Course Overview

He He

NEW YORK UNIVERSITY

January 24, 2023

Table of Contents

Logistics

A brief history

Challenges in NLP

Logistics

He He (lecturer)

Nitish Joshi (TA)

Aditya Kashilkar (grader)

Aayush Agrawal (grader)

- Best way to communicate with us: **Campuswire** (link and code on Brightspace).
- Office hours will be on Zoom (details on website), or in person by appointment.
- Midterm will be online through Gradescope.
- Let us know if you have accessibility needs.

What this course is (not) about

- It’s not about specific NLP applications (QA, dialogue etc.)
  - Unified approaches to various NLP problems
  - Hands-on experience in building NLP systems through assignments and the course project

- It’s not about fundamental machine learning
  - Focus on unique challenges in language data
  - Formalize NLP tasks as statistical learning problems

- Changes from previous offerings
  - − Structured prediction (HMM, PCFG, CRF)
  - + Large language models

What we expect you to know

- **Linear algebra**: vector space, vector norm, dot product, gradient etc.
- **Probability and statistics**: conditional probability, expectation, Bayes rule etc.
- **Basic machine learning**: loss function, gradient descent, logistic regression etc.
- **Programming**: read and write Python code, use Numpy, HPC, and deep learning libraries (Pytorch, Huggingface etc.)

Course project

An important component of the course (more on this later)

- Related to NLP (doesn’t have to be in the scope of this course)
- New algorithms or models for existing problems
- Applications of NLP or ML techniques to a problem
- Analysis of well-known approaches that leads to new insight
- ML Reproducibility Challenge 2021 (`https://paperswithcode.com/rc2021`)

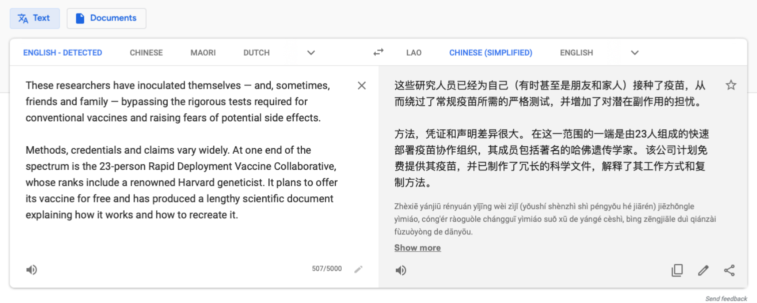
Table of Contents

Logistics

A brief history

Challenges in NLP

Products powered by NLP technologies



These researchers have inoculated themselves — and, sometimes, friends and family — bypassing the rigorous tests required for conventional vaccines and raising fears of potential side effects.

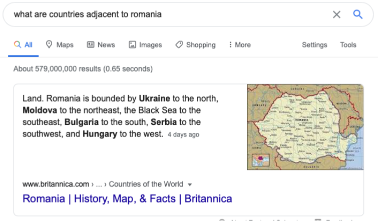
Methods, credentials and claims vary widely. At one end of the spectrum is the 23-person Rapid Deployment Vaccine Collaborative, whose ranks include a renowned Harvard geneticist. It plans to offer its vaccine for free and has produced a lengthy scientific document explaining how it works and how to recreate it.

507/5000

这些研究人员已经为自己（有时甚至是朋友和家人）接种了疫苗，从而绕过了常规疫苗所需的严格测试，并增加了对潜在副作用的担忧。

方法，凭证和声明差异很大。 在这一范围的一端是由23人组成的快速部署疫苗协作组织，其成员包括著名的哈佛遗传学家。 该公司计划免费提供其疫苗，并已制作了冗长的科学文件，解释了其工作方式和复制方法。

Zhèxiē yánjiū rényuán yǐjīng wèi zìjǐ (yǒushí shènzhì shì péngyǒu hé jiārén) jiēzhǒngle yìmiáo, cóng'ér ràoguòle chángguī yìmiáo suǒ xū de yángé cèshì, bìng zēngjiāle duì qiánzài fùzuòyòng de dānyōu.



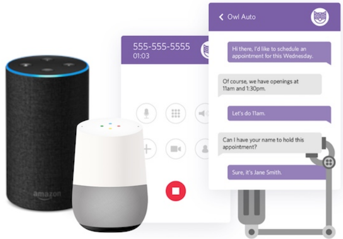
what are countries adjacent to romania

About 579,000,000 results (0.65 seconds)

Land. Romania is bounded by **Ukraine** to the north, **Moldova** to the northeast, the Black Sea to the southeast, **Bulgaria** to the south, **Serbia** to the southwest, and **Hungary** to the west. 4 days ago

www.britannica.com › ... › Countries of the World

Romania | History, Map, & Facts | Britannica



Hi there, I'd like to schedule an appointment for this Wednesday.

