CS 505: Introduction to Natural Language Processing

Wayne Snyder
Boston University

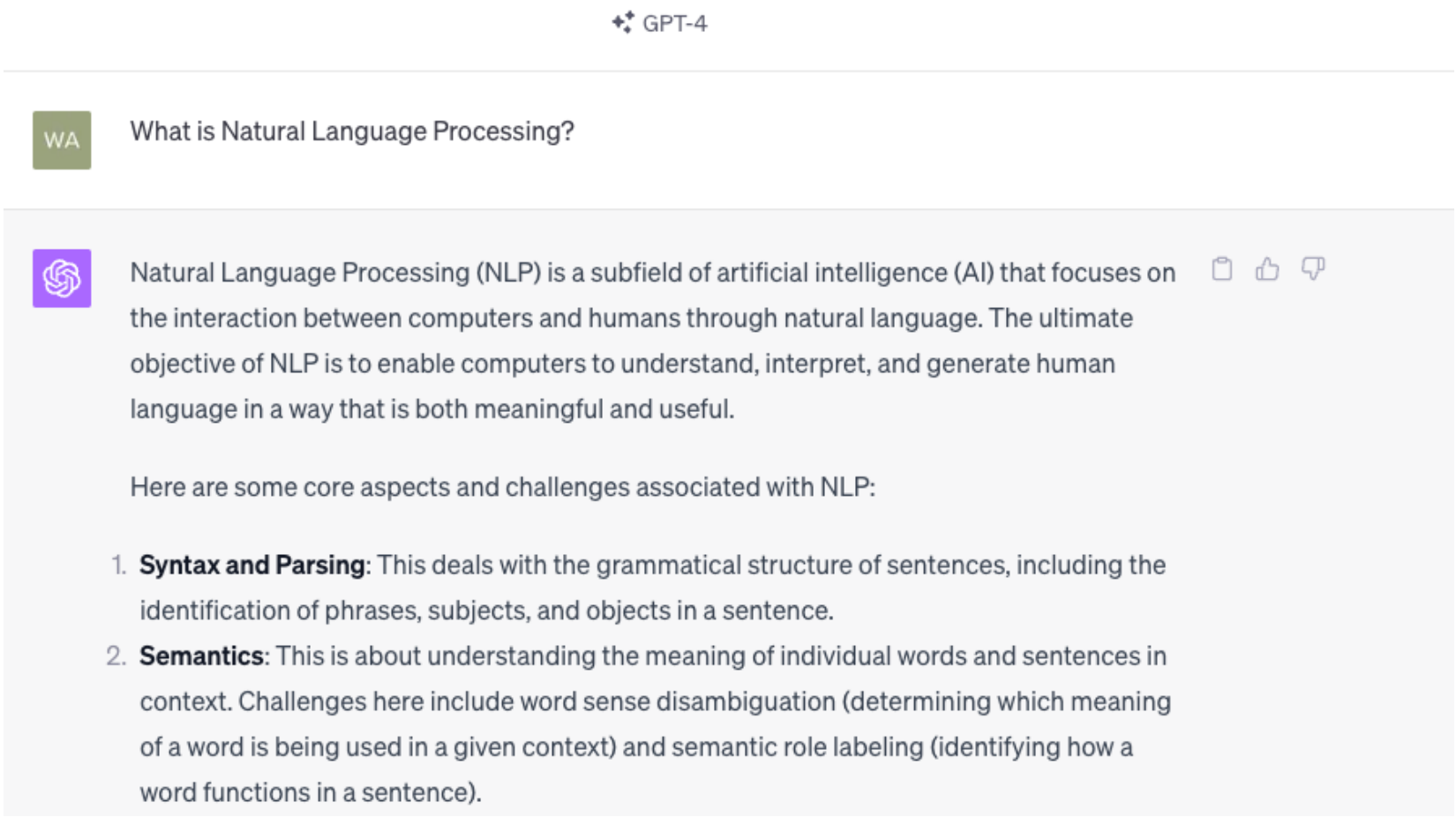
Lecture 2: What is Natural Language Processing?

