SC4002 CE4045 CZ4045
Natural Language Processing

Introduction

Dr. Sun Aixin

Course Instructors

- Dr. Sun Aixin
  - Email: axsun@ntu.edu.sg
  - Email subject: **[SC4002 CE/CZ4045]**
  - Homepage: https://personal.ntu.edu.sg/axsun/
  - Research Interests: Information Retrieval, Recommender System, NLP

- Dr. Wang Wenya
  - Email: wangwy@ntu.edu.sg
  - Email subject: **[SC4002 CE/CZ4045]**
  - Homepage: https://personal.ntu.edu.sg/wangwy/
  - Research Interests: Reasoning for Natural Language Processing (or multimodal learning), including logic reasoning, commonsense reasoning, knowledge integration etc.

Outline for today

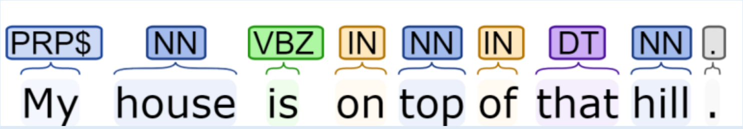
- Teaches key theory and methods for NLP
  - Word-level analysis, parsing, semantics, etc.
  - Learn techniques which can be used in **practical**, robust systems that can (partly) understand human language

- This is **not** a language course
  - Computational methods of processing natural languages
  - But, you are expected to have knowledge of (basic) English grammar

- Course Expectation
  - Preparation
  - Evaluation
  - Learning objective

Pre-requisites

- Basic understanding on English grammar,
  - e.g., verb, noun phrase, preposition



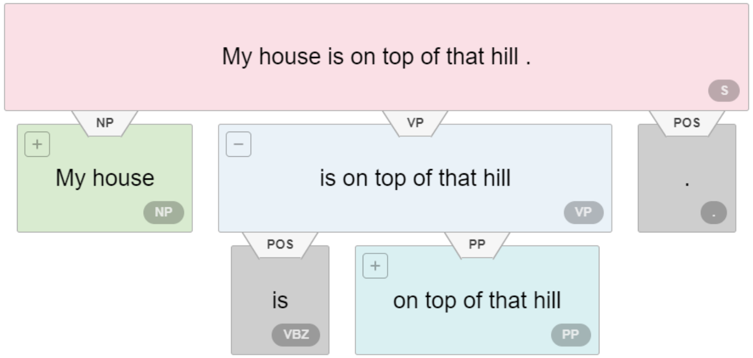
My house is on top of that hill .

possessive pronoun
noun
verb
preposition
determiner

- Basic algorithm and data structure analysis,
  - e.g., dynamic programming

- Basic probability concepts,
  - e.g., conditional probability

$$P(B|A) = \frac{P(A,B)}{P(A)}$$



- **Decent** programming skills

Preparation

- Machine learning?
  - Machine learning knowledge can be helpful for assignment and some parts of lecture
  - Not everyone has the same skills
    - Assumes some ability to learn missing knowledge

- What kind of computation?
  - **Lots of** statistics (the first half)!
  - Some rules based on **linguistic theory** (the first half)
  - **Introduction to Deep Learning** for NLP (the second half)

Learning Objective

- You will learn natural language processing at a **basic level**, establishing a solid understanding on the theory of morphological, syntactic, and semantic analysis.

- With that, you will gain skills to apply the NLP techniques to **real-world problems** by using **NLP packages and toolkits**.

- Upon completion of the course, you should be able to:
  - **Understand and analyze** the linguistic characteristics of written English
  - **Design and develop** an NLP system to analyze and process a general corpus
  - **Troubleshoot** for **domain-specific** NLP applications