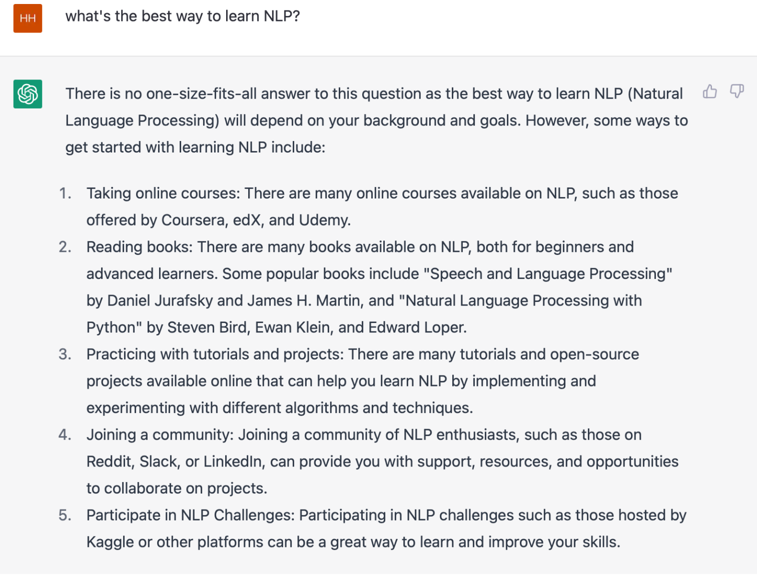
Of course, we have openings at 11am and 1:30pm.

Let's do 11am.

Can I have your name to hold this appointment?

Sure, it's Jane Smith.

A single natural language interface for everything

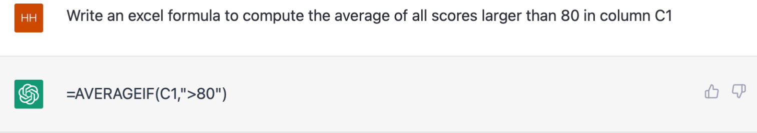


what's the best way to learn NLP?

There is no one-size-fits-all answer to this question as the best way to learn NLP (Natural Language Processing) will depend on your background and goals. However, some ways to get started with learning NLP include:

1. Taking online courses: There are many online courses available on NLP, such as those offered by Coursera, edX, and Udemy.
2. Reading books: There are many books available on NLP, both for beginners and advanced learners. Some popular books include "Speech and Language Processing" by Daniel Jurafsky and James H. Martin, and "Natural Language Processing with Python" by Steven Bird, Ewan Klein, and Edward Loper.
3. Practicing with tutorials and projects: There are many tutorials and open-source projects available online that can help you learn NLP by implementing and experimenting with different algorithms and techniques.
4. Joining a community: Joining a community of NLP enthusiasts, such as those on Reddit, Slack, or LinkedIn, can provide you with support, resources, and opportunities to collaborate on projects.
5. Participate in NLP Challenges: Participating in NLP challenges such as those hosted by Kaggle or other platforms can be a great way to learn and improve your skills.

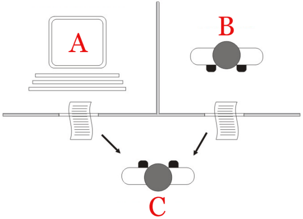
A single natural language interface for everything



Write an excel formula to compute the average of all scores larger than 80 in column C1

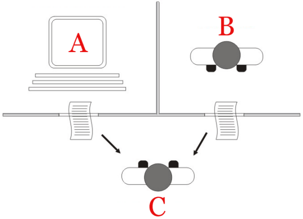
=AVERAGEIF(C1,">80")

Language is at the core of AI: the imitation game



"I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9*, to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning."* Turing (1950)

Language is at the core of AI: the imitation game



"I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9*, to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning."* Turing (1950)

🤔 Is humanlikeness the ultimate goal?

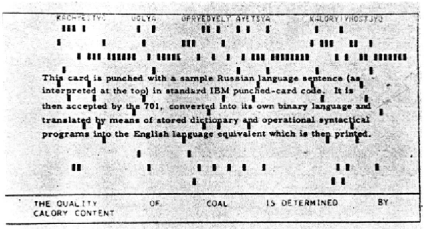
ELIZA

- Built by Joseph Weizenbaum at MIT in 1964 to demonstrate the *superficiality* of human-machine communication.
- Surprisinly, people were convinced that ELIZA had human intelligence.

Human: Well, my boyfriend made me come here.
ELIZA: Your boyfriend made you come here?
Human: He says I’m depressed much of the time.
ELIZA: I am sorry to hear you are depressed.
Human: It’s true. I’m unhappy.
ELIZA: Do you think coming here will help you not to be unhappy?

