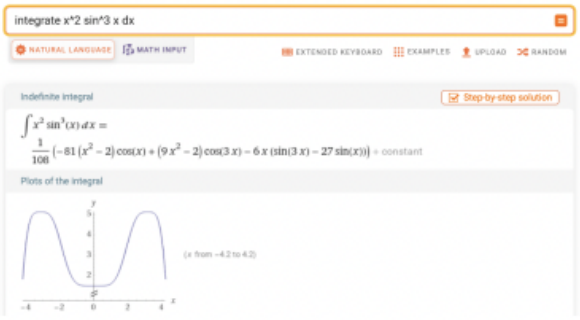
viii. Entity Linking: recognizing and disambiguating named entities to a knowledge base (e.g., Wikidata).

ix. Relation prediction: identifying a named relation between two named semantic entities.

m. Text-to-Text Generation
    i. Machine Translation: translating from one language to another.
    ii. Text Generation: creating text from a prompt or subject phrase that appears indistinguishable from human-written text.
    iii. Lexical Normalization: translating/transforming a non-standard text to a standard register.
    iv. Paraphrase Generation: creating an output sentence that preserves the meaning of input but includes variations in word choice and grammar.
    v. Text Simplification: making a text easier to read and understand, while preserving its main ideas and approximate meaning.
    vi. Text Summarization

n. Text Summarization
    i. Topic Modeling: identifying abstract “topics” underlying a collection of documents.
    ii. Keyword Extraction: identifying the most relevant terms to describe the subject of a document
    iii. Text Summarization: Reducing size of document while preserving the most important information
    iv. Use cases for Text Summarization:
        1. Summaries for busy executives or (students!)
        2. Summaries of articles, books, chapters
        3. Automatic Table of Contents or Indices
        4. Downstream from Speech-to-Text systems:
            a. Notetaking of meetings, lectures
            b. Abstracts of podcasts, YouTube videos
            c. Automatic summary of customer phone calls
o. Chatbots and Question Answering
    i. Slot Filling or Cloze Task: aims to extract the values of certain types of attributes (or slots, such as cities or dates) for a given entity from texts.
    ii. Chatbots: Conversation agents (started with Eliza in early 1060's!)
    iii. Dialog Management: managing of state and flow of conversations.
    iv. Question Answering: Responding to textual queries with textual answers
        1. Extractive QA: The model extracts the answer from a knowledge source, such as a knowledge graph, database, or document (next slide).
        2. Open Generative QA: The model generates free text directly based on the (global) context.
        3. Closed Generative QA: The model generates free text directly based only on the question.
p. Reasoning with Text
    i. Logical Relationship of two sentences/documents:
        1. Entailment
        2. Temporal sequence
        3. Specialization
    ii. Subsystem of text generation at scale
    iii. Text-to/from-First Order Logic: Translate between text and expressions in first-order logic:
    iv. Use cases:
        1. Teaching logic
        2. Game/puzzle solving
        3. Interface to automated theorem prover

4. Prolog
5. Planner
6. Wolfram Alpha



q. Text-to-Data and Data-to-Text

i.

ii.

An example of some of the images created by Imagen, Google's text-to-image AI generator.

iii. Text-to-Image: generating photo-realistic images which are semantically consistent with the text descriptions.

iv. Image captioning: Generate captions for input images

v. Video-to-Text: Generating text describing a sequence of images

vi. Text-to-Speech: Human-like reading of input text.

vii. Speech-to-Text: transcribing speech to text