



Introduction to NLP

CS 780/880 Natural Language Processing

Samuel Carton, University of New Hampshire

This lecture

1. Introductions
2. What is NLP?
3. Why should you care about NLP
4. Structure of this course



Who am I

This guy:



My research: human-centered and interpretable NLP

- Why do the models do what they do and how can people work with them more effectively?



What is NLP?



What is NLP?

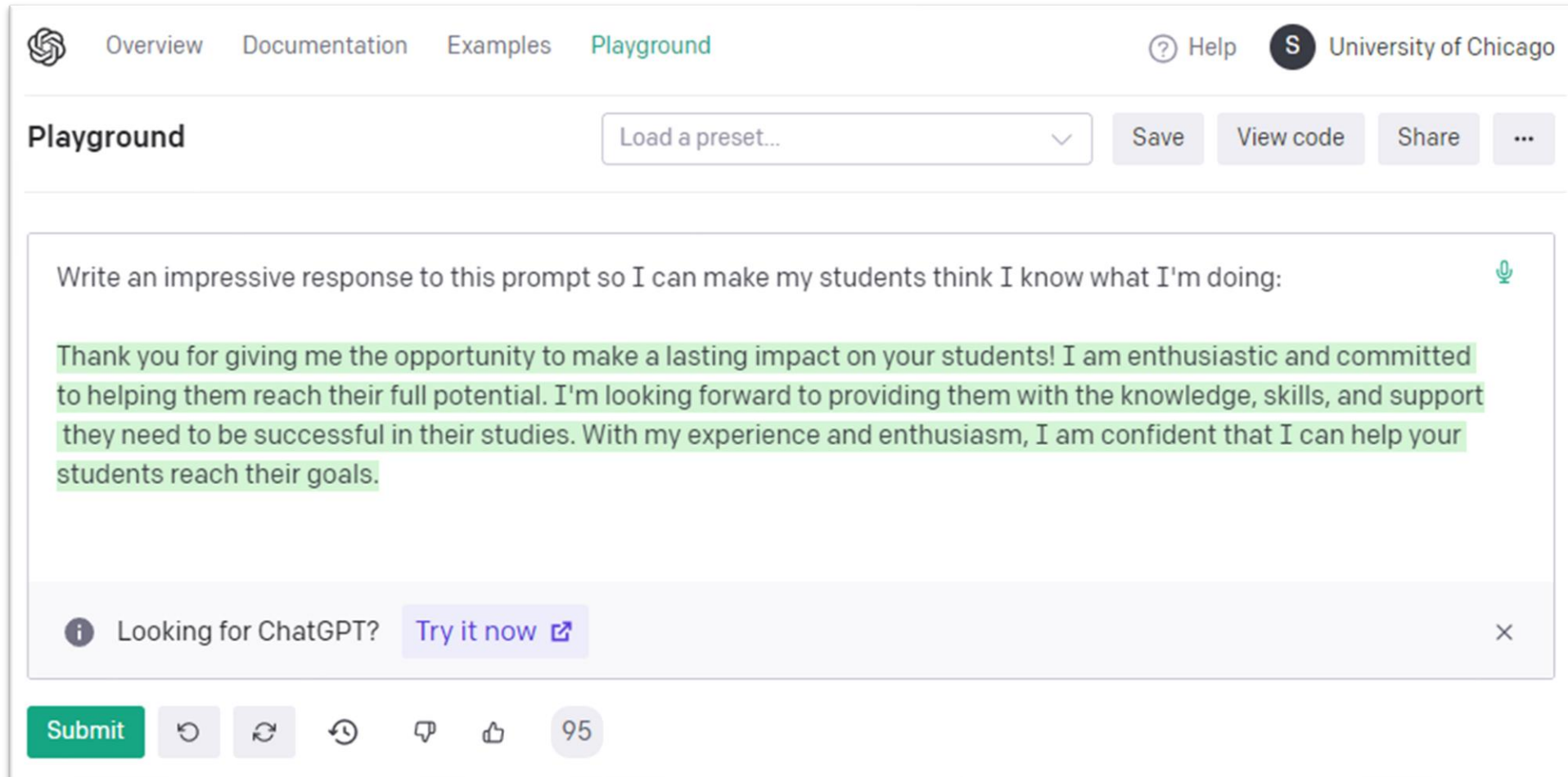
Natural Language Processing

Wikipedia: “an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language”

Me: any attempt to represent and work with written text numerically or statistically



NLP example: chat bot



The screenshot shows the OpenAI Playground interface. At the top, there are navigation links: Overview, Documentation, Examples, and Playground (highlighted). On the right, there is a Help icon and a user profile for 'University of Chicago'. Below the navigation, the 'Playground' title is followed by a 'Load a preset...' dropdown menu and buttons for 'Save', 'View code', 'Share', and a three-dot menu. The main chat area contains a prompt: 'Write an impressive response to this prompt so I can make my students think I know what I'm doing:'. The model's response is highlighted in green: 'Thank you for giving me the opportunity to make a lasting impact on your students! I am enthusiastic and committed to helping them reach their full potential. I'm looking forward to providing them with the knowledge, skills, and support they need to be successful in their studies. With my experience and enthusiasm, I am confident that I can help your students reach their goals.' At the bottom of the chat area, there is a notification: 'Looking for ChatGPT? Try it now' with a link icon and a close button. Below the chat area, there is a 'Submit' button and a row of icons for undo, redo, refresh, copy, and thumbs up, along with a '95' score indicator.



NLP example: chat bot detector

The screenshot shows the Writer AI Content Detector interface. The browser address bar displays `writer.com/ai-content-detector/`. The page features a navigation menu with links for Product, Use cases, Pricing, Customers, and Resources, along with buttons for 'Try for free' and 'Sign in'. The main heading is 'AI Content Detector', followed by a sub-heading 'AI Content Detector' and a paragraph explaining that search engines penalize pages ranking if they detect AI-generated content. The interface includes a text input area with a URL field containing 'https://', a text area containing a paragraph of text, and a character count of '380/1500 CHARACTERS'. A prominent 'Analyze text' button is visible. On the right, a large '15%' score is displayed, indicating the percentage of human-generated content. Below the score is a progress bar and a message: 'You should edit your text until there's less detectable AI content.' The Writer logo is in the top left, and a 'Privacy - Terms' link is in the bottom left.