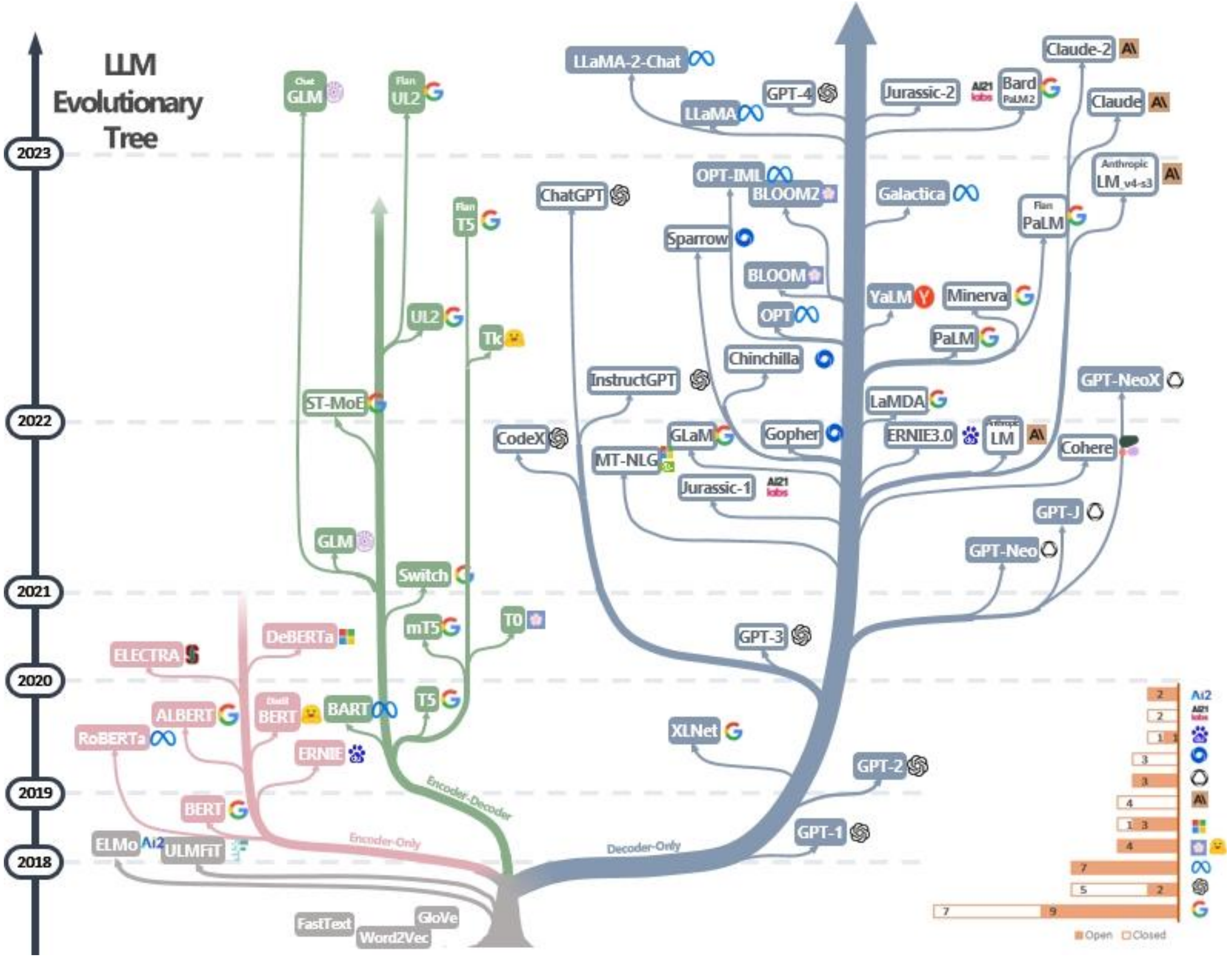
Expectations

- You are **willing to learn NLP**
  - There are a lots of happenings on NLP

- You are expected to participate.

- You are expected to
  - Read **lecture slides** for reference only
  - Read **textbook, reference papers,** and other related materials
  - **Enjoy** assignment and programming!