What is Natural Language Processing (NLP)?

I wonder who we could ask about this…..?

What is Natural Language Processing (NLP)?

I wonder who we could ask about this…..?



What is Natural Language Processing (NLP)?

3. **Pragmatics**: This refers to understanding language in context, considering things like speaker intent and how meaning can change based on the situation or context.
4. **Discourse**: This involves understanding how consecutive sentences relate to each other, for example, identifying when a sentence refers back to something mentioned previously.
5. **Speech Processing**: This deals with understanding and generating spoken language, including tasks like speech recognition, speech synthesis, and speaker identification.
6. **Information Retrieval**: This is about extracting relevant information from large text corpora, as is done by search engines.
7. **Machine Translation**: This involves translating text from one language to another automatically.
8. **Named Entity Recognition (NER)**: This is the task of identifying and classifying proper names (e.g., "Barack Obama" as a person, "London" as a location) in a text.
9. **Sentiment Analysis**: This involves determining the sentiment or mood of a piece of text, like identifying if a product review is positive, negative, or neutral.
10. **Text Generation**: This is about generating coherent and contextually relevant text based on a given input, a task for which models like GPT-3 and GPT-4 are known.

Regenerate

What is Natural Language Processing (NLP)?

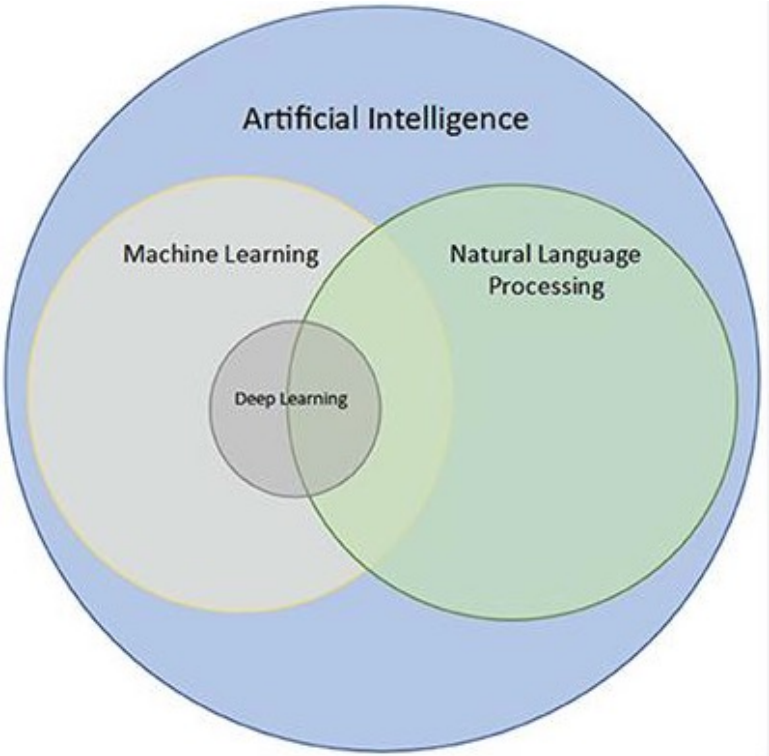
Advancements in deep learning and neural network architectures in recent years, especially the rise of transformers and models like BERT, GPT, and their variants, have led to significant progress in many NLP tasks.

However, despite these advances, true natural language understanding remains a challenging goal. Humans use a vast amount of background knowledge, context, and reasoning abilities when processing language, and capturing all of these nuances in a machine remains an active area of research.

Hm… I couldn’t have said it better myself!