WRITER

Product ▾ Use cases ▾ Pricing Customers Resources ▾ [Try for free](#) [Sign in](#)

AI Content Detector

Some search engines **penalize** your page **ranking** if **they detect content that reads like it was entirely produced by AI**. **Evaluate your text with this detector** and decide if you want to make adjustments before you publish your content.

380/1500 CHARACTERS [Analyze text](#)

15%

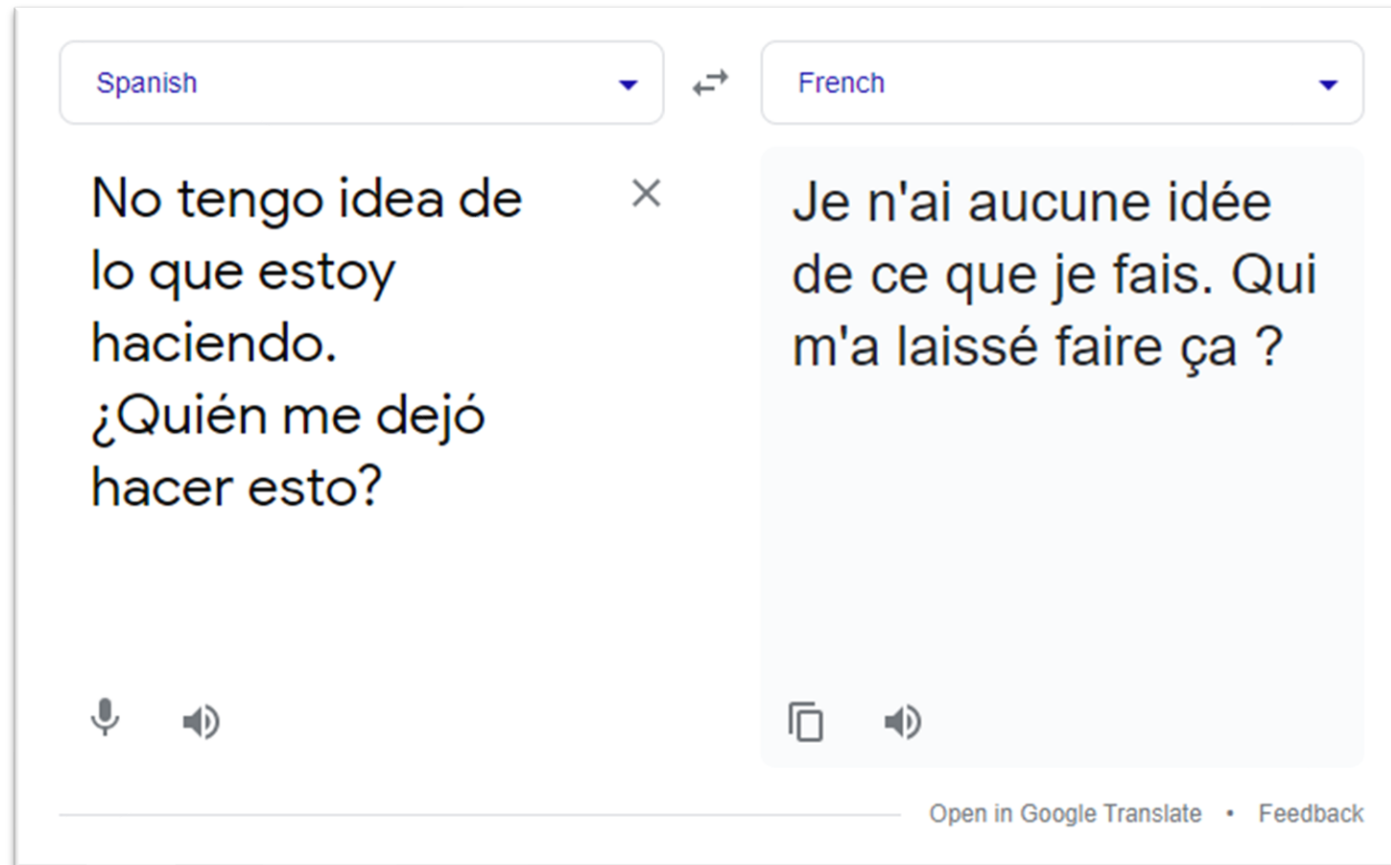
HUMAN-GENERATED CONTENT

You should edit your text until there's less detectable AI content.

[Privacy - Terms](#)



NLP example: machine translation



The screenshot displays the Google Translate interface. On the left, the source language is set to Spanish, and the text reads: "No tengo idea de lo que estoy haciendo. ¿Quién me dejó hacer esto?". On the right, the target language is set to French, and the translated text reads: "Je n'ai aucune idée de ce que je fais. Qui m'a laissé faire ça ?". Below the text are icons for voice input/output and a copy icon. At the bottom, there are links for "Open in Google Translate" and "Feedback".