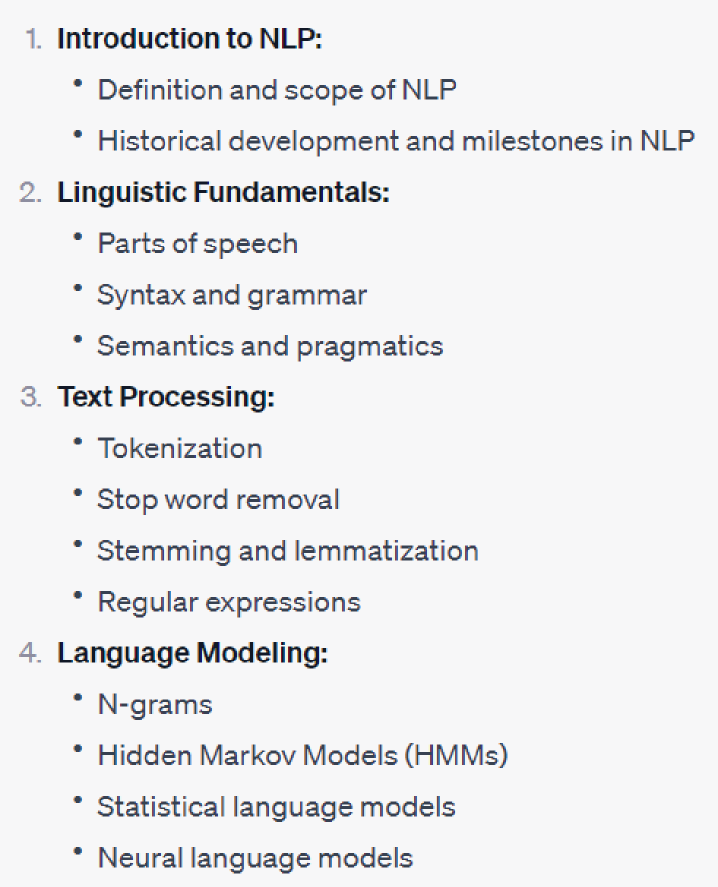
There are a lot of happenings in LLM



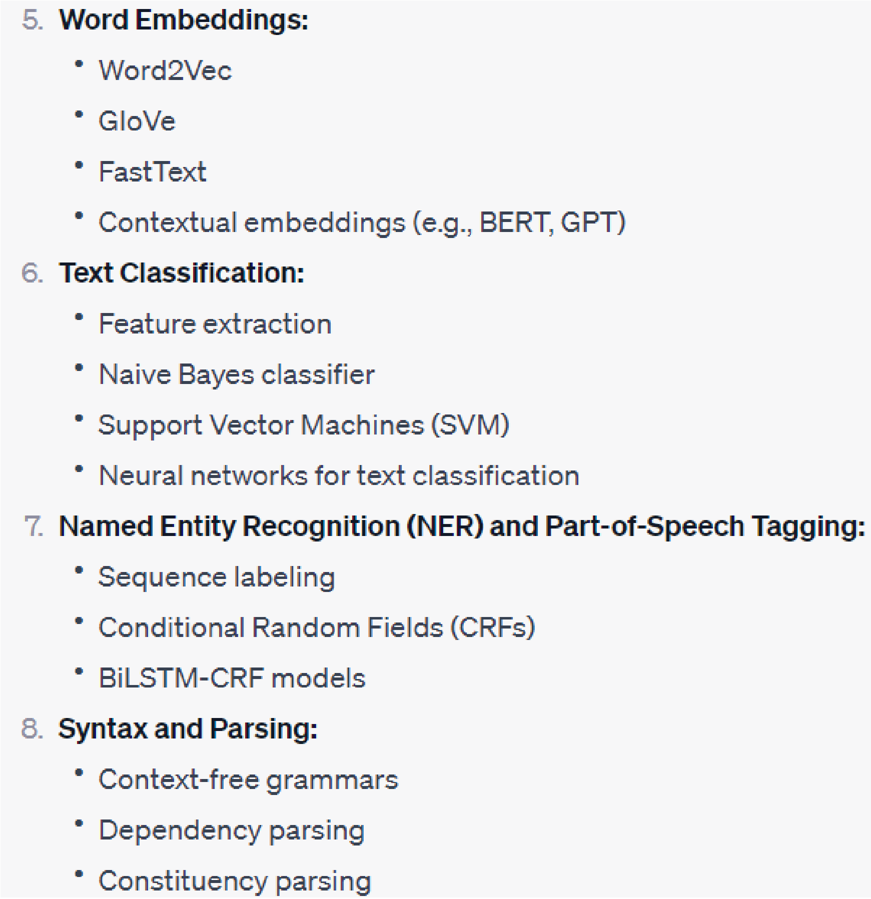
https://github.com/Mooler0410/LLMsPracticalGuide/blob/main/imgs/tree.jpg

What topics to be covered?

- ➢ https://chat.openai.com/share/1427fe3d-f449-4a09-b974-829e9eb9d828

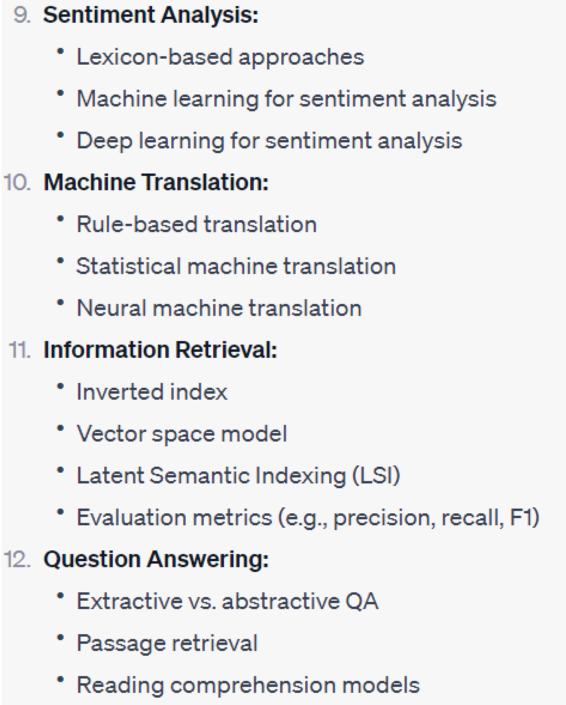


1. **Introduction to NLP:**
   - Definition and scope of NLP
   - Historical development and milestones in NLP
2. **Linguistic Fundamentals:**
   - Parts of speech
   - Syntax and grammar
   - Semantics and pragmatics
3. **Text Processing:**
   - Tokenization
   - Stop word removal
   - Stemming and lemmatization
   - Regular expressions
4. **Language Modeling:**
   - N-grams
   - Hidden Markov Models (HMMs)
   - Statistical language models
   - Neural language models