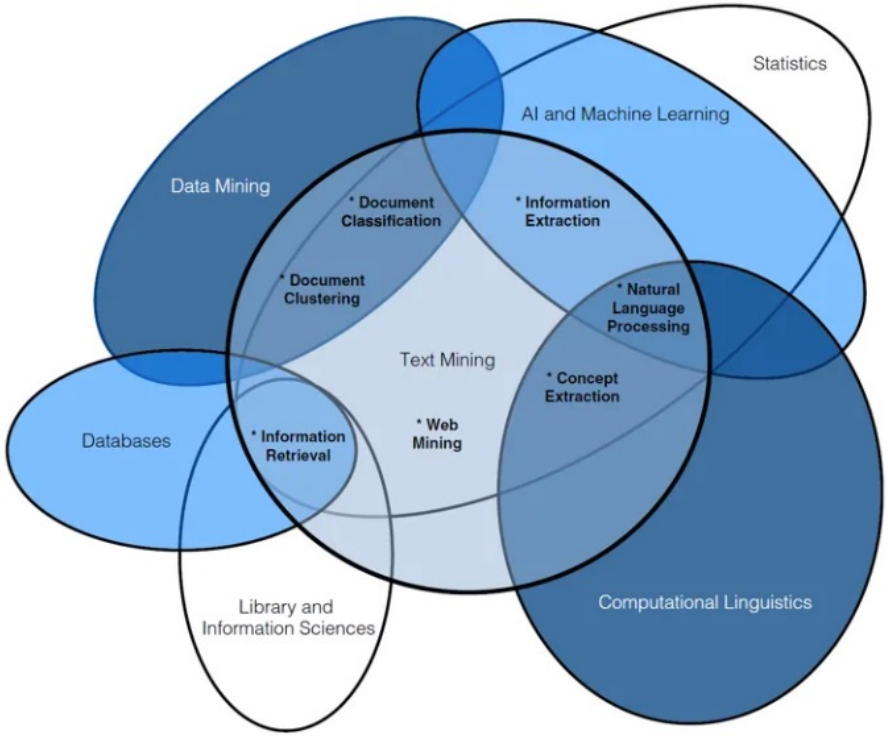
NLP and Aritificial Intelligence

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):



NLP and Artificial Intelligence

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:

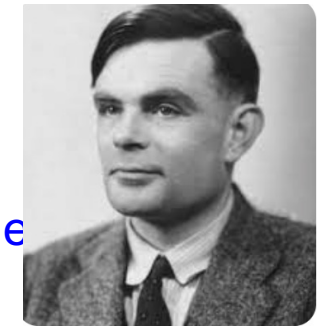


Note: this diagram, already somewhat out of date, does not include speech processing.

In the last few years, NLP and Computational Linguistics have diverged, with CL being associated with using CS/NLP techniques to answer questions interesting to linguists.

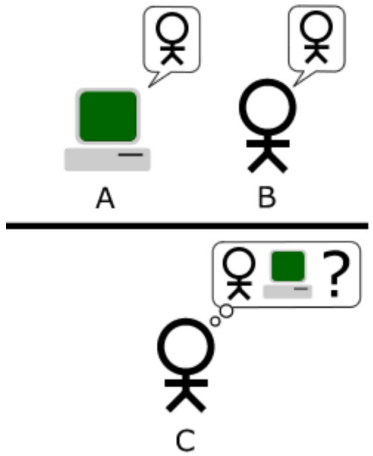
Most NLP professionals know little about linguistics!

NLP in Artificial Intelligence



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 provide by Alan Turing, with the famous “Turing Test”:

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.



VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

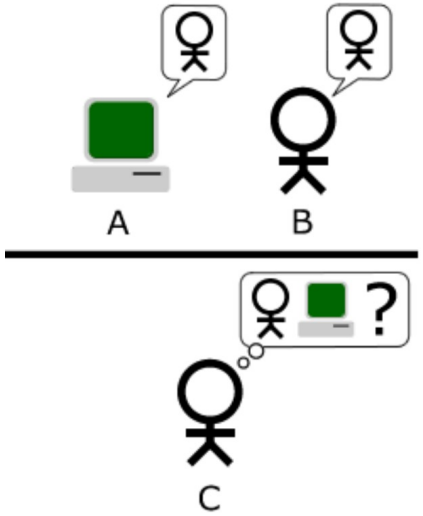
BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, ‘Can machines think?’ This should