Early rule-based systems: the Georgetown-IBM experiment

- The Russian-English machine translation program:



KACHYESTVO UGLYA OPRYEDYELY AYETSYA KALORYIYNOSTJYU

This card is punched with a sample Russian language sentence (as interpreted at the top) in standard IBM punched-card code. It is then accepted by the 701, converted into its own binary language and translated by means of stored dictionary and operational syntactical programs into the English language equivalent which is then printed.

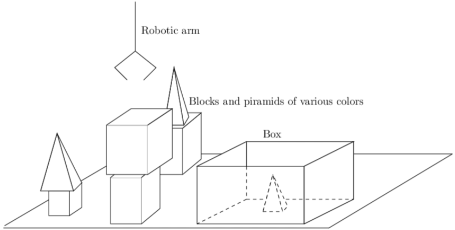
THE QUALITY OF COAL IS DETERMINED BY
CALORY CONTENT

- A vocabulary of 250 words
- Using 6 grammar rules, e.g.,

```
 If first code is 110, is third code associated with preceding
complete word equal to 21? If so, reverse order of appearance of
words in output (i.e., word carrying 21 should follow that carrying
110)---otherwise, retain order.
```

Approaching AI as a whole: SHRDLU

- Built by Terry Winograd at MIT in 1968.
- A person instructs the computer to build structures in a blocks world.
- Require many capabilities: grounding, coreference resolution, question answering, semantic parsing etc.



Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I don’t understand which pyramid you mean.
Person: (changing their mind) Find a block which is taller than the one you are holding and put it into the box.
Computer: By ”it”, i assume you mean the block which is taller than the one i am holding.

Limitations of early systems

- Optimism in the 50’s and 60’s: working on tasks that are too complex at that time

  “Within the very near future—much less than twenty-five years—we shall have the technical capability of substituting machines for any and all human functions in organizations.”

- Disappointing results due to
  - **Limited computation**: hardware has limited speed and memory
  - **Combinatorial explosion**: algorithms are intractable in realistic settings
  - **Underestimated complexity**: ambiguity, commonsense knowledge etc.

The rise of statistical learning in the 80's

- Notable progress in MT from IBM (neglected knowlege of linguistics).
- HMMs widely used for speech recognition.
  "Every time I fire a linguist, the performance of the speech recognizer goes up."—Frederick Jelinek.
- The paradigm shift: expert knowledge + rules → data + features
- Statistical learning is the main driving force of NLP today.

The deep learning tsunami

- Before deep learning (around 2015), NLP is mostly about structured prediction and feature engineering.
- Neural networks can automatically learn good features/representations for a task
- The paradigm shift: features $\rightarrow$ network architectures + embeddings
- Almost all NLP models are neural networks nowadays.

Models and data keep getting larger

- Since around 2018, Transformer-based pretrained models have become the standard.
- Pre-training on large data provides useful representations for many downstream tasks.
- The paradigm shift: architecture design → transfer learning (fine-tuning)
- More recently, a single natural language interface for all tasks (e.g., ChatGPT by OpenAI).
- The paradigm shift: transfer learning → instructing / prompting

Table of Contents

Logistics

A brief history

Challenges in NLP

Why is language hard?

Why is language hard?

- **Discrete**
  - How to define metrics?

I work at NYU.	vs	I work for NYU.
This is good.	vs	This is actually good.

  - How to define transformations?

The food is okay.	→	The food is awesome!
They made a brief return to Cambridge to drop the book.	→	They returned.

  - In general, hard to represent text as mathematical objects.

Why is language hard?

- **Compositional**
  - The whole is built from parts (chars, words, sentences, paragraphs, documents...)
  - How to generalize when we don't see all possible combinations?
  - An example from [Lake et al., 2018]

    Vocabulary:

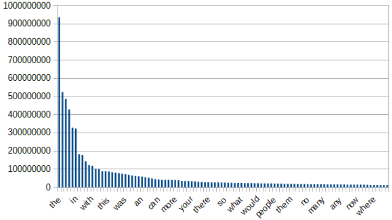
    $$\{\text{jump, walk, turn, once, twice, left, right, before, after, and}\}$$

    Sentences:

    jump
    jump left
    jump left and walk right
    jump left after walk right once before turn left twice
    ...

Why is language hard?

- **Sparse**
    - How to handle the long tail?
    - Zipf's law: word frequency $\propto \frac{1}{\text{rank}}$



    - Many linguistic phenomena follow Zipf's law
      BoA's financial assistant Erica:
      The bank "learned [that] there are over 2,000 different ways to ask us to move money."[1]

[1] https://www.aiqudo.com/2019/06/28/voice-success-story-erica-bank-america/

Why is language hard?

- **Ambiguous**
  - How to interpret meaning in context?

    Bass: fish? guitar? frequency? (word sense disambiguiation)

    I shot an elephant in my pajamas: who is in the pajamas? (PP attachment)

    The spirit is willing but the flesh is weak.
    → The vodka is strong but the meat is rotten.