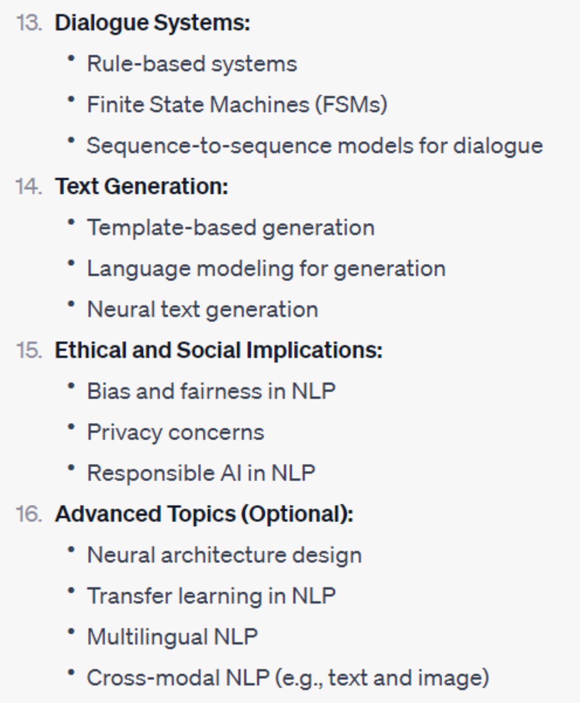


5. **Word Embeddings:**
   - Word2Vec
   - GloVe
   - FastText
   - Contextual embeddings (e.g., BERT, GPT)
6. **Text Classification:**
   - Feature extraction
   - Naive Bayes classifier
   - Support Vector Machines (SVM)
   - Neural networks for text classification
7. **Named Entity Recognition (NER) and Part-of-Speech Tagging:**
   - Sequence labeling
   - Conditional Random Fields (CRFs)
   - BiLSTM-CRF models
8. **Syntax and Parsing:**
   - Context-free grammars
   - Dependency parsing
   - Constituency parsing

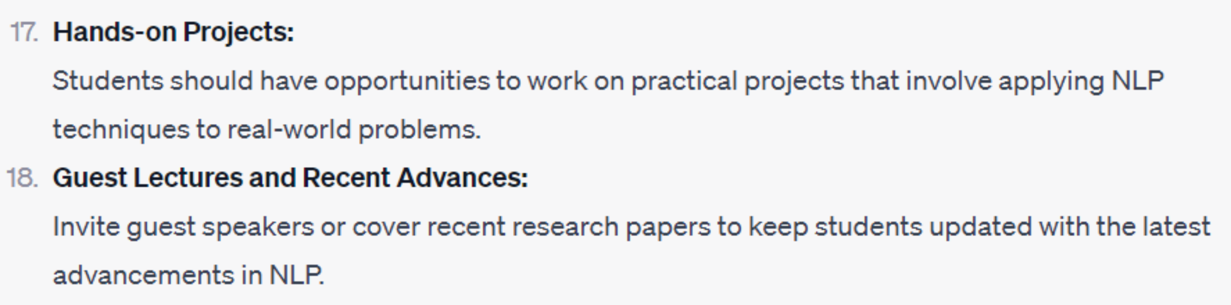
What topics to be covered?



9. **Sentiment Analysis:**
   - Lexicon-based approaches
   - Machine learning for sentiment analysis
   - Deep learning for sentiment analysis
10. **Machine Translation:**
    - Rule-based translation
    - Statistical machine translation
    - Neural machine translation
11. **Information Retrieval:**
    - Inverted index
    - Vector space model
    - Latent Semantic Indexing (LSI)
    - Evaluation metrics (e.g., precision, recall, F1)
12. **Question Answering:**
    - Extractive vs. abstractive QA
    - Passage retrieval
    - Reading comprehension models

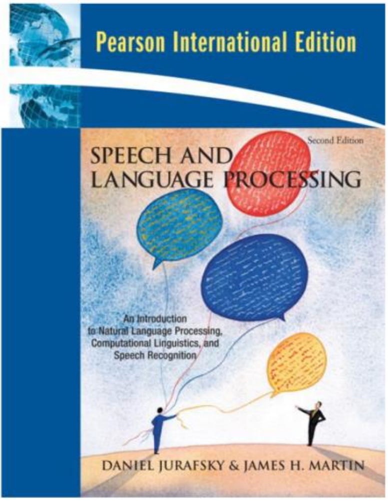


13. **Dialogue Systems:**
    - Rule-based systems
    - Finite State Machines (FSMs)
    - Sequence-to-sequence models for dialogue
14. **Text Generation:**
    - Template-based generation
    - Language modeling for generation
    - Neural text generation
15. **Ethical and Social Implications:**
    - Bias and fairness in NLP
    - Privacy concerns
    - Responsible AI in NLP
16. **Advanced Topics (Optional):**
    - Neural architecture design
    - Transfer learning in NLP
    - Multilingual NLP
    - Cross-modal NLP (e.g., text and image)



17. **Hands-on Projects:**
    Students should have opportunities to work on practical projects that involve applying NLP techniques to real-world problems.
18. **Guest Lectures and Recent Advances:**
    Invite guest speakers or cover recent research papers to keep students updated with the latest advancements in NLP.

Topics


Pearson International Edition
Second Edition
SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
DANIEL JURAFSKY & JAMES H. MARTIN

- First half of lecture https://web.stanford.edu/~jurafsky/slp3/
  - Regular Expression (Chap 2)
  - Text Normalization and Edit Distance (Chap 2)
  - N-grams Language Model (Chap 3)
  - POS Tagging, Hidden Markov Model, Named Entities (Chap 8, A)
  - Constituency Grammars and Constituency Parsing (Chap 17)
  - Dependency Parsing (Chap 18)
- Second half of lecture
  - Introduction to Machine Learning and Deep Learning
  - Word vectors, language modeling
  - Sequence modeling, sequence-to-sequence learning
  - Attentions and transformers
  - Pretraining and natural language generation
  - Prompting and in-context learning

Course Evaluation

- Evaluation Objective
  - Spread evaluation over the whole course, not just one exam or one report

- Main Components
  - 15% Mid-term Quiz (for first half, Week 8 tutorial slot)
  - 35% Group Assignment (to be released around recess week)
  - 50% Final exam

- Tutorial
  - Starts from Week 3

Goals of the field of NLP

- Computers would be a lot more useful if they could
  - handle our email, do our library research, chat to us…
  - Google: google booking demo https://www.youtube.com/watch?v=D5VN56jQMWM

- But someone has to work on the hard problems!
  - How can we tell computers about language?
  - Help them learn it as kids do?