Spanish

French

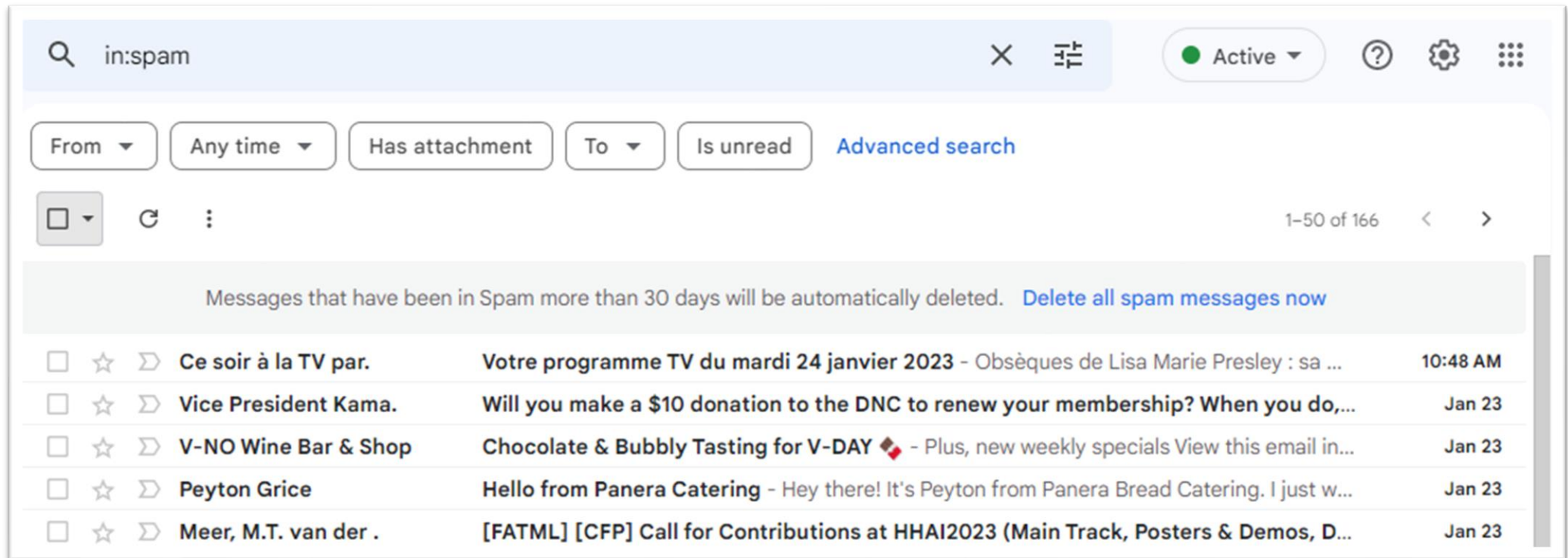
No tengo idea de lo que estoy haciendo. ¿Quién me dejó hacer esto?

Je n'ai aucune idée de ce que je fais. Qui m'a laissé faire ça ?

Open in Google Translate • Feedback



NLP example: spam detection

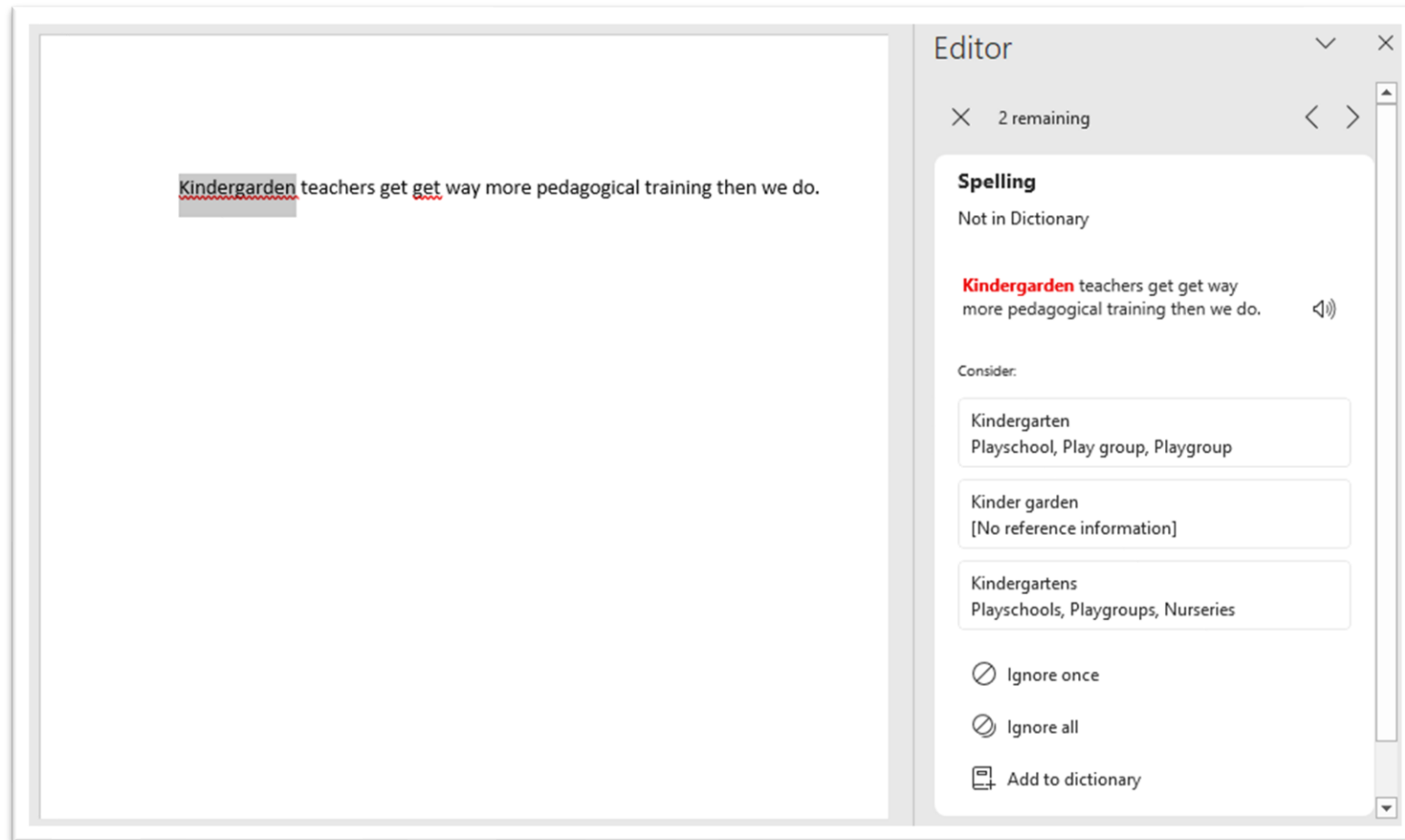


The screenshot shows an email inbox with a search bar at the top containing "in:spam". Below the search bar are several filter buttons: "From", "Any time", "Has attachment", "To", and "Is unread", along with an "Advanced search" link. A notification bar states: "Messages that have been in Spam more than 30 days will be automatically deleted. [Delete all spam messages now](#)". The inbox list contains five messages, each with a checkbox, star, and envelope icon. The messages are:

Sender	Subject	Date
Ce soir à la TV par.	Votre programme TV du mardi 24 janvier 2023 - Obsèques de Lisa Marie Presley : sa ...	10:48 AM
Vice President Kama.	Will you make a \$10 donation to the DNC to renew your membership? When you do,...	Jan 23
V-NO Wine Bar & Shop	Chocolate & Bubbly Tasting for V-DAY - Plus, new weekly specials View this email in...	Jan 23
Peyton Grice	Hello from Panera Catering - Hey there! It's Peyton from Panera Bread Catering. I just w...	Jan 23
Meer, M.T. van der .	[FATML] [CFP] Call for Contributions at HHAI2023 (Main Track, Posters & Demos, D...	Jan 23



NLP example: grammar and spell-checker

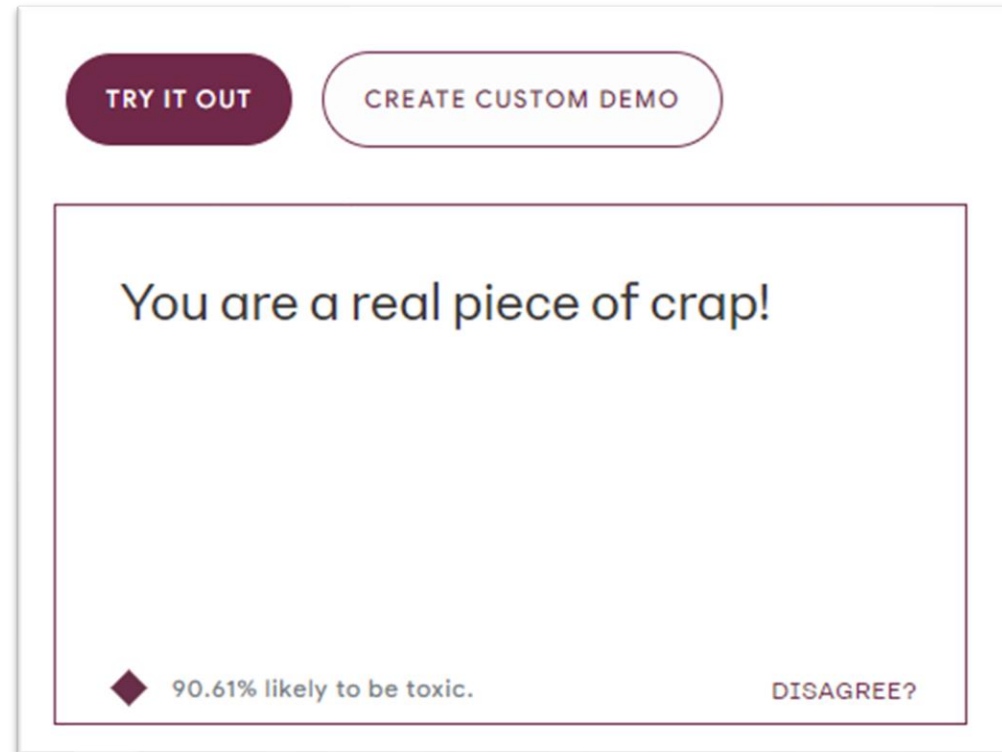


The image shows a screenshot of a text editor window titled "Editor". The main text area contains the sentence: "Kindergarden teachers get get way more pedagogical training then we do." The word "Kindergarden" is highlighted with a grey background and has a red wavy underline underneath it. To the right of the text area is a sidebar panel titled "Editor" with a close button (X) and a "2 remaining" indicator. The sidebar is titled "Spelling" and shows "Not in Dictionary". Below this, the word "Kindergarden" is listed in red, followed by the sentence "Kindergarden teachers get get way more pedagogical training then we do." with a speaker icon. Underneath, it says "Consider:" and lists three suggestions in rounded rectangular boxes: "Kindergarten" (with sub-suggestions "Playschool, Play group, Playgroup"), "Kinder garden" (with "[No reference information]"), and "Kindergartens" (with sub-suggestions "Playschools, Playgroups, Nurseries"). At the bottom of the sidebar are three options: "Ignore once", "Ignore all", and "Add to dictionary", each with a radio button.



NLP example: hate speech detection

perspectiveapi.com



The screenshot shows the Perspective API interface. At the top, there are two buttons: "TRY IT OUT" (a dark red button) and "CREATE CUSTOM DEMO" (a white button with a red border). Below the buttons is a large text input area containing the sentence "You are a real piece of crap!". At the bottom of the input area, there is a red diamond icon followed by the text "90.61% likely to be toxic." and a "DISAGREE?" link.



