NLP For Economists

Lecture 1: Course and NLP overview

Sowmya Vajjala

MGSE - LMU Munich
Guest Course, October 2022

5th October 2022

About Us

Course Overview

NLP Overview

Representing Text as a Vector

NLP in Economics: An Overview

About me

- I work as a full time researcher at the National Research Council, Canada in Digital Technologies Research Center.
- 2020: Wrote a book on NLP for O'Reilly (http://www.practicalnlp.ai/)
- 2018-19: Senior Data Scientist in software engineering companies in Toronto
- 2016-18: Assistant Professor (tenure track) at Iowa State University, USA
- 2011-15: PhD student at University of Tuebingen, Germany

Teaching experience

- 2021: NLP without a readymade annotated dataset, Hauptseminar at University of Tuebingen (online).
- 2020: this course, first edition https://econnlpcourse.github.io/
- 2016-18: Introductory and advanced courses in python programming, NLP and data science.

About you

- What do you do? (Masters/PhD in XXX)
- Why are you enrolled in this course?
- What do you want to do later?

Course Background

1. Natural Language Processing (NLP) is all about making computers understand and interact with humans, in their language(s).

Course Background

1. Natural Language Processing (NLP) is all about making computers understand and interact with humans, in their language(s).
2. NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.

Course Background

1. Natural Language Processing (NLP) is all about making computers understand and interact with humans, in their language(s).
2. NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.
3. NLP is also widely used as a research method in many disciplines, including economics.

Course Background

1. Natural Language Processing (NLP) is all about making computers understand and interact with humans, in their language(s).
2. NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.
3. NLP is also widely used as a research method in many disciplines, including economics.
4. From measuring news sentiment and connecting it to the nation's economy to extracting key information from policy documents, there are many uses of NLP in economics research.

Course Objectives

- Introduce you to some common NLP methods
- Illustrate how to use Python programs for NLP
- Understand how to approach a new NLP problem
- Explore the various use cases for NLP in Economics research

Expected Learning Outcomes

- An overview of various NLP methods.
- Given a problem description in economics research involving textual data, understand what methods from NLP can be used to solve the problem, and design a step by step process to solve it.

What the course can't do

- Don't expect to become an NLP expert with one compact course.
- Contents are not tailor made for individual preferences, but there is a group discussion and a short paper you can submit at the end, which gives you opportunities to explore your specific interests related to this topic.
- The course won't teach you programming in general.

Pre-requisities

1. some familiarity with Python, installing and using software on your computers
2. advantage: some familiarity with machine learning

How we learn and grow

आचार्यात् पादमादत्ते पादं शिष्यः स्वमेधया ।
सब्रह्मचारिभ्यः पादं पादं कालक्रमेण च ॥

One fourth from the teacher, one fourth from own intelligence,
One fourth from classmates, and one fourth only with time.

AchAryAt pAdamAdatte, pAdam shiShyaH swamedhayA |
sa-brahmachAribhyaH pAdam, pAdam kAlakrameNa cha ||

आचार्यात् पादमादत्ते पादं शिष्यः स्वमेधया ।
सब्रह्मचारिभ्यः पादं पादं कालक्रमेण च ॥

Source

Course Logistics

Meeting and Location

- This is an online course.
- Duration: 2 weeks
- Dates: 5th, 6th, 7th; 11th, 12th, 13th October; 3-6 pm CET (9-12 noon EST).
- Location: Zoom.
  Meeting ID: 942 6228 4672
  Passcode: 107380
- I will be available for another week afterwards for an optional review session or any one-on-one conversation (email me for a 1-1 meeting if needed).

Course Website

- https://econnlpcourse.github.io/
- Syllabus and reading list is available there.
- Lecture slides and any exercise/assignment descriptions will be uploaded there.

Course Format

- 5 Online (live) video lectures, which will be recorded (I haven't figured out how to share privately yet. If I don't find a way, I will delete them after the course).
- I will provide some practice exercises with lectures 2-5. They are not graded.
- 6th lecture is a session where you present in small groups of 2-4 people. Pick a paper from the reading list on the website (or any other relevant paper) and present a brief discussion in a live session (10-15 minutes per group).
- We can have an optional review session on 17th or 18th October, if needed.

Deadlines, Grade Requirements etc

1. Grade: Pass/Fail
2. What counts for the grade: attending the lectures, participating in the group discussion, submitting a short term paper at the end.
3. Term paper: Short (2 page document) comprising of a problem statement for a research question in your discipline that you think can benefit from using NLP methods, outlining a potential way to solve it based on what you learnt in this course.
4. Deadline: 27th October 2022.

Meetings Plan (Tentative)

Topic	Date	Time
Overview (course, NLP, econNLP)	5/10	3-6 pm CET
Python and Text	6/10	3-6 pm CET
NLP Methods: overview	7/10	3-6 pm CET
Zooming in: Text Classification	11/10	3-6 pm CET
NLP with little or no labeled data	12/10	3-6 pm CET
NLP and Economics - Group Discussions	13/10	3-6 pm CET
Revision	17 or 18 Oct	3-5 pm CET (Optional)
1-1 Sessions (15-30min each)	17-20th Oct	3-5 pm CET (Optional)

Textbooks

1. "Speech and Language Processing" by Jurafsky and Martin (2/3 editions)
2. "Practical Natural Language Processing" by Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta and Harshit Surana.
3. NLTK book
4. For Python: "Python for Everybody" Charles Severance

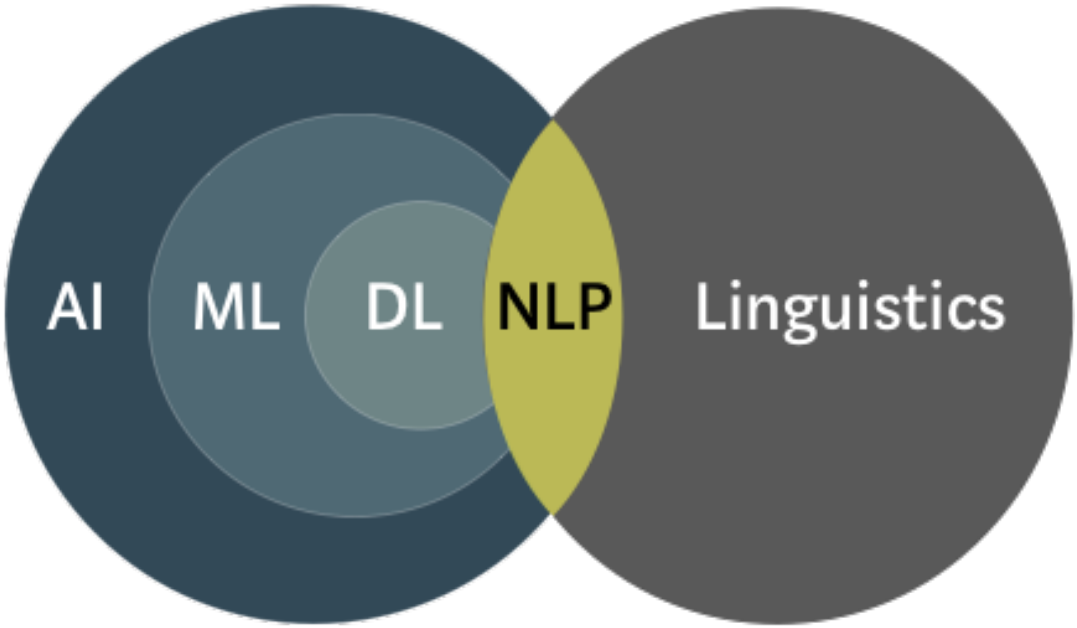
(Details on how to access these books are in the course website)

NLP Overview

1. **What is NLP?**
2. The many faces of NLP
3. What makes NLP Challenging?
4. Some common NLP Tasks: An overview
5. Approaches to NLP

note: images without source attribution are taken from our book: practicalnlp.ai

What is NLP?