- In this course we seek to identify many of the **open research problems (?)** in Natural Language Processing

What/where is NLP?

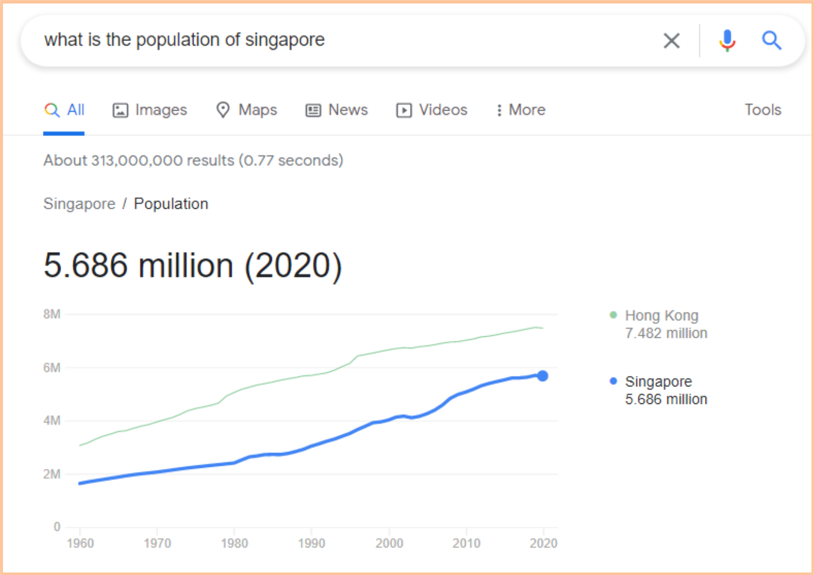
- Goals can be very far reaching …
  - True text understanding (does ChatGPT understand all your questions?)
  - Real-time participation in spoken dialogs

- Or very down-to-earth … (the Web, business documents)
  - Finding the price of products on the web
  - Sentiment detection about products or stocks
  - Extracting facts or relations from documents

- These days, the latter predominate
  - NLP becomes increasingly practical, and it is increasingly engineering oriented, and Large Language Models are the main backbones

Example of down-to-earth Applications

- Machine translation,
  - e.g., https://translate.google.com/
  - ChatGPT and other language models
- Question answering directly through Web search or chat interfaces
- Information extraction:
  - Extracting product information from the Web
- Text analytics
  - Sentiment Analysis
  - Document summarization

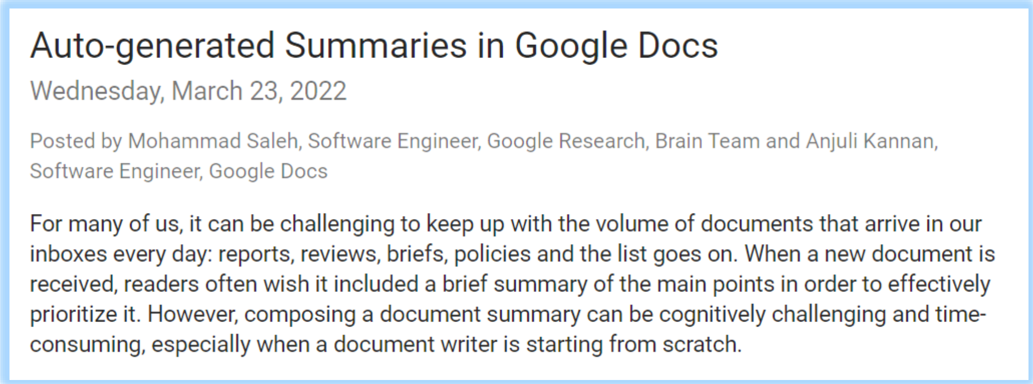


what is the population of singapore

About 313,000,000 results (0.77 seconds)

Singapore / Population

5.686 million (2020)



Auto-generated Summaries in Google Docs

Wednesday, March 23, 2022

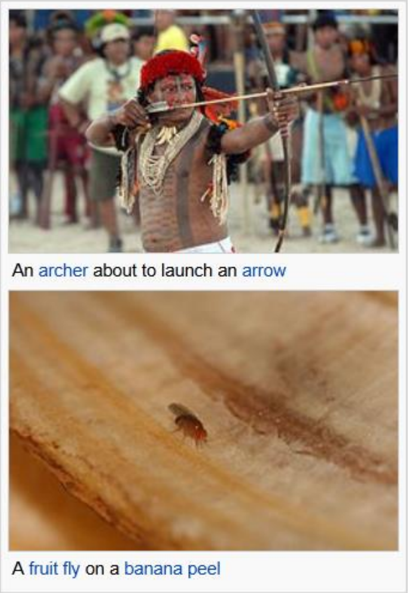
Posted by Mohammad Saleh, Software Engineer, Google Research, Brain Team and Anjuli Kannan, Software Engineer, Google Docs

For many of us, it can be challenging to keep up with the volume of documents that arrive in our inboxes every day: reports, reviews, briefs, policies and the list goes on. When a new document is received, readers often wish it included a brief summary of the main points in order to effectively prioritize it. However, composing a document summary can be cognitively challenging and time-consuming, especially when a document writer is starting from scratch.