Turing Test

Turing gave several examples of the kind of conversations that might take place:



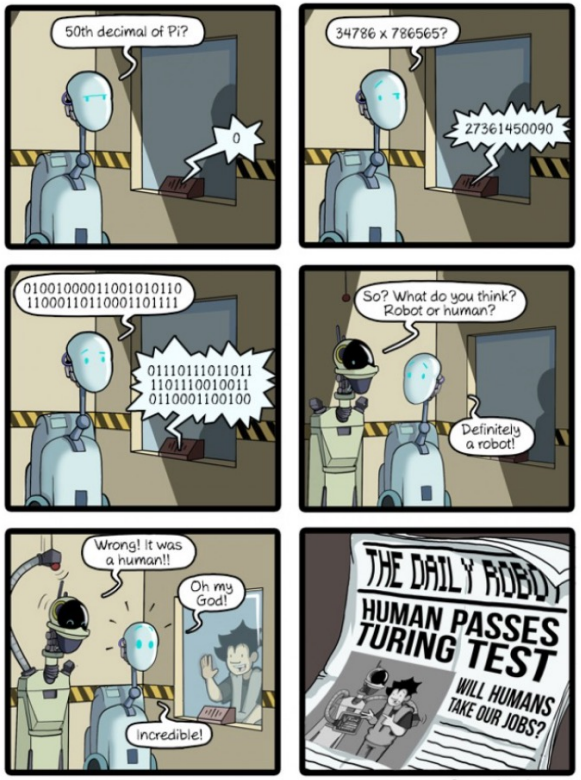
Q : Please write me a sonnet on the subject of the Forth Bridge.
A : Count me out on this one. I never could write poetry.
Q : Add 34957 to 70764
A : (Pause about 30 seconds and then give as answer) 105621.
Q : Do you play chess?
A : Yes.
Q : I have K at my K1, and no other pieces. You have only K at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

Turing Test

This has been a part of our culture for a long time....

"The good news,Dave, is that the computer's passed the Turing test. The bad news is that you've failed. "

Reverse Turing Test



ChatBots and the Turing Test

Throughout the history of AI, the Turing Test has been a semi-serious benchmark to aim for, starting with Joseph Weizenbaum's Eliza:

https://web.njit.edu/~ronkowit/eliza.html

The Loebner Prize was an annual contest (discontinued in 2020), to award the best human-simulation in a Turing Test:

Loebner Prize Gold Medal

(Solid 18 carat, not *gold-plated* like the Olympic "*Gold*" medals)

And now we have chatGPT, with which we have started, and will end, our course!

https://openai.com/blog/chatgpt/

Question: Do you think chatGPT passes the Turing Test?

History of Natural Language Processing

- **1940s –1950s: Foundations**
  - Development of formal language theory (Chomsky, Backus, Naur, Kleene)
  - Information theory (Shannon)
- **1957 – 1970s:**
  - Use of formal grammars as basis for natural language processing (Chomsky, Kaplan)
  - Use of logic and logic-based programming (Minsky, Winograd, Colmerauer, Kay)
- **1970s – 1983:**
  - Probabilistic methods for early speech recognition (Jelinek, Mercer)
  - Discourse modeling (Grosz, Sidner, Hobbs)
- **1983 – 1993:**
  - Finite state models (morphology) (Kaplan, Kay)
- **1993 – present:**
  - Strong integration of different techniques, different areas.

Most Important Tasks in NLP

Two minutes NLP — 33 important NLP tasks explained

Information Retrieval, Knowledge Bases, Chatbots, Text Generation, Text-to-Data, Text Reasoning, etc.

From the Medium post by Fabnio Chiusano, 12/7/2021

(Posted on the class web site)

Most Important Tasks in NLP

Text Preprocessing

- **Spelling Correction:** Finding most likely re-spelling of a word not in the dictionary
- **Normalization:**
  - Tokenization,
  - Stemming,
  - Lemmatization
- **Part Of Speech (POS) tagging:** 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.
- **Word Sense Disambiguation:** associating words in context with their most suitable entry in a pre-defined sense inventory (typically WordNet).
- **Grammatical Error Correction:** correcting different kinds of errors in text such as spelling, punctuation, grammatical, and word choice errors.