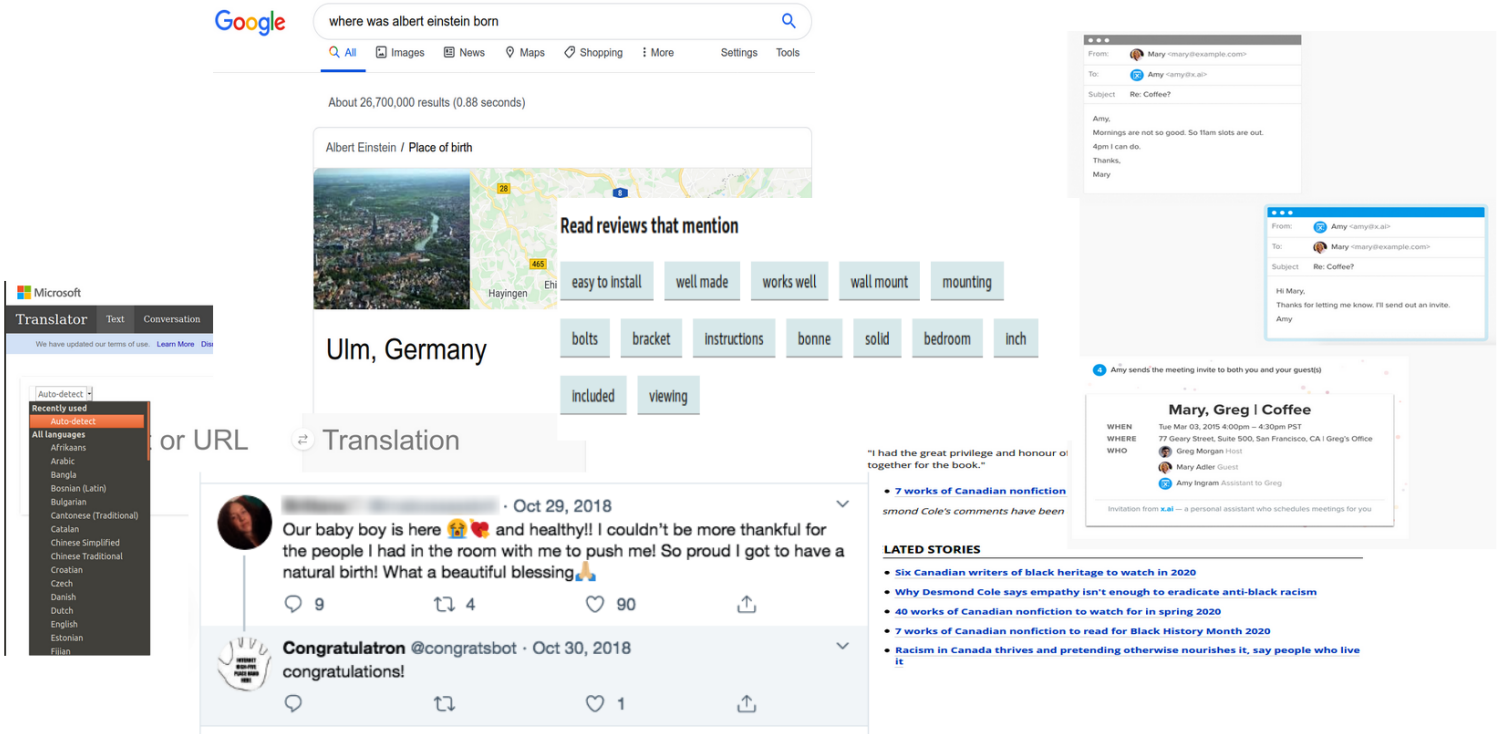
source

NLP is all about understanding and modeling human language using computational methods.

History of NLP

1. Foundational ideas: 40s and 50s. WWII and Beyond.
2. Main NLP problem of that time (and even now): Machine Translation
3. First few decades: Work focused on the development of speech recognition systems, logic based language understanding systems, creating elaborate grammars to teach human language to computers, rule based systems, and automatic language generation.
4. Late 90s on: Advent of statistical methods and machine learning
5. 2010s: Deep learning
6. Last few years: more into interpretable models, discussion about ethical issues, introspection about the field etc.

Everyday NLP Applications



But, we also know it is far from being perfect!

Where is NLP used elsewhere?

1. NLP is used as a method to answer research questions in many disciplines.
2. NLP sometimes plays a major role in discipline specific challenges, going beyond being just a research method.
3. In Google Scholar, I saw mentions of NLP methods in journals as diverse as Asian studies & History to Clinical Oncology.

The Many Faces of NLP

The many faces of NLP

Three broad groups:

1. NLPers: NLP researchers in academia and industry
2. Other researchers who use NLP methods in their research
3. Industry professionals developing NLP based applications

Data Scientist

Building real-world NLP systems

Product Manager

Applying NLP to their products & domains

Annotator

Or linguists creating NLP data

ML Engineer

Or data engineer applying MLOps to scale NLP systems

Business Leader

CXOs, VPs, founders incorporating NLP in their business

NLP Research: Summary

- generally heavy on algorithms/methods based on machine learning/deep learning
- relatively less focus on corpus creation, evaluation beyond standard tests, and rule engineering
- recent emphasis on understanding potential biases in models, ethics of NLP research, interpretability of models etc

What does NLP in Industry look like?

- large r&d teams building NLP focused products, for their own use as well as for third parties (Google Cloud NLP, Amazon Comprehend, IBM Watson, Microsoft NLP etc)
- software teams where NLP contributes to existing product functionalities
- speech to text/text to speech software, transcription tools etc.
- language learning/teaching/assessment software

and so on.

NLP in Other Disciplines: An Overview

Where is NLP used in other disciplines?

- NLP is used as a method to answer research questions in many disciplines.
- NLP sometimes plays a major role in discipline specific challenges, going beyond being **just** a research method.
- In Google Scholar, I saw mentions of NLP methods in journals as diverse as Asian studies & History to Clinical Oncology.
- Let us see a few examples (only paper titles, to give a big picture overview).

NLP in other disciplines - examples

Medical Informatics: Chen, L., Gu, Y., Ji, X., Sun, Z., Li, H., Gao, Y., & Huang, Y. (2020). [Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning](https://doi.org/10.1093/jamia/ocz141). Journal of the American Medical Informatics Association : JAMIA, 27(1), 56–64.
Plant Science: Braun, I. R., & Lawrence-Dill, C. J. (2019). [Automated methods enable direct computation on phenotypic descriptions for novel candidate gene prediction](https://www.frontiersin.org/articles/10.3389/fpls.2019.01629/full). Frontiers in Plant Science, 10, 1629.
Civil Engineering: Le, T., & David Jeong, H. (2017). [NLP-based approach to semantic classification of heterogeneous transportation asset data terminology](https://par.nsf.gov/servlets/purl/10069437). Journal of Computing in Civil Engineering, 31(6), 04017057.
Economics: Hansen, S., McMahon, M., & Prat, A. (2018). [Transparency and deliberation within the FOMC: a computational linguistics approach](https://academic.oup.com/qje/article/133/2/801/4582916). The Quarterly Journal of Economics, 133(2), 801-870.
Political Science: Benoit, K., Munger, K., & Spirling, A. (2019). [Measuring and explaining political sophistication through textual complexity](https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12423). American Journal of Political Science, 63(2), 491-508.
Urban planning: Plunz, R. A., Zhou, Y., Vintimilla, M. I. C., Mckeown, K., Yu, T., Uguccioni, L., & Sutto, M. P. (2019). [Twitter sentiment in New York City parks as measure of well-being. Landscape and urban planning](https://www.sciencedirect.com/science/article/pii/S0169204618305863), 189, 235-246.
Cultural Heritage: Machidon, O. M., Tavčar, A., Gams, M., & Duguleană, M. (2020). [CulturalERICA: A conversational agent improving the exploration of European cultural heritage](https://www.sciencedirect.com/science/article/pii/S1296207418308136). Journal of Cultural Heritage, 41, 152-165.