Natural Language Processing (NLP) is difficult

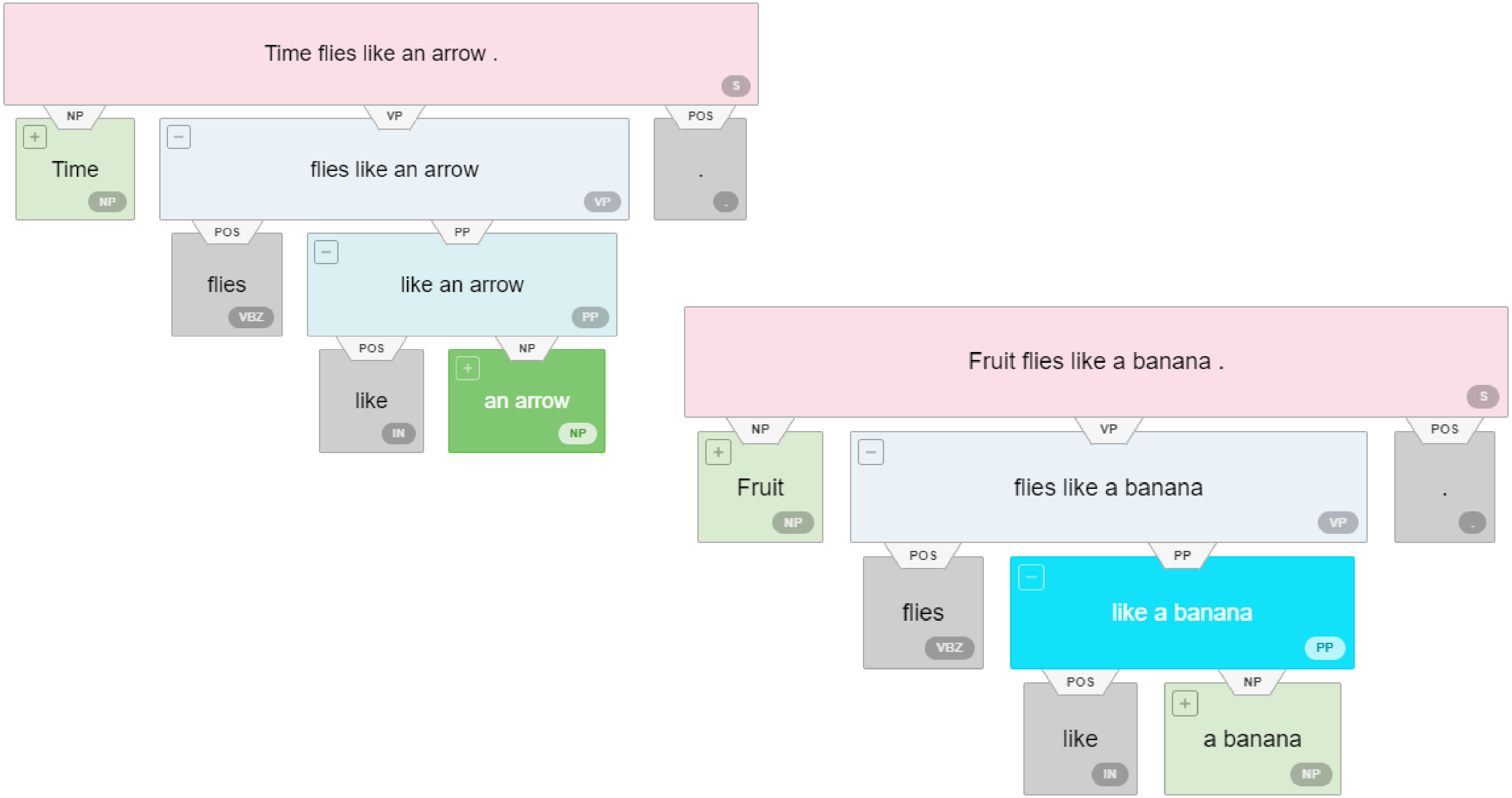
- Natural language is hugely **ambiguous**;
  - We don’t often come up with exact solutions/algorithms

Time *flies* like an arrow.
Fruit *flies* like a banana.