NLP example: information extraction

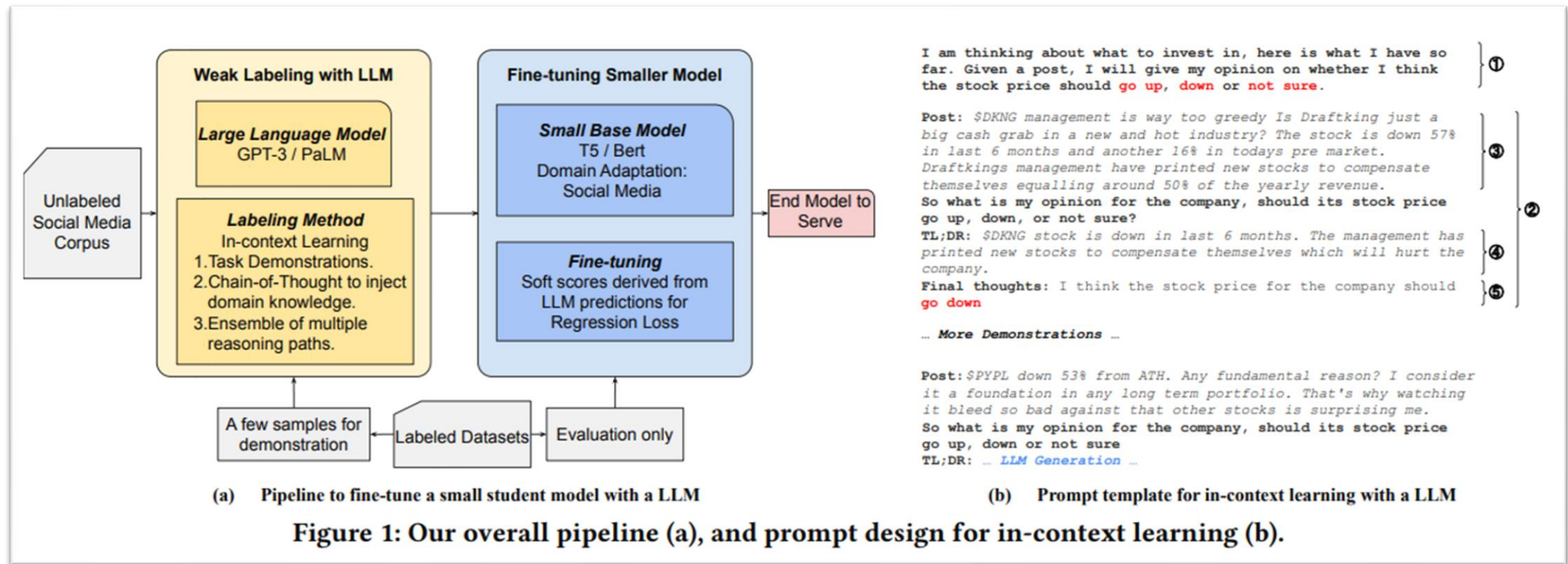
P2- Na₂/3Ni₁/4Ti_xMn₃/4-xO₂ was prepared through a simple solid state method. The precursor solution was prepared by mixing desirable amount of Ni(CH₃COO)₂*4H₂O, Mn(CH₃COO)₂*4H₂O and CH₃COONa and titanium citrate solution. The obtained mixture was heated at 400 degC for 12 h. The ground powder was ball-milled for 1 h and was subsequently calcinated at 900 degC in air for 12 h to synthesize Na₂/3Ni₁/4Ti_xMn₃/4-xO₂ (x=0, 0.05, 0.10, 0.15, 0.20, 0.30).

Figure 1: Part of an example synthesis procedure included in the dataset with entity annotations from Zhao et al. (2015). Colors represent entity types and underlines represent span boundaries. Colors: Target, Nonrecipe-operation, Unspecified-Material, Operation, Material, Condition-Unit, Number.

Tim O’Gorman, Zach Jensen, Sheshera Mysore, Kevin Huang, Rubayat Mahbub, Elsa Olivetti, and Andrew McCallum. 2021. MS-Mentions: Consistently Annotating Entity Mentions in Materials Science Procedural Text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*



NLP example: predicting the stock market



Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2022. What do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis. 0 citations (Semantic Scholar/arXiv) [2023-01-24] arXiv:2212.11311 [cs].



NLP vs. machine learning

A few basic types of data:

- Text → NLP
- Images → Computer Vision (CS 855)
- Tabular data
- Time series
- Sound

NLP: machine learning as applied to text (?)

Other miscellaneous terms: “Data science”, “data mining”, “text mining”, “text analytics”



NLP vs. linguistics

Linguistics: study of the structure of language and comparative differences between languages

In the past: used to inform the design of NLP systems

- Recursive neural nets (Socher et al. 2013) a great example

Now: increasingly irrelevant to design of NLP models

- Though NLP can still be used to study it!



NLP vs. information retrieval (IR)

Lots of overlap with NLP, but its own field with its own practices and conferences

- Also, doesn't only retrieve text

Converging with NLP as “Neural IR” comes to dominate



Okay, but what **is** NLP

Mostly representations & models

How to **represent** text numerically so that computers can work with it

Models that can learn from humans how to do stuff with text

- Generate it
- Classify it
- Translate it
- Understand it
- Transform it
- Etc.



Why you should care



Why should you care

1. Super cool

2. Lots of job and research opportunities

3. Hottest area of AI; taking over the world; indelibly transforming the future; nothing will ever be the same again; what is happening right now; I want to get off Mr. Bones wild ride; etc.



ChatGPT

Reddit posts from **today**, January 24, 2023

30  31.2k   **ChatGPT passes MBA exam given by a Wharton professor.** Artificial Intelligen... (nbcnews.com)
submitted 21 hours ago by SUPRVLLAN to r/technology
2097 comments share save hide give award report crosspost

5  170   **ChatGPT got 50% more marks on data science assignment than me. What's next?** Discussion (self.datascience)
submitted 7 hours ago by rifat_monzur to r/datascience
141 comments share save hide give award report crosspost

35  129   **Universities offered software that detects ChatGPT essays** Artificial Intelligen... (theregister.com)
submitted 9 hours ago by Parking_Attitude_519 to r/technology
60 comments share save hide give award report crosspost

86  2621   **Scores of Stanford students used ChatGPT on final exams** Artificial Intelligen... (stanforddaily.com)
submitted 1 day ago by Parking_Attitude_519 to r/technology
727 comments share save hide give award report crosspost

ChatGPT is at
capacity right now

[Get notified when we're back](#)

Write a guided meditation about the status of ChatGPT.