NLP in other disciplines: Summary

- Clearly, there are many more. I just sampled a few examples, from even fewer disciplines!
- Existing NLP tools + rules is a commonly used approach in some disciplines.
- Doing user studies and using a small set of manually annotated documents for validation of approaches is also a common method.
- In some fields (e.g., medical informatics), we also see state of the art deep learning and NLP.

How are the faces of NLP different from each other?

- NLP researchers focus on developing new methods, using standard corpora/evaluation procedures, and comparing against SOTA.
- Industry professionals focus on end users, end to end system development and maintenance.
  "If you think Machine Learning will give you a 100% boost, then a heuristic will get you a 50% of the way there" - Martin Zinkevich, Google
- Other discipline researchers are concerned how to use NLP methods to address their own research questions.

Questions?

My questions:

- What role does NLP play in your daily life?
- What do you want to do with NLP in your work?

What makes NLP challenging?
(what sort of issues pose problems for a computer?)

Language is ambiguous

Some ambiguous sentences

- Newspaper headlines
  - "Children make delicious snacks"
  - "Dead expected to rise"
  - "Republicans grill IRS chief over lost emails"
- Normal, grammatical sentences can be ambiguous too:
  - "I saw a man on a hill with a telescope."
  - "Look at the man with one eye"

We are not even talking about ambiguities involving speech or alternative interpretations due to stress/emphasis on some word.

Some types of ambiguity

1. Lexical ambiguity: due to multiple meanings or senses of word usage
   e.g., He stood near the **bank**
2. Structural ambiguity: due to syntactic structure
   e.g., I saw the man on the hill with telescope.
3. Semantic ambiguity: more interpretations possible
   e.g., John and Mary are married (to each other? or to different people?)
4. Referential ambiguity
   e.g., She dropped the *plate* on the *table* and broke **it**
5. Ambiguity due to the use of non-literal language
   e.g., Time flies like an arrow

Good source to read more: http://cs.nyu.edu/faculty/davise/ai/ambiguity.html

"common" knowledge for humans

Look at these two sentences:
Dog bit man.
Man bit dog.
- For a computer, both of them are linguistically the same.
We know only the first one is "normal" English sentence because we have "world knowledge".

Few more challenges

- Language is diverse: many different forms of documents such as news, tweets, legal texts etc.
- Language is creative: its use changes over time, and vocabulary gets richer.
- There are many different languages in the world
- Many spelling variations, slangs, sarcasm etc.

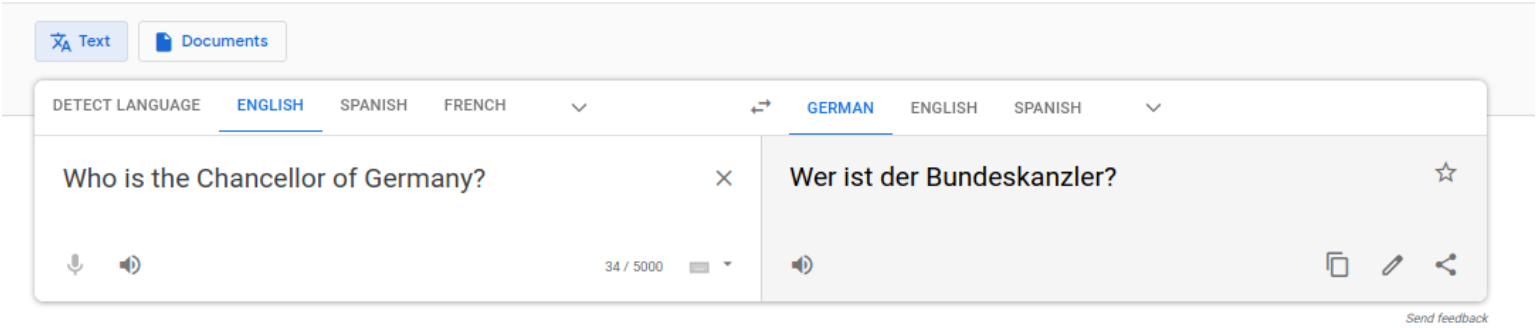
- NLP solutions should account for all these things!

So, the summary is:

perfect NLP is hard to achieve because of all these issues that come up when we start using computers to analyze language!