An archer about to launch an arrow

A fruit fly on a banana peel

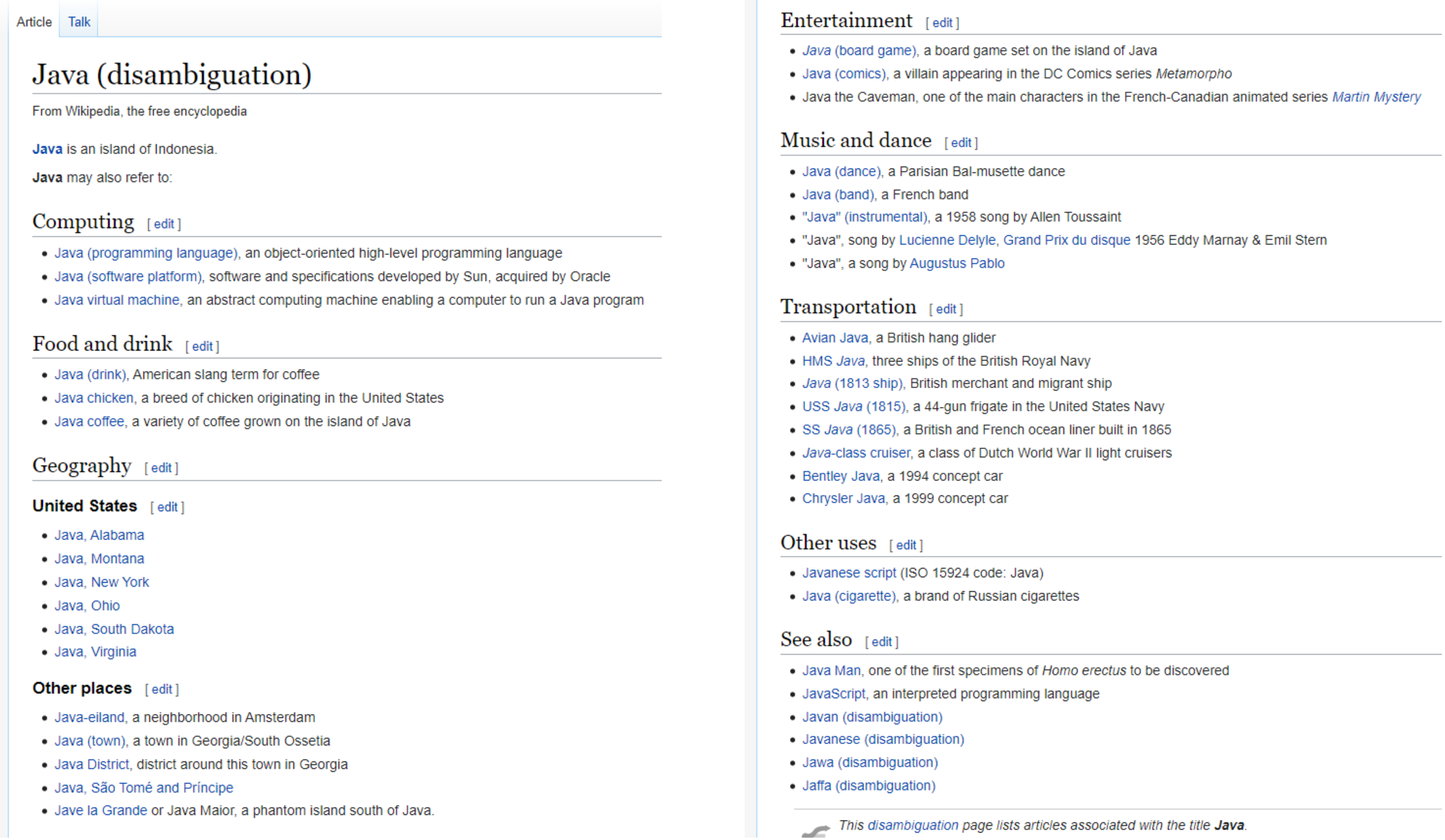
Results from https://demo.allennlp.org/constituency-parsing



https://chat.openai.com/share/870b8d17-f9ae-4445-b3f8-b7084e2015b5

What is “Java”?

https://en.wikipedia.org/wiki/Java_(disambiguation)



Article Talk

Java (disambiguation)

From Wikipedia, the free encyclopedia

Java is an island of Indonesia.

Java may also refer to:

Computing [edit]

- Java (programming language), an object-oriented high-level programming language
- Java (software platform), software and specifications developed by Sun, acquired by Oracle
- Java virtual machine, an abstract computing machine enabling a computer to run a Java program

Food and drink [edit]

- Java (drink), American slang term for coffee
- Java chicken, a breed of chicken originating in the United States
- Java coffee, a variety of coffee grown on the island of Java

Geography [edit]

United States [edit]

- Java, Alabama
- Java, Montana
- Java, New York
- Java, Ohio
- Java, South Dakota
- Java, Virginia

Other places [edit]

- Java-eiland, a neighborhood in Amsterdam
- Java (town), a town in Georgia/South Ossetia
- Java District, district around this town in Georgia
- Java, São Tomé and Príncipe
- Jave la Grande or Java Maior, a phantom island south of Java.

Entertainment [edit]

- *Java* (board game), a board game set on the island of Java
- Java (comics), a villain appearing in the DC Comics series *Metamorpho*
- Java the Caveman, one of the main characters in the French-Canadian animated series *Martin Mystery*

Music and dance [edit]

- Java (dance), a Parisian Bal-musette dance
- Java (band), a French band
- "Java" (instrumental), a 1958 song by Allen Toussaint
- "Java", song by Lucienne Delyle, Grand Prix du disque 1956 Eddy Marnay & Emil Stern
- "Java", a song by Augustus Pablo

Transportation [edit]