Most Important Tasks in NLP

Classification

Text Classification: assigning a category to a sentence or document.

Methods:

- Rule-based: Human-designed rules on keywords and phrases:
- Machine Learning: Naive Bayes, Logistic Regression, Support Vector Machines,
- Deep-Learning (e.g., BERT)
- Hybrid Methods: Add rules downstream from ML approach to deal with exceptions.

Applications:

- Spam Detection
- Topic Labeling (where to store this data for later retrieval)
- Customer Feedback
- Urgency Detection (how important is this email)
- Intent Detection (what is the reason behind this customer feedback)
- Language Detection (e.g., before input to machine translation system)
- Deep fake detection ("I am not a robot") 
- Sentiment Analysis (next slide)

Most Important Tasks in NLP

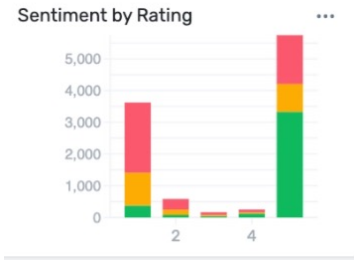
Classification

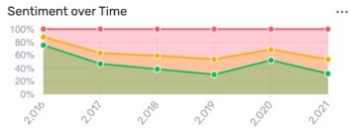
Sentiment Analysis: identifying the position of a piece of text in some scale of sentiment.

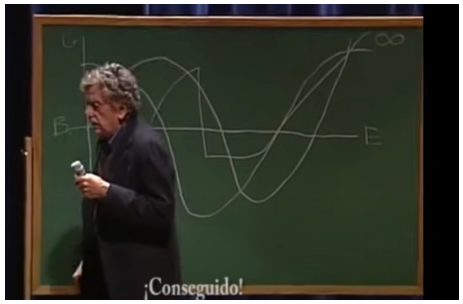
Position may be categorical (2 stars out of 5) or continuous in some range (2.3 on a scale 0 .. 10)

Types of sentiment:

- Positive – Negative
- Aspect or point of view or bias (e.g., political)
- Intent detection
- Emotion Detection
  - Happiness
  - Excited/enthusiastic
  - Frustration or Anger
- Friendship, affection, love or sexual attraction
- Humorous
- Irony
- Hate speech and Fake News detection (next slide)







Source: https://monkeylearn.com/sentiment-analysis/

Most Important Tasks in NLP

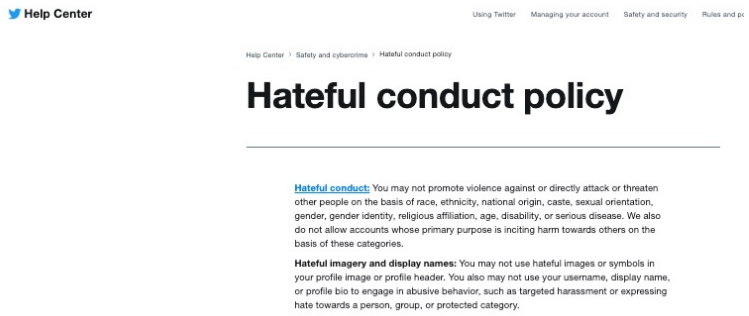
Classification: Fake News and Hate Speech Detection

Fake News Detection: detecting and filtering out texts containing false and misleading information.

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.

Hate Speech Detection: detecting if a piece of text contains hate speech.