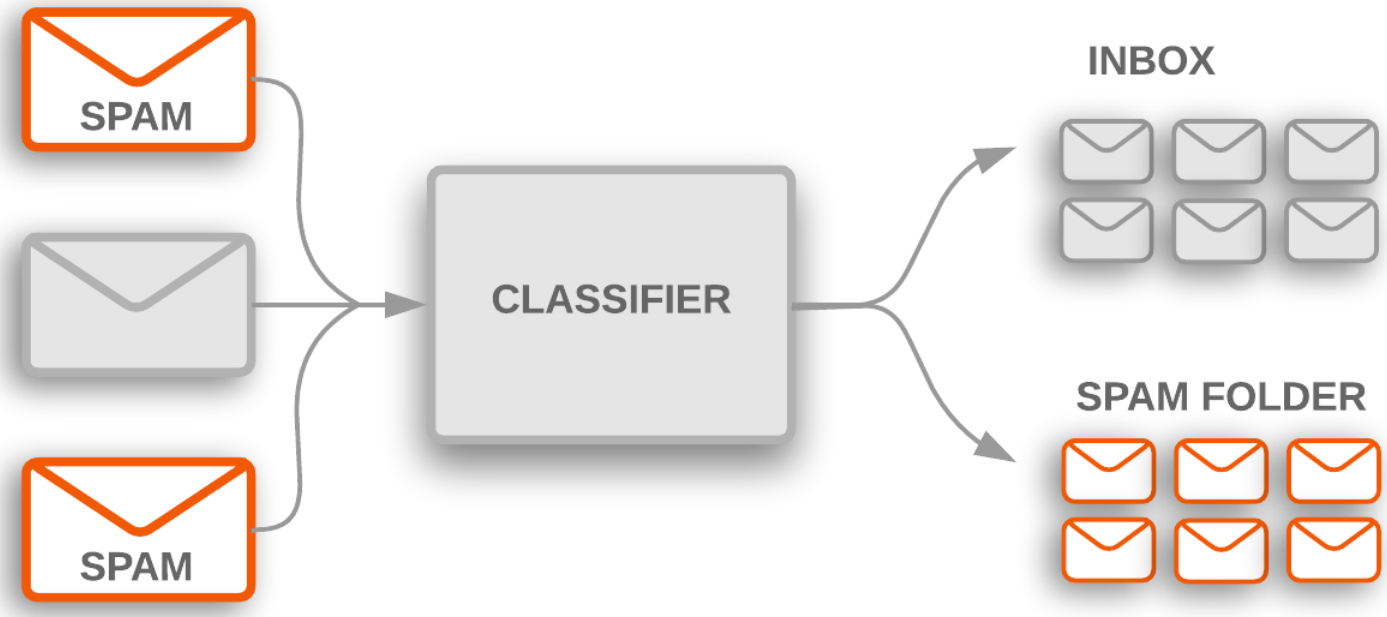
Some NLP Tasks

Machine Translation



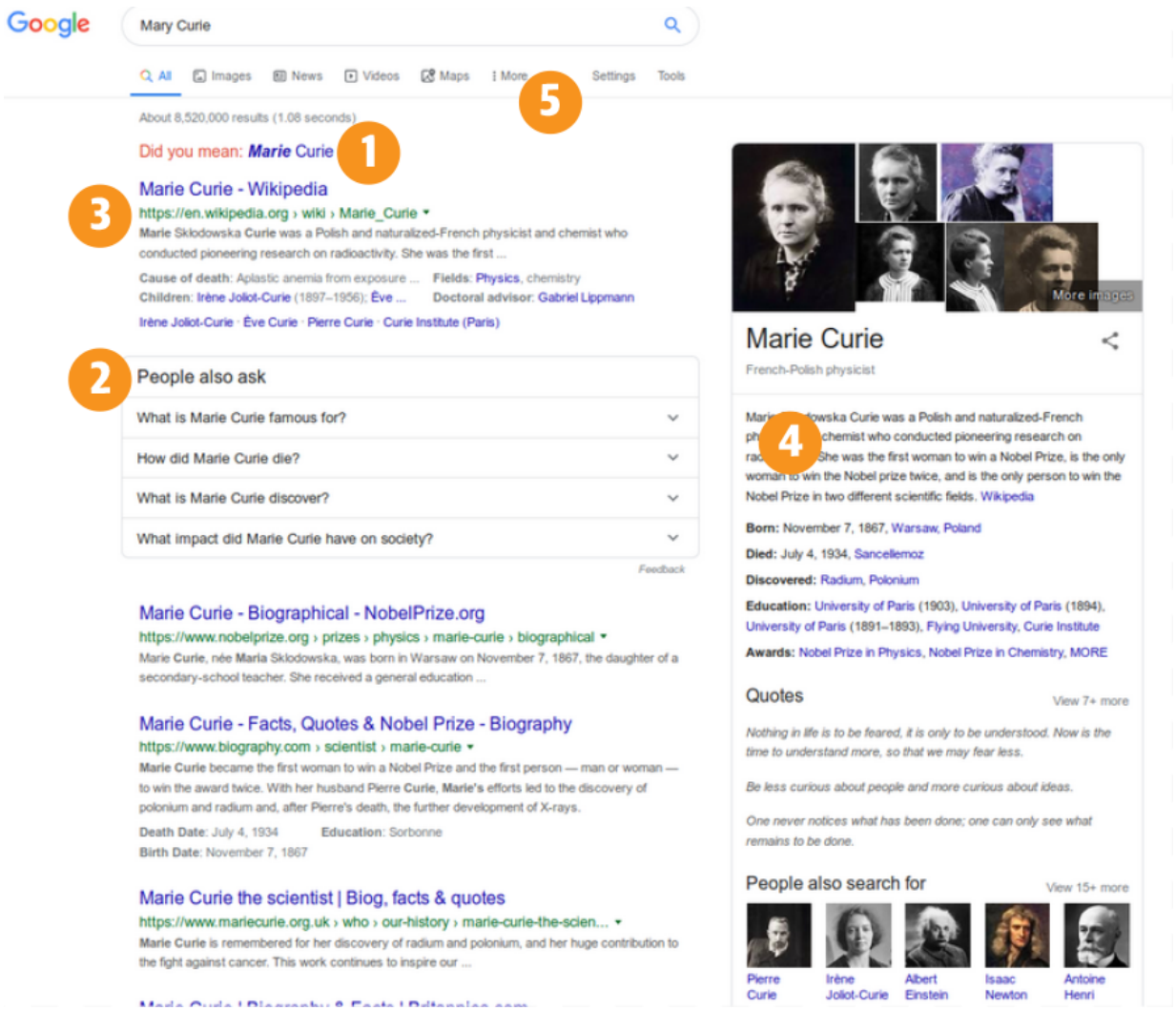
source: https://translate.google.com

Text Classification

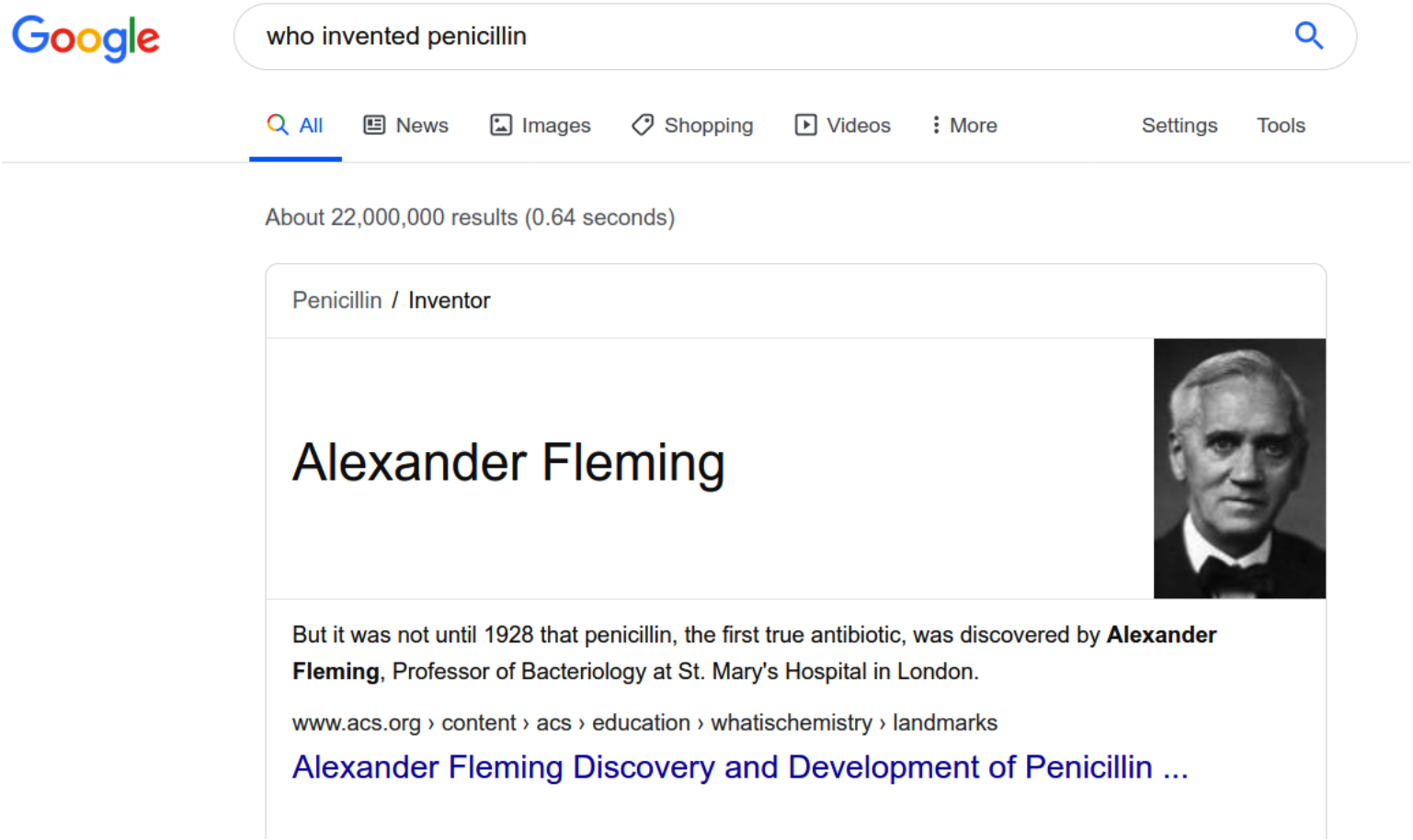


source: https://developers.google.com/machine-learning/guides/text-classification