As you read this message, please take a moment



ChatGPT

Simple idea, same as any language model: pick the most likely next word based on the previous words

But models are getting **so good** that we're moving toward everyone having a personal secretary in their pocket. It can write:

- Essays for school
- Text for written take-home exams
- Code for assignments
- Emails
- Poetry

Just think about how this is going to change society!



GPT-3

Write a poem from the perspective of a potato about what it is like being fried:

Frying in oil, so hot and so searing
My skin is crisp, a golden brown hue
The aroma of my cooking fills the air
As I transform from tuber to stew.

My insides are tender and creamy white
My outside is crunchy and so delight
The heat of the oil is overwhelming
But I'm not one to ever fight.

My journey has been a long one
From the ground to the fryer
I've seen many things along the way
But I'm still in for a bit of surprise.



GPT-3

Write a recommendation letter for a student in my natural language processing class who got a good grade, participated very enthusiastically, and did an excellent final project about detecting hate speech on social media:

To Whom It May Concern:

I am writing to provide a recommendation for [Student Name], who recently took my Natural Language Processing course.

[Student Name] was an outstanding student throughout the course. He received excellent grades and was very enthusiastic in class discussions and activities. His final project was particularly impressive; he created an algorithm that was able to detect hate speech on social media. His creativity and dedication to the course was evident in the results.

I highly recommend [Student Name] for any role that requires advanced knowledge of natural language processing. He has a strong ability to understand and apply concepts and theories, and his creativity and enthusiasm for the subject make him an ideal candidate for any position that requires these skills.

Sincerely,

[Your Name]



GPT-3

Write a short essay about the relative merits on Kantian versus utilitarian ethics in medical decision-making:

Kantian and utilitarian ethics are two prominent ethical theories that are often used to guide medical decision-making. Both theories have advantages and disadvantages, so it is important to consider both when making medical decisions.

Kantian ethics are based on the concept of universalizability; that is, any action taken should be one that could be universally applied to all persons in a similar situation. This means that medical decisions should be made on the basis of respecting the autonomy of the patient and considering the ethical implications of each action. Utilitarian ethics, on the other hand, are based on the principle of maximizing the overall good of society. This means that medical decisions should be made with the aim of achieving the greatest good for the greatest number of people.

Kantian ethics have the advantage of emphasizing the importance of respecting the autonomy of the patient. This means that medical decisions are made with the interests of the patient in mind, which is important for ensuring that a patient's rights and dignity are respected. On the other hand, utilitarian ethics can be criticized for being overly focused on the good of society rather than the well-being of the individual patient. This can lead to medical decisions being made that are not in the best interests of the patient.

In conclusion, while both Kantian and utilitarian ethics have advantages and disadvantages, it is important to consider both when making medical decisions. By considering both perspectives, medical professionals can ensure that decisions are made in the best interests of the patient, while also having regard for the good of society.



GPT-3

Write an implementation of Bubble sort in Python using bird names for all the variable names:

```
def robin_sort(pigeon_list):
    # Iterate through the list of pigeons
    for sparrow in range(len(pigeon_list) - 1):
        # Compare adjacent elements
        for eagle in range(len(pigeon_list) - sparrow - 1):
            # If current element is greater than the next element, swap them
            if pigeon_list[eagle] > pigeon_list[eagle + 1]:
                blue_jay = pigeon_list[eagle]
                pigeon_list[eagle] = pigeon_list[eagle + 1]
                pigeon_list[eagle + 1] = blue_jay

    # Return the sorted list
    return pigeon_list
```