Facebook/Meta:


Help Center

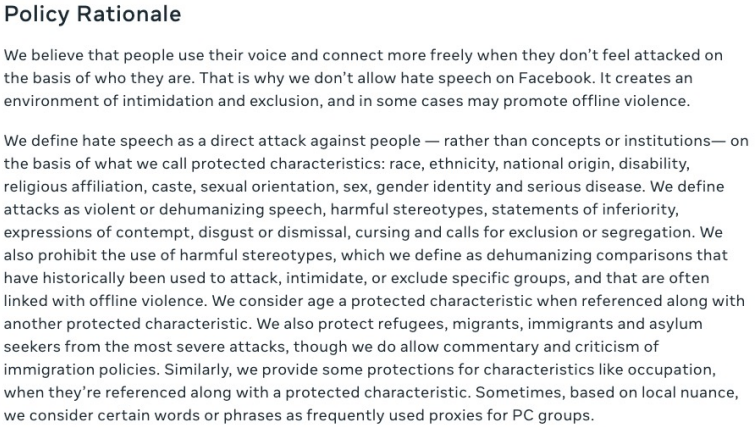
Using Twitter Managing your account Safety and security Rules and p

Help Center › Safety and cybercrime › Hateful conduct policy

Hateful conduct policy

Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.


Policy Rationale

We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That is why we don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence.

We define hate speech as a direct attack against people — rather than concepts or institutions— on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic. Sometimes, based on local nuance, we consider certain words or phrases as frequently used proxies for PC groups.

Most Important Tasks in NLP

Fake News and Hate Speech Detection

Why is NLP so necessary in this?

- 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....

Most Important Tasks in NLP

Information Retrieval

(An old subject, even before Google made it the the most popular text-processing task.)

Resource Retrieval from text queries/questions

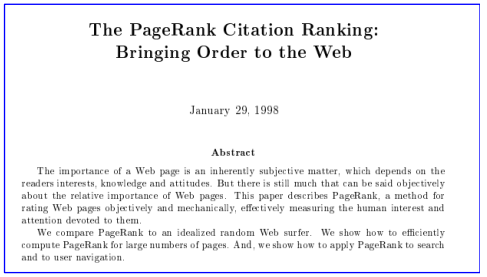
- Resource could be
  - Highly structured (relational database, code)
  - Semi-structured (Markup Languages (XML), labeled documents)
  - Unstructured (documents)
- Database search from keywords
- Google search
- Backend to Speech to Text systems (siri)
- Question Answering (next slide)

Sentence/document similarity: determining how "similar" two texts are

- Notion of "similar" is variable (similar topic, similar sentiment, ...)
- Relationship to IR:
  - How similar is text query to a document?
  - "Retrieve more documents similar to this one"
- Create a map/graph of documents similar to given sentence/document
- Plagiarism/copyright infringement

Document Ranking: Rank documents as to some criterion (e.g., PageRank)

- How well does this document satisfy my query?
- How important/authoritative is this document?



The PageRank Citation Ranking:
Bringing Order to the Web

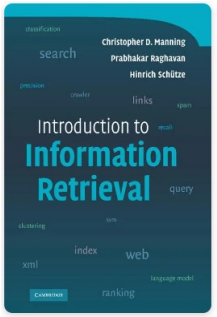
January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

1999



2008

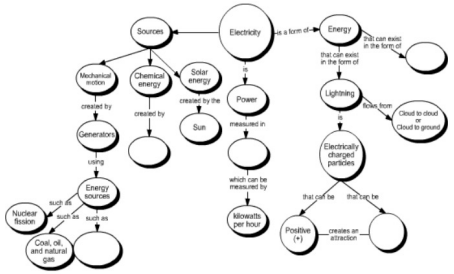
Net worth of Google in 2022: $1.135 Trillion.