Search


Google
Mary Curie
All
Images
News
Videos
Maps
More
Settings
Tools
About 8,520,000 results (1.08 seconds)
Did you mean: Marie Curie
1
5
3
Marie Curie - Wikipedia
https://en.wikipedia.org › wiki › Marie_Curie
Marie Sklodowska Curie was a Polish and naturalized-French physicist and chemist who conducted pioneering research on radioactivity. She was the first ...
Cause of death: Aplastic anemia from exposure ...
Fields: Physics, chemistry
Children: Irène Joliot-Curie (1897–1956); Ève ...
Doctoral advisor: Gabriel Lippmann
Irène Joliot-Curie · Ève Curie · Pierre Curie · Curie Institute (Paris)
2
People also ask
What is Marie Curie famous for?
How did Marie Curie die?
What is Marie Curie discover?
What impact did Marie Curie have on society?
Feedback
Marie Curie - Biographical - NobelPrize.org
https://www.nobelprize.org › prizes › physics › marie-curie › biographical
Marie Curie, née Maria Sklodowska, was born in Warsaw on November 7, 1867, the daughter of a secondary-school teacher. She received a general education ...
Marie Curie - Facts, Quotes & Nobel Prize - Biography
https://www.biography.com › scientist › marie-curie
Marie Curie became the first woman to win a Nobel Prize and the first person — man or woman — to win the award twice. With her husband Pierre Curie, Marie's efforts led to the discovery of polonium and radium and, after Pierre's death, the further development of X-rays.
Death Date: July 4, 1934
Education: Sorbonne
Birth Date: November 7, 1867
Marie Curie the scientist | Biog, facts & quotes
https://www.mariecurie.org.uk › who › our-history › marie-curie-the-scien...
Marie Curie is remembered for her discovery of radium and polonium, and her huge contribution to the fight against cancer. This work continues to inspire our ...
More images
Marie Curie
French-Polish physicist
4
owska Curie was a Polish and naturalized-French
chemist who conducted pioneering research on
She was the first woman to win a Nobel Prize, is the only
woman to win the Nobel prize twice, and is the only person to win the
Nobel Prize in two different scientific fields. Wikipedia
Born: November 7, 1867, Warsaw, Poland
Died: July 4, 1934, Sancellemoz
Discovered: Radium, Polonium
Education: University of Paris (1903), University of Paris (1894), University of Paris (1891–1893), Flying University, Curie Institute
Awards: Nobel Prize in Physics, Nobel Prize in Chemistry, MORE
Quotes
View 7+ more
Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.
Be less curious about people and more curious about ideas.
One never notices what has been done; one can only see what remains to be done.
People also search for
View 15+ more
Pierre Curie
Irène Joliot-Curie
Albert Einstein
Isaac Newton
Antoine Henri

Question Answering



Information Extraction

SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

¿¿

- Who is Luca Mestri?

Information Extraction

SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

¿¿

- Who is Luca Mestri? needs: Named Entity Recognition and Linking, Relation extraction
- What is the article about?

Information Extraction

SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

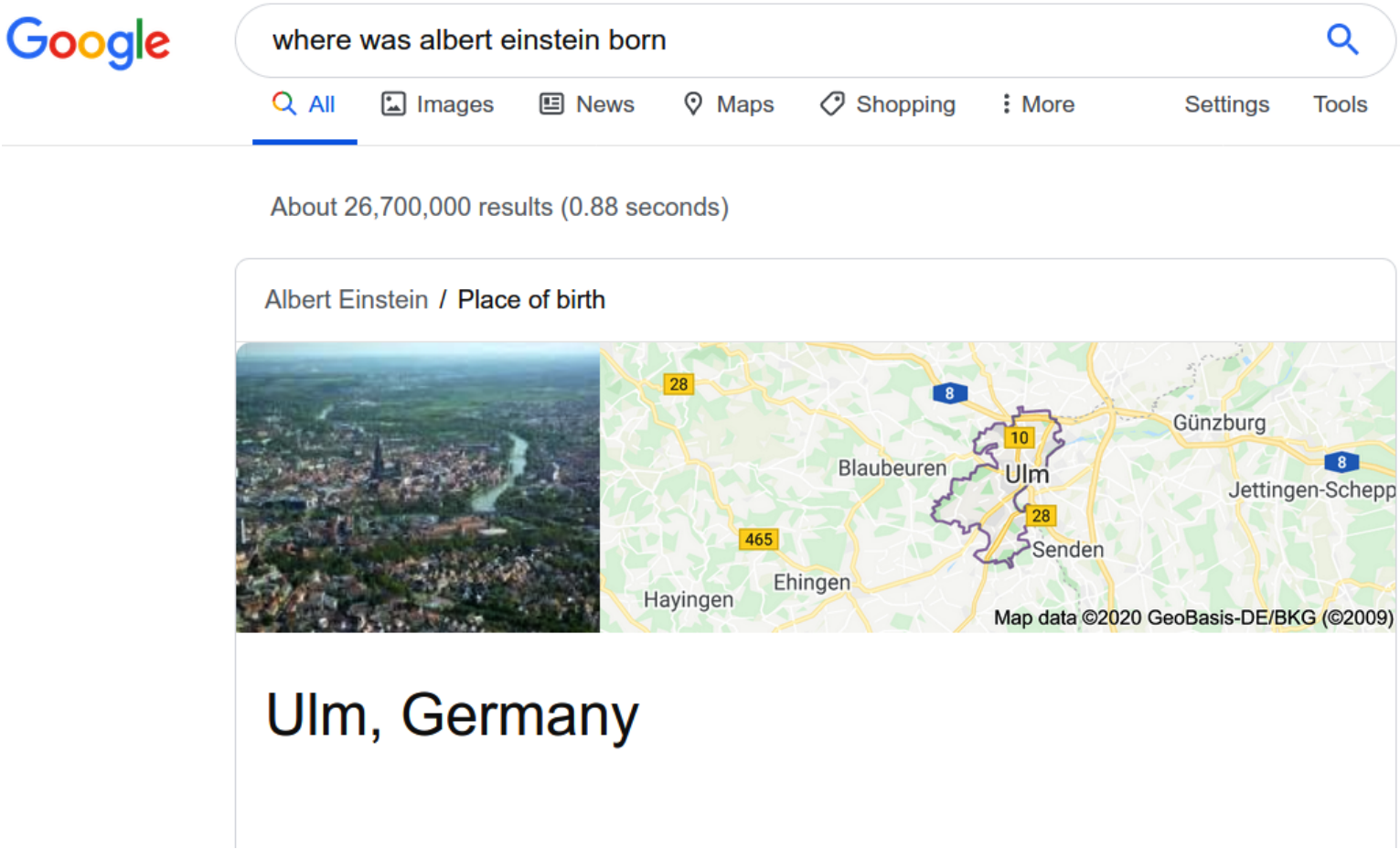
"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

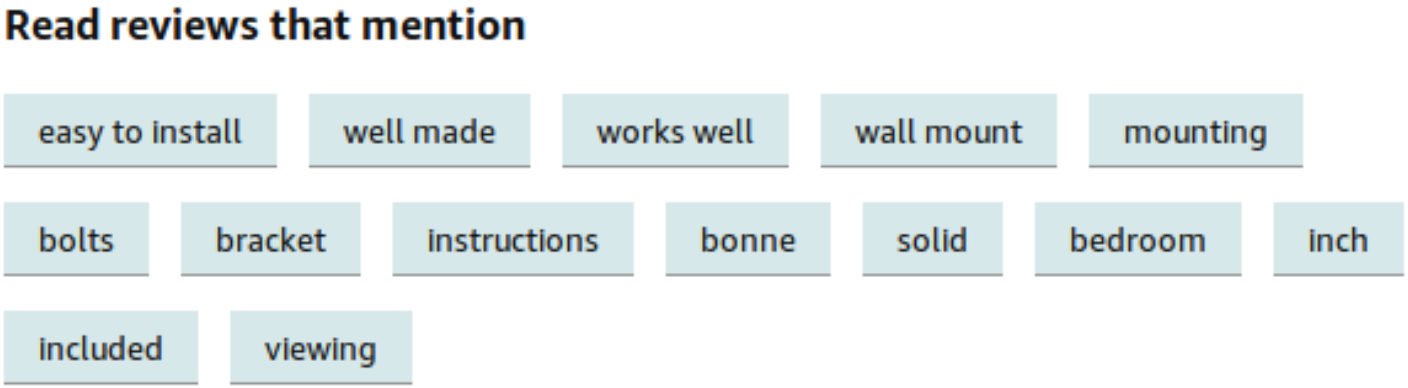
¿¿

- Who is Luca Mestri? needs: Named Entity Recognition and Linking, Relation extraction
- What is the article about? needs: Key phrase extraction, event extraction