Research opportunities

<input type="checkbox"/>	☆	▷	ioanna...@gmail.com	ML News/.../ACL Portal	[ML-news] AIDA Short Courses: "Nvidia DLI – Fundamentals of Deep Learning", 2nd February 2023 - vision and natural language processing. Yo...	3:40 AM
<input type="checkbox"/>	☆	▷	E. Niemann UKP Lab	ML News/.../ACL Portal	[ML-news] Science Manager (Postdoc) in AI and Natural Language Processing, UKP Lab, TU Darmstadt, Germany - Ubiquitous Knowledge Proc...	Jan 23
<input type="checkbox"/>	☆	▷	Rainer Schmidt	ML News/.../ACL Portal	[AISWorld] CFP Workshop on Social and Human Aspects of Business Process Management (BPMS2'23) - mining, natural language processing, la...	Jan 23
<input type="checkbox"/>	☆	▷	Nuno Moniz	ML News/.../ACL Portal	[ML-news] EPIA 2023 - 22nd EPIA Conference on AI (Azores, Portugal) - TeMA - Natural Language Processing, Text Mining and Applications PSDM...	Jan 22
<input type="checkbox"/>	☆	▷	Rémi Eyraud	ML News/.../ACL Portal	[ML-news] TAYSIR Competition: Transformers+RNN: Algorithms to Yield Simple and Interpretable Representations - domains like natural langua...	Jan 21
<input type="checkbox"/>	☆	▷	Caio Filippo Corro	ML News/.../ACL Portal	[ML-news] CFP - TAL journal - special issue "Robustness and limitations of natural language processing models" - limitations of natural langua...	Jan 20
<input type="checkbox"/>	☆	▷	José G. Moreno	ML News/.../ACL Portal	[ML-news] Postdoc position / Multi-modal Pretrained Language Models for Health Care - modal Pretrained Language Models for Health Care == ...	Jan 20
<input type="checkbox"/>	☆	▷	Aris Filos-Ratsikas	ML News/.../ACL Portal	[ML-news] ECAI-23 joint call for papers/Workshops/tutorials - Speech, Natural Language Processing and Understanding (NLP) - Search, Planning ...	Jan 20
<input type="checkbox"/>	☆	▷	E. Niemann UKP Lab	ML News/.../ACL Portal	[ML-news] Postdoctoral Research Fellow in Computational Linguistics/Natural Language Processing at MBZUAI, UAE; additionally affiliated ...	Jan 20
<input type="checkbox"/>	☆	▷	E. Niemann UKP Lab	ML News/.../ACL Portal	[ML-news] Postdoctoral Research Fellow in Computational Linguistics/Natural Language Processing at the UKP Lab, TU Darmstadt, Germany -	Jan 20
<input type="checkbox"/>	☆	▷	Gargi Gupta	ML News/.../ACL Portal	[ML-news] [ML-News] 1st World Conference on Explainable AI - call for papers - of signal processing methods Finite state machines for explaina...	Jan 20
<input type="checkbox"/>	☆	▷	Frenay Benoit	ML News/.../ACL Portal	[ML-news] ML applied to Sign Language: Special Session at ESANN'23 - to Sign Language" at ESANN 2023* European Symposium on Artificial Ne...	Jan 20
<input type="checkbox"/>	☆	▷	hady...@gmail.com	ML News/.../ACL Portal	[ML-news] #TheWebConf 2023 - Call for Contributions to 22 Workshops - Workshop on Natural Language Processing for Social Media (SocialNLP...	Jan 19
<input type="checkbox"/>	☆	▷	Raoul Koudijs	ML News/.../ACL Portal	[AISWorld] [CFP] Call for Papers: 7th International Joint Conference on Rules and Reasoning (RuleML+RR 2023) - approaches to natural langua...	Jan 18
<input type="checkbox"/>	☆	▷	rafika boutalbi	ML News/.../ACL Portal	[ML-news] [CFP] 1st ICDAR International workshop on Machine vision and NLP for Document Analysis (VINALDO) - for image processing are a...	Jan 18
<input type="checkbox"/>	☆	▷	Jennifer D'Souza	ML News/.../ACL Portal	[ML-news] [1st Call] Call for Workshop & Tutorial Proposals SEMANTICS 2023 - Policies - Natural Language Processing and Computational Lingui...	Jan 18
<input type="checkbox"/>	☆	▷	alfredo vellido	ML News/.../ACL Portal	[ML-news] [CFP] Special Session IJCNN 2023 - The Coming of Age of Explainable AI (XAI) and ML - Logic and Natural Language Processing Imp...	Jan 18
<input type="checkbox"/>	☆	▷	lacer Calixto	ML News/.../ACL Portal	[ML-news] SemEval Task on Visual Word Sense Disambiguation (Visual-WSD): Evaluation starts! - vision and language models, as well as other s...	Jan 17

<https://groups.google.com/g/ml-news>



Research opportunities






Search interface for research opportunities. The search criteria include "natural language processing" and "Location". The search results show 37 matching records. The results are displayed in a table with columns for the institution, job title, location, and job type.

Search criteria: "natural language processing", Location, Remote positions only, Choose a category...

Search Jobs

Professional Postdoc

Search completed. Found 37 matching records. [RSS](#) [Reset](#)

 Instructional Professor – Computational Social Science <i>MA Programs - Social Sciences Division</i>	The University of Chicago Chicago, Illinois	Professional 4 days ago
 Tenure-Track and Research Faculty Positions <i>Department of Computer Science</i>	Toyota Technological Institute at Chicago Chicago, Illinois	Professional 1 week ago
 Instructional Professor (open rank) in Data Science <i>The Data Science Institute</i>	University of Chicago Chicago, Illinois	Professional 1 month ago
 Assistant Professor, Tenure Track, Data Science <i>iSchool</i>	University of Arizona Tucson, Arizona	Professional 1 month ago
 Tenure-Stream Faculty Positions <i>SCI-Department of Computer Science</i>	University of Pittsburgh Pittsburgh, PA	Professional 1 month ago


<https://cra.org/ads/>





Job opportunities


Jobs based on your profile


151 results

**Senior Scientist**
Pandora
New York, NY (Remote)
\$135K/yr - \$170K/yr (from job description)
1 company alumni
15 minutes ago

**Summer 2023: Reinforcement Learning Intern (PhD)**
iRobot
Bedford, MA (On-site)
3 school alumni
Promoted

**Unqork Engineer**
Accenture
Boston, MA
1 connection
Promoted · 4 applicants

**Principal Software Engineer - Remote**
Liberty Mutual Insurance
Dover, NH
Your profile matches this job
Promoted · 1 applicant

**Data Science Specialist**
Harvard University

Who We Are

SiriusXM and its brands (Pandora, Stitcher, SXM Media, AdsWizz, Simplecast, and SiriusXM Connected Vehicle Services) are leading a new era of audio entertainment and services by delivering the most compelling subscription and ad-supported audio entertainment experience for listeners -- in the car, at home, and anywhere on the go with connected devices. Our vision is to shape the future of audio, where everyone can be effortlessly connected to the voices, stories and music they love wherever they are.

This is the place where a diverse group of emerging talent and legends alike come to share authentic and purposeful songs, stories, sounds and insights through some of the best programming and technology in the world. Our critically-acclaimed, industry-leading audio entertainment encompasses music, sports, comedy, news, talk, live events, and podcasting. No matter their individual role, each of our employees plays a vital part in bringing SiriusXM's vision to life every day.

SiriusXM is the leading audio entertainment company in North America, and the premier programmer and platform for subscription and digital advertising-supported audio products. SiriusXM's platforms collectively reach approximately 150 million listeners, the largest digital audio audience across paid and free tiers in North America, and deliver music, sports, talk, news, comedy, entertainment and podcasts. Pandora, a subsidiary of SiriusXM, is the largest ad-supported audio entertainment streaming service in the U.S. SiriusXM's subsidiaries Stitcher, Simplecast and AdsWizz make it a leader in podcast hosting, production, distribution, analytics and monetization. The Company's advertising sales organization, which operates as SXM Media, leverages its scale, cross-platform sales organization and ad tech capabilities to deliver results for audio creators and advertisers. SiriusXM, through Sirius XM Canada Holdings, Inc., also offers satellite radio and audio entertainment in Canada. In addition to its audio entertainment businesses, SiriusXM offers connected vehicle services to automakers.