- Avian Java, a British hang glider
- HMS *Java*, three ships of the British Royal Navy
- *Java* (1813 ship), British merchant and migrant ship
- USS *Java* (1815), a 44-gun frigate in the United States Navy
- SS *Java* (1865), a British and French ocean liner built in 1865
- *Java*-class cruiser, a class of Dutch World War II light cruisers
- Bentley Java, a 1994 concept car
- Chrysler Java, a 1999 concept car

Other uses [edit]

- Javanese script (ISO 15924 code: Java)
- Java (cigarette), a brand of Russian cigarettes

See also [edit]

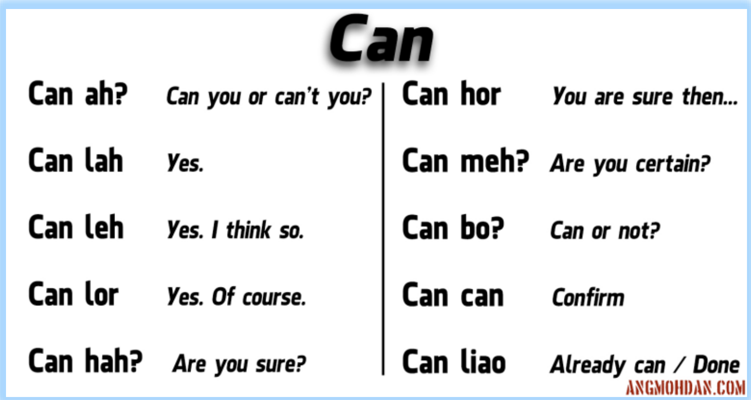
- Java Man, one of the first specimens of *Homo erectus* to be discovered
- JavaScript, an interpreted programming language
- Javan (disambiguation)
- Javanese (disambiguation)
- Jawa (disambiguation)
- Jaffa (disambiguation)

*This disambiguation page lists articles associated with the title **Java**.*

What is the meaning: I made her duck

- **Some of the possible meanings**
  - I cooked waterfowl for her
  - I cooked waterfowl belonging to her
  - I created the (plaster?) duck she owns
  - I caused her to quickly lower her head and body
  - I waved my magic wand and turned her into undifferentiated waterfowl

- A closer look:
  - "**duck**": can be a noun or verb
  - "**her**": can be a possessive pronoun ("of her") or dative pronoun ("for her")
  - "**make**": can mean "create" or "cook", and about 100 other things as well



Can

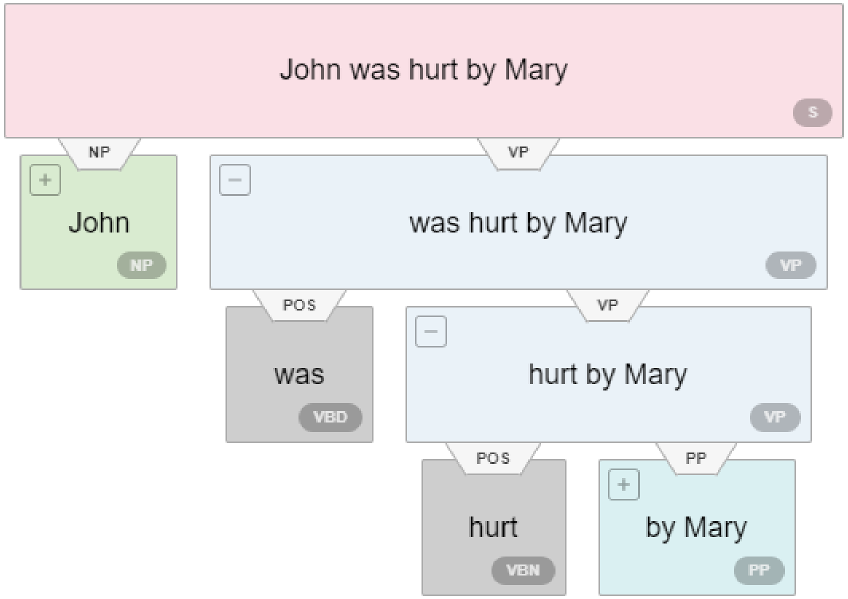
Can ah?	*Can you or can't you?*	Can hor	*You are sure then...*
Can lah	*Yes.*	Can meh?	*Are you certain?*
Can leh	*Yes. I think so.*	Can bo?	*Can or not?*
Can lor	*Yes. Of course.*	Can can	*Confirm*
Can hah?	*Are you sure?*	Can liao	*Already can / Done*

ANGMOHDAN.COM

More specific tasks

- Word-level: Locate all verbs and verbs only
  - "the tower collapsed as a result of safety violations"
  - Is 'result' here a noun or a verb?

- Syntactic-level:
  - Answer: "Who hurt John?"
  - Given:
  - "Mary hurt John."
  - "John was hurt by Mary."
  - "The guy who loved Mary hurt John."
  - "Mary is not sure of who hurt John."



More specific tasks

- Semantic level:
  - Answer: "Who killed John?"
  - Given: "Mary assassinated John"

  - Answer: "Who snores?"
  - Given: "Everyone who smokes snores, and John smokes."

- Discourse level:
  - Answer: "Who killed John?"
  - Given: Mary threw John into sea. [some other sentences] He drowned.

Course Web Page: NTULearn

- Lecture notes, tutorials, announcements, etc.

- Slides cannot replace the textbook/reading materials
  - They are at most a guideline.