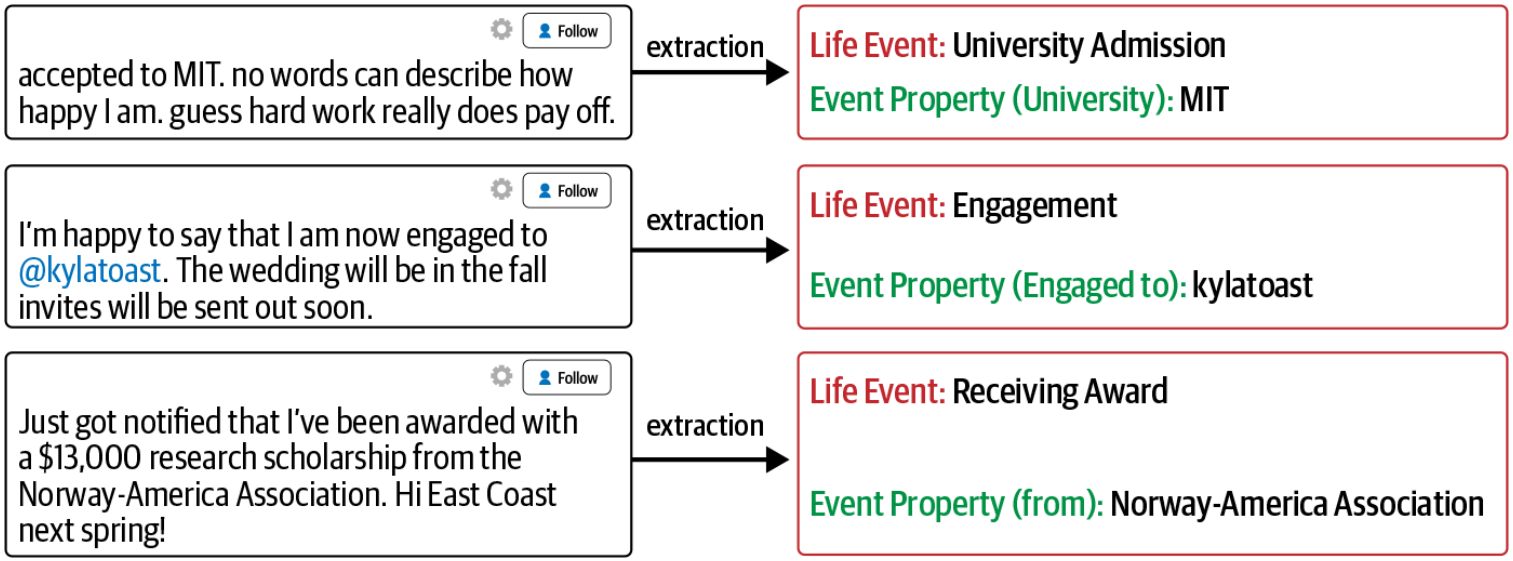
Named Entity Extraction/Linking



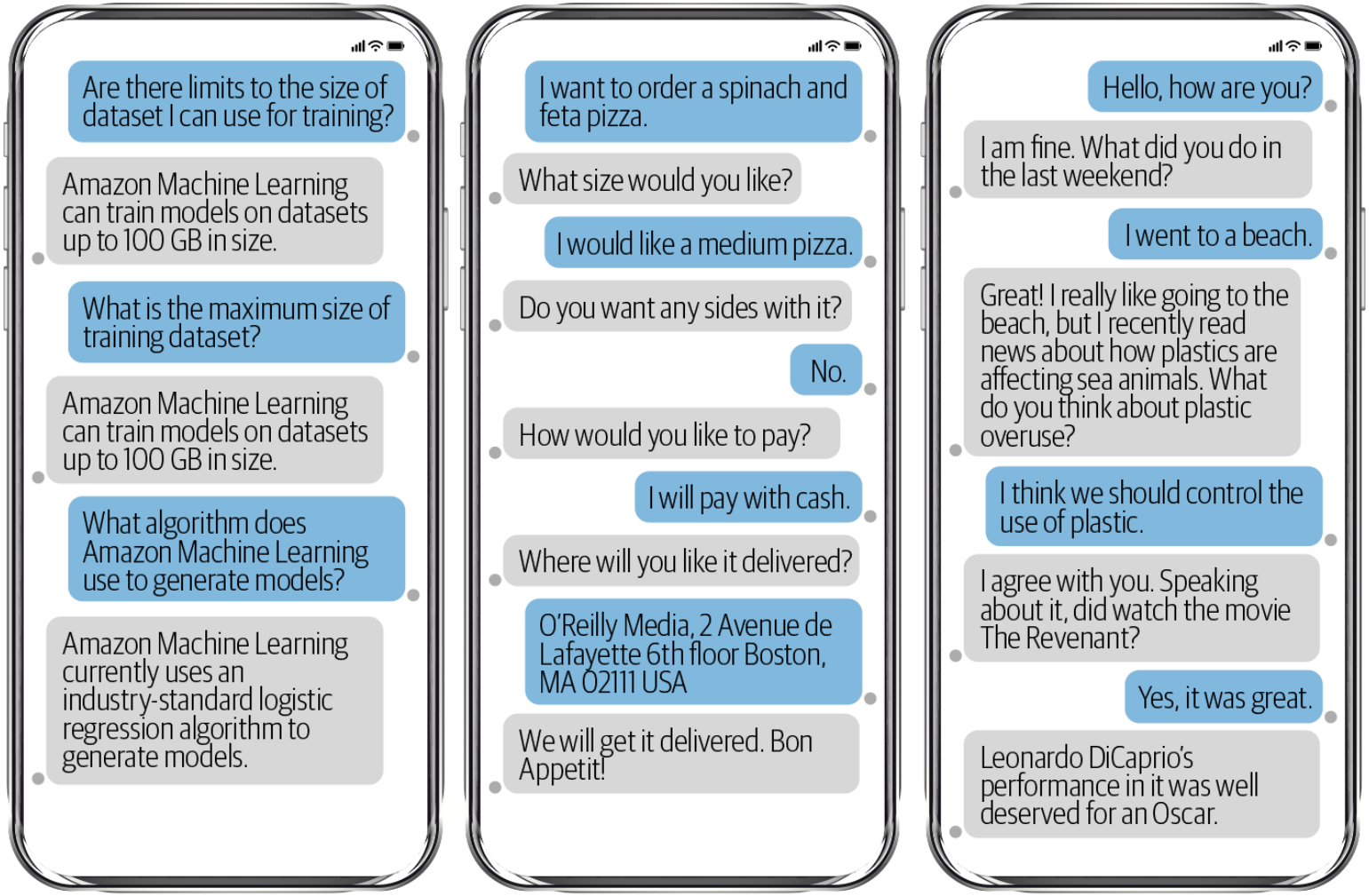
Key Phrase Extraction



Event/Relation Extraction



Chatbots



FAQ Bot **Flow-Based Bot** **Open-Ended Bot**

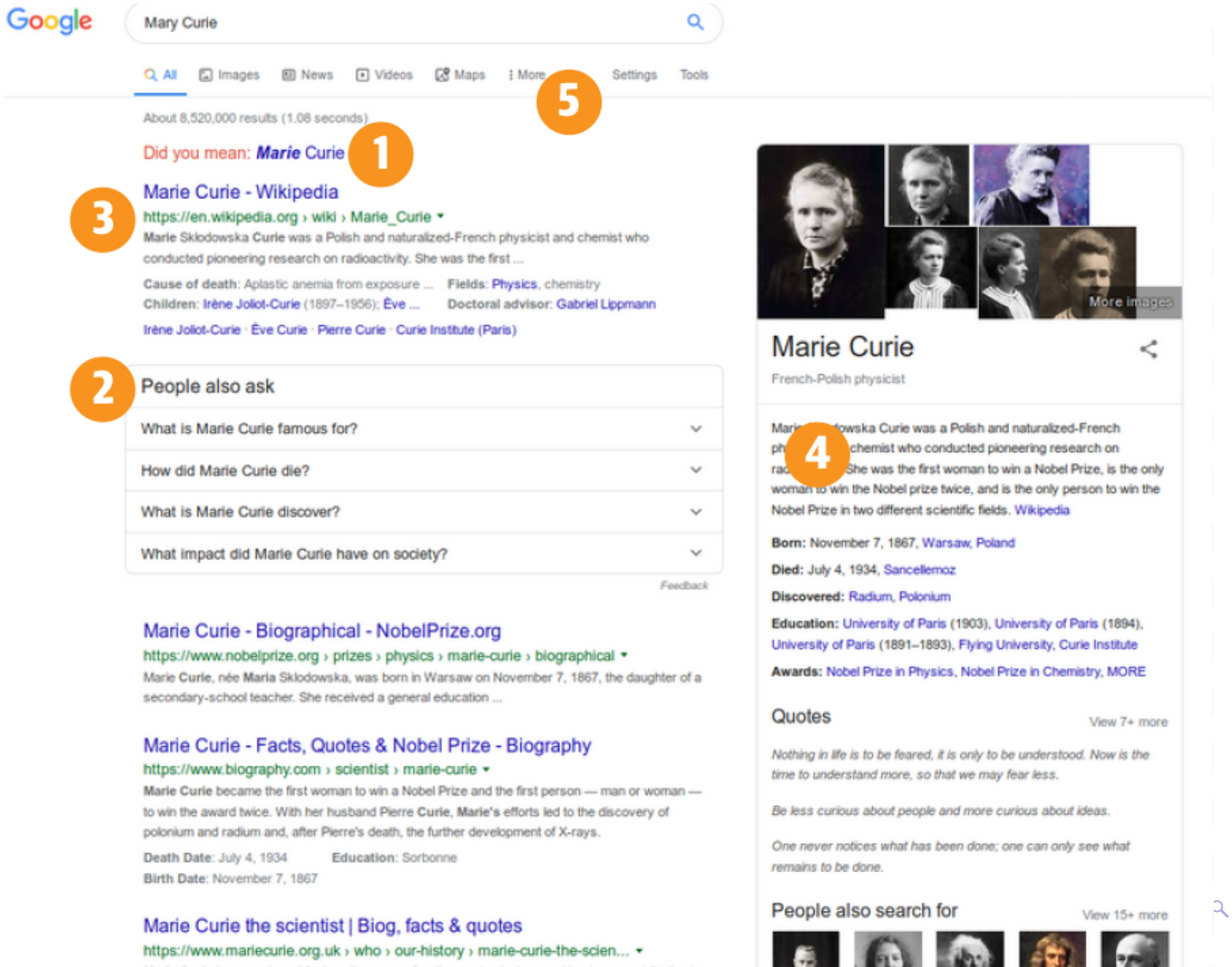
Some other tasks

- text summarization
- text recommendation
- topic modeling
- text to speech/speech to text conversion
- language generation (e.g., automatically generating weather reports, chatbots, etc.)

.... and so on.

A real world example: revisiting search

What levels of language processing are involved? What NLP tasks can be seen here?



Approaches to NLP

1. heuristics based NLP
2. machine learning
3. deep learning
4. combination of all or some of the above three

(More on these later)

Heuristics

1. Word lists, lexicons etc.
2. Rule engineering based on observed patterns of language use.
3. Regular expressions

...

Machine Learning

- Machine learning is a collection of methods that learn a task automatically by looking at lots (and lots) of examples
- Usually involves a feature extraction step, where we need to specify what kind of patterns (e.g., words? heuristics? pos tags?) should a machine learn.
- Different kinds of machine learning algorithms are used in NLP - naive bayes, logistic regression (text classification), conditional random fields (sequence tagging), clustering etc.
- We will briefly see how some of them are useful for NLP as we progress with the course.

Deep Learning

...a sub-field of machine learning

- Deep learning is inspired by learning in humans and other biological forms, but "deep" refers to the number of layers in the artificial neural network of the deep learning method.

Deep Learning

...a sub-field of machine learning

- Deep learning is inspired by learning in humans and other biological forms, but "deep" refers to the number of layers in the artificial neural network of the deep learning method.
- Some popular deep learning architectures used in NLP in the past 5 years:
  - Convolutional Neural Networks are typically used in image processing related problems, and are useful to automatically extract features (words/word sequences) in NLP tasks.
  - Recurrent Neural Networks are more popular than CNNs for NLP, as they are capable of modeling sequential data (like language).
  - Transformers (e.g., BERT) model the textual context, but not sequentially. It uses a phenomenon called **attention** to model the context.

Deep Learning

...a sub-field of machine learning

- Deep learning is inspired by learning in humans and other biological forms, but "deep" refers to the number of layers in the artificial neural network of the deep learning method.
- Some popular deep learning architectures used in NLP in the past 5 years:
  - Convolutional Neural Networks are typically used in image processing related problems, and are useful to automatically extract features (words/word sequences) in NLP tasks.
  - Recurrent Neural Networks are more popular than CNNs for NLP, as they are capable of modeling sequential data (like language).
  - Transformers (e.g., BERT) model the textual context, but not sequentially. It uses a phenomenon called **attention** to model the context.

Note: Transformers have become very popular in NLP in recent 2-3 years across tasks and languages.

Deep Learning and Transfer Learning

- Deep learning and Machine learning methods in general expect you have to have a lot of data for the problem you want to solve.
- Sometimes, we may have a lot of data on some related problem, but very small amount of data on a particular problem.
- Transfer learning is all about how to use existing data/knowledge for a new problem.

Transfer learning has also become a popular method in NLP in the recent 2-3 years.

How does one build NLP solutions?

Option 1: Utilize Existing Services

We don't always have to build everything ourselves. We can use a pay as you go service from a third party provider. An example: Microsoft's machine translation API

```
import os, requests, uuid, json
subscription_key = "XXXX"
endpoint = "https://api-nam.cognitive.microsofttranslator.com"
path = '/translate?api-version=3.0'
params = '&to=de' #From English to German (de)
constructed_url = endpoint + path + params
headers = {
    'Ocp-Apim-Subscription-Key': subscription_key,
    'Content-type': 'application/json',
    'X-ClientTraceId': str(uuid.uuid4())
}

body = [{'text' : 'How good is Machine Translation?'}]
request = requests.post(constructed_url, headers=headers, json=body)
response = request.json()

print(json.dumps(response, sort_keys=True, indent=4, separators=(',', ': ')))
```

source: Practical NLP, Ch 7

Output


```
[
    {
    "detectedLanguage": {
          "language": "en",
          "score": 1.0
    },
    "translations": [
          {
               "text": "Wie gut ist maschinelle Ãbersetzung?",
               "to": "de"
          }
    ]
    }
]
```

Advantages and Disadvantages

- Advantage: You don't have to worry about setting stuff up, hiring a large NLP team, maintaining the NLP system etc.
- Disadvantages:
  1. This only works if you have problem that exactly meets the specifications of such an available API
  2. No possibility of customization/modification
  3. Depending on how much you use, costs may escalate

Note: You still have to think whether this approach is a long term solution for your problem.

Short Exercise

- Spend 5 minutes on the internet and look for NLP services from Cloud providers such as Google, Microsoft, Amazon, Watson etc.
- Make a note of what sort of services do they provide for NLP applications.