How You'll Make An Impact

As a Senior Scientist on the Content Science team, you will build products and services to understand, organize, and recommend music and non-music content reaching an audience of 100M+ listeners across SiriusXM, Pandora and Stitcher. You will join a team of scientists with diverse expertise in machine learning, natural language understanding, recommendation systems, and statistics.

<https://www.linkedin.com>

The logo for the University of New Hampshire, featuring a blue shield with a white 'N' and 'H' inside.

28

Intellectually interesting

My interest: understand the hidden patterns in written text and communication

Want to major in English? Do NLP instead.



This course



Basic structure

Syllabus: <https://shcarton.github.io/cs780spring2023/>

Structure

- Biweekly lecture
- Attendance “required”
- Office hours after class on Tuesday

Assignments

- Weekly homeworks
- Final project with milestone & poster presentation
- (Maybe) mock interview
- No midterm/final



Homework assignments

Written and completed in Google Colab (Python)

Individual work*

Late policy: -10% per day, up to 5 days later

Further details TBD



Topics

Compare/contrast with CMU

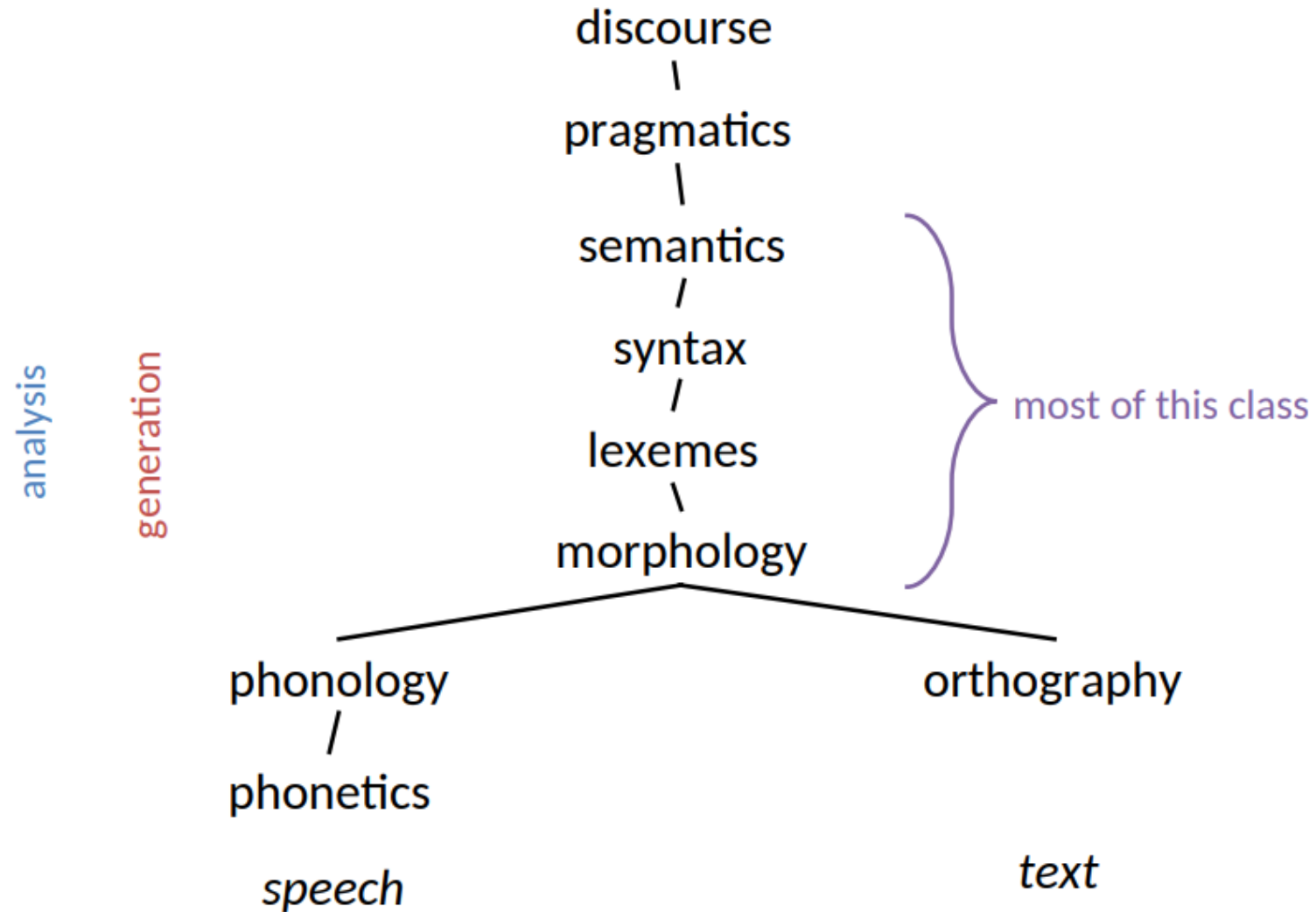
- Basic NLP: <http://demo.clab.cs.cmu.edu/NLP/>
- Advanced NLP: <http://www.phontron.com/class/anlp2022/>
- Neural nets for NLP: <https://www.phontron.com/class/nn4nlp2021/>

This course

- Very little linguistics
- Almost entirely ML
- First 4 weeks: classical statistical/probabilistic models
- Next 7 weeks: neural methods up through transformer-based models(e.g. ChatGPT)
- Last few weeks: special topics (interpretability, fairness, etc.)



Levels of linguistic representation



Rec letters & research opportunities

Happy to write letters and supervise research projects—part of my job!

Recommendation letters

- Come to class & participate
- Let me know at least 6 weeks before you need the letter
 - Same goes for all letter writers
 - Ideally let me know now-ish so I can pay special attention to you

Research

- No summer support for the time being
- Independent study or volunteer basis