Most Important Tasks in NLP

Entities, Relations, and Knowledge Graphs

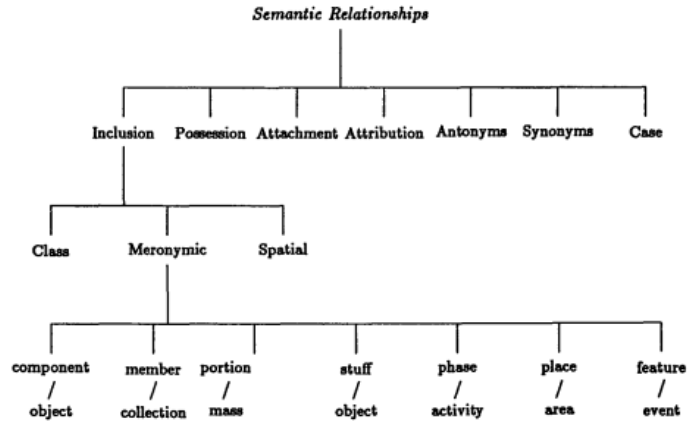
- Named Entity Recognition: tagging entities in text with their corresponding type, typically in BIO notation.)
- Coreference Resolution: clustering mentions in text that refer to the same underlying real-world entities.
- Relation extraction: extracting semantic relationships from a text, e.g.,
  - Is-A
  - Has-A
  - Son-Of
  - Part-Of
  - Size-of
  - etc., etc., etc.
- Build a graph structure:
  - Knowledge Graph
  - Concept Map
  - Mind Map
- Graphs can be used to enhance other NLP tasks: search, similarity, question answering, etc
- Entity Linking: recognizing and disambiguating named entities to a knowledge base (e.g., Wikidata).
- Relation prediction: identifying a named relation between two named semantic entities.

Concept Map about Electricity:

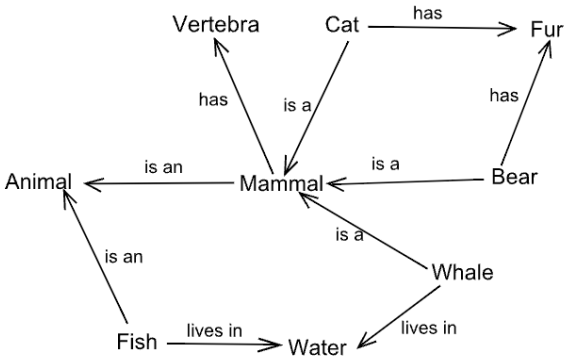


Most Important Tasks in NLP

Entities, Relations, and Knowledge Graphs



Based on Landis et al. (1987); Winston et al. (1987); Chaffin et al. (1988).



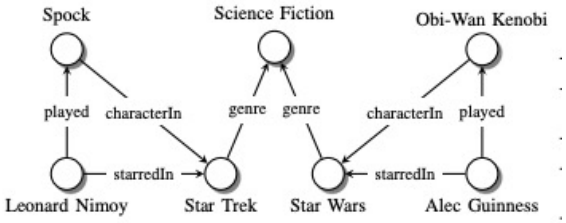


Fig. 1. Sample knowledge graph. Nodes represent entities, edge labels represent types of relations, edges represent existing relationships.

Most Important Tasks in NLP

Text-to-Text Generation

- **Machine Translation:** translating from one language to another.
  - Covered in lecture – Transformer technology transformed this task
- **Text Generation:** creating text from a prompt or subject phrase that appears indistinguishable from human-written text.
  - Covered in lecture – Use language models, Large Language Models (GPT) have transformed this task
- **Lexical Normalization:** translating/transforming a non-standard text to a standard register.
- **Paraphrase Generation:** creating an output sentence that preserves the meaning of input but includes variations in word choice and grammar.
- **Text Simplification:** making a text easier to read and understand, while preserving its main ideas and approximate meaning.
- **Text Summarization** (next slide)

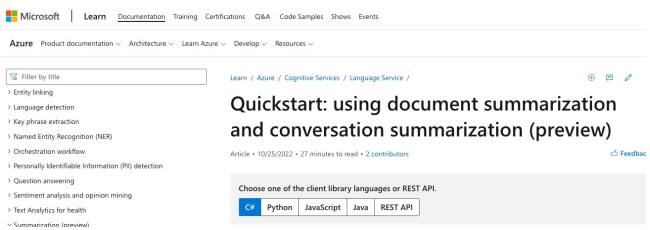
How Large Language Models are Transforming Machine-Paraphrased Plagiarism