What is more likely to happen, though?

- Say you have custom problem (e.g., classify customer tweets about your product into "actionable" and "non-actionable" ones.)
- Will any of these off the shelf solutions work?

What is more likely to happen, though?

- Say you have custom problem (e.g., classify customer tweets about your product into "actionable" and "non-actionable" ones.)
- Will any of these off the shelf solutions work?
- What do we need to design a solution for this problem??

What is more likely to happen, though?

- Say you have custom problem (e.g., classify customer tweets about your product into "actionable" and "non-actionable" ones.)
- Will any of these off the shelf solutions work?
- What do we need to design a solution for this problem??
- First of all, we need "data". What Data? for what?

What is more likely to happen, though?

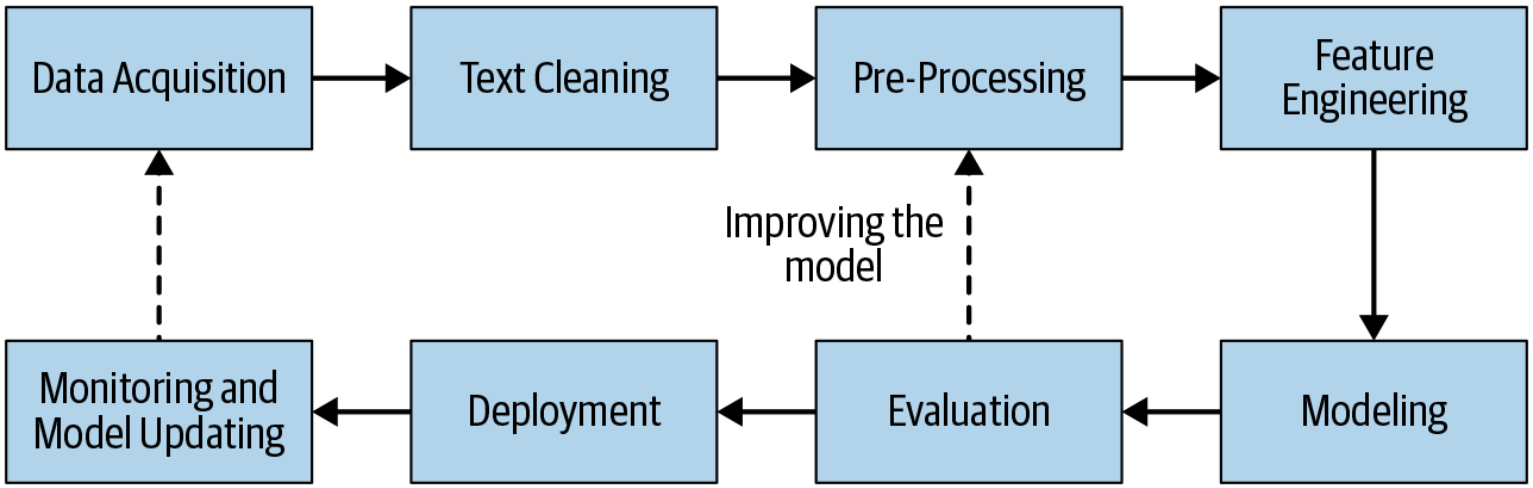
- Say you have custom problem (e.g., classify customer tweets about your product into "actionable" and "non-actionable" ones.)
- Will any of these off the shelf solutions work?
- What do we need to design a solution for this problem??
- First of all, we need "data". What Data? for what?
  1. To "train" NLP models
  2. To "evaluate" and validate NLP models

What is more likely to happen, though?

- Say you have custom problem (e.g., classify customer tweets about your product into "actionable" and "non-actionable" ones.)
- Will any of these off the shelf solutions work?
- What do we need to design a solution for this problem??
- First of all, we need "data". What Data? for what?
  1. To "train" NLP models
  2. To "evaluate" and validate NLP models
- ... and then?

NLP System Development Pipeline

A Common Case



Data Acquisition

- Where do we get our data from?

Data Acquisition

- Where do we get our data from?
- Let us say you are working in a software company, and they asked you to develop a customer ticket routing system that looks at a ticket text, and classifies it into three categories: technical, sales, other. Where will your data come from?

Data Acquisition

- Where do we get our data from?
- Let us say you are working in a software company, and they asked you to develop a customer ticket routing system that looks at a ticket text, and classifies it into three categories: technical, sales, other. Where will your data come from?
- In an ideal scenario, we have some historical data of customer tickets along with this routing information.
- But what if they were just routing to the right team, but not storing that information anywhere in the past? We don't have the training data we want!

How do we get our Data?

- use a public NLP dataset (e.g., https://huggingface.co/datasets/)

How do we get our Data?

- use a public NLP dataset (e.g., https://huggingface.co/datasets/)
- scrape the data from the web (e.g., customer support forums of other products, if they are available and are tagged with categories)
- work together with your customer support team and gradually collect the data you want.
- using pattern matching and other such methods to create some data to train your own models and gradually improve this baseline.
- setting up data annotation experiments

.... and so on.

Text Extraction and Cleaning

- What, according to you, is the format of data you see in NLP?

Text Extraction and Cleaning

- What, according to you, is the format of data you see in NLP?
- Data can come in all forms: PDF, Docs, HTML files, scanned png files, tables etc.
- Text extraction and cleanup refers to the process of extracting raw text from the input data by removing all the other non-textual information.
- Text extraction may not involve NLP per se, but it defines the rest of your NLP pipeline. Bad text extraction = Bad NLP system.

Text Extraction: Some facts

- PDF to text conversion is hard and imperfect. Not all pdfs can be efficiently parsed.

Text Extraction: Some facts

- PDF to text conversion is hard and imperfect. Not all pdfs can be efficiently parsed.
- When we are extracting text from images, We may see some characters not rendered properly, or some words extracted with spelling mistakes etc
- If all of the "dataset" is a large collection of pdf documents with scanned text along with some tables: it is like the worst of all NLP worlds :-)
- We don't have a single solution that works for all cases. There are tools like Amazon Textract, but they are not perfect.

Text Pre-processing

- Sentence segmentation and word tokenization

Text Pre-processing

- Sentence segmentation and word tokenization
- Stop word removal, stemming and lemmatization, removing digits/punctuation, lowercasing, etc.

Text Pre-processing

- Sentence segmentation and word tokenization
- Stop word removal, stemming and lemmatization, removing digits/punctuation, lowercasing, etc.
- Normalization, language detection, code mixing, transliteration, etc.

Text Pre-processing

- Sentence segmentation and word tokenization
- Stop word removal, stemming and lemmatization, removing digits/punctuation, lowercasing, etc.
- Normalization, language detection, code mixing, transliteration, etc.
- POS tagging, parsing, coreference resolution, etc.

Questions so far?

Feature Engineering

- The goal of feature engineering is to capture the characteristics of the text into a numeric vector that can be understood by the ML algorithms.
- There are primarily two ways of feature extraction in NLP:
  1. hand crafted features (e.g., number of words/sentences in a document, number of spelling errors/sentence etc.)
  2. automatically extracted features (e.g., Bag of words, Bag of N-grams etc.)

Modeling

- Heuristics based systems (e.g., regular expressions, rule based matching etc)
- Learning from data: Machine learning (e.g., logistic regression, support vector machines etc.), deep learning
- Model ensemble (combining predictions from multiple models)
- Model cascade (using one model's prediction as input to another model)

.....

Evaluation

How well is my NLP approach doing?

- Intrinsic evaluation focuses on intermediary objectives, while extrinsic focuses on evaluating performance on the final objective.

Evaluation

How well is my NLP approach doing?

- Intrinsic evaluation focuses on intermediary objectives, while extrinsic focuses on evaluating performance on the final objective.
- Consider a email spam classification system:
  1. Intrinsic evaluation will focus on measuring the system performance using precision and recall.

Evaluation