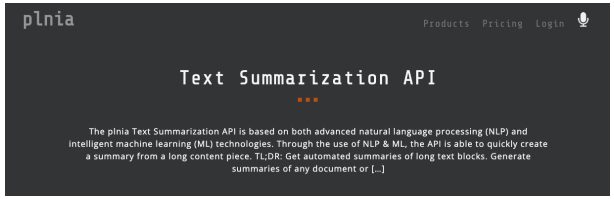
Jan Philip Wahle[♣*], **Terry Ruas**[*], **Frederic Kirstein**[♠*], **Bela Gipp**[*]
[*]Georg-August-Universität Göttingen, Germany
[♠]Mercedes-Benz Group AG, Germany
[♣]wahle@gipplab.org

Most Important Tasks in NLP

Text Summarization

- **Topic Modeling:** identifying abstract “topics” underlying a collection of documents.
- **Keyword Extraction:** identifying the most relevant terms to describe the subject of a document
- **Text Summarization:** Reducing size of document while preserving the most important information
- Use cases for Text Summarization:
  - Summaries for busy executives or (students!)
  - Summaries of articles, books, chapters
  - Automatic Table of Contents or Indices
  - Downstream from Speech-to-Text systems:
    - Notetaking of meetings, lectures
    - Abstracts of podcasts, YouTube videos
    - Automatic summary of customer phone calls





Most Important Tasks in NLP

Chatbots and Question Answering

- **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.
- **Chatbots:** Conversation agents (started with Eliza in early 1060's!)
- **Dialog Management:** managing of state and flow of conversations.
- **Question Answering:** Responding to textual queries with textual answers
  - **Extractive QA**: The model extracts the answer from a knowledge source, such as a knowledge graph, database, or document (next slide).
  - **Open Generative QA:** The model generates free text directly based on the (global) context.
  - **Closed Generative QA:** The model generates free text directly based only on the question.

Most Important Tasks in NLP

Reasoning with Text

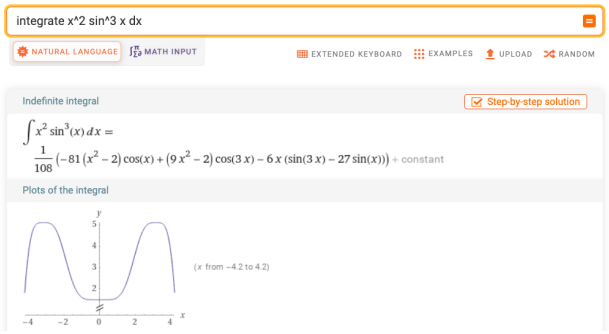
- Logical Relationship of two sentences/documents:
  - Entailment
  - Temporal sequence
  - Specialization
- Subsystem of text generation at scale

10. Logic Puzzle: A farmer wants to cross a river and take with him a wolf, a goat and a cabbage. He has a boat, but it can only fit himself plus either the wolf, the goat or the cabbage. If the wolf and the goat are alone on one shore, the wolf will eat the goat. If the goat and the cabbage are alone on the shore, the goat will eat the cabbage. How can the farmer bring the wolf, the goat and the cabbage across the river without anything being eaten?

- Text-to/from-First Order Logic: Translate between text and expressions in first-order logic:

No student failed Chemistry, but at least one student failed History.
¬∃x (Student(x) ∧ Failed(x,Chemistry)) ∧ ∃x (Student(x) ∧ Failed (x,History))

- Use cases:
  - Teaching logic
  - Game/puzzle solving
  - Interface to automated theorem prover
    - Prolog
    - Planner
    - Wolfram Alpha



Most Important Tasks in NLP

Text-to-Data and Data-to-Text

- **Text-to-Image:** generating photo-realistic images which are semantically consistent with the text descriptions.
- **Image captioning:** Generate captions for input images
- **Video-to-Text:** Generating text describing a sequence of images
- **Text-to-Speech:** Human-like reading of input text.
- **Speech-to-Text:** transcribing speech to text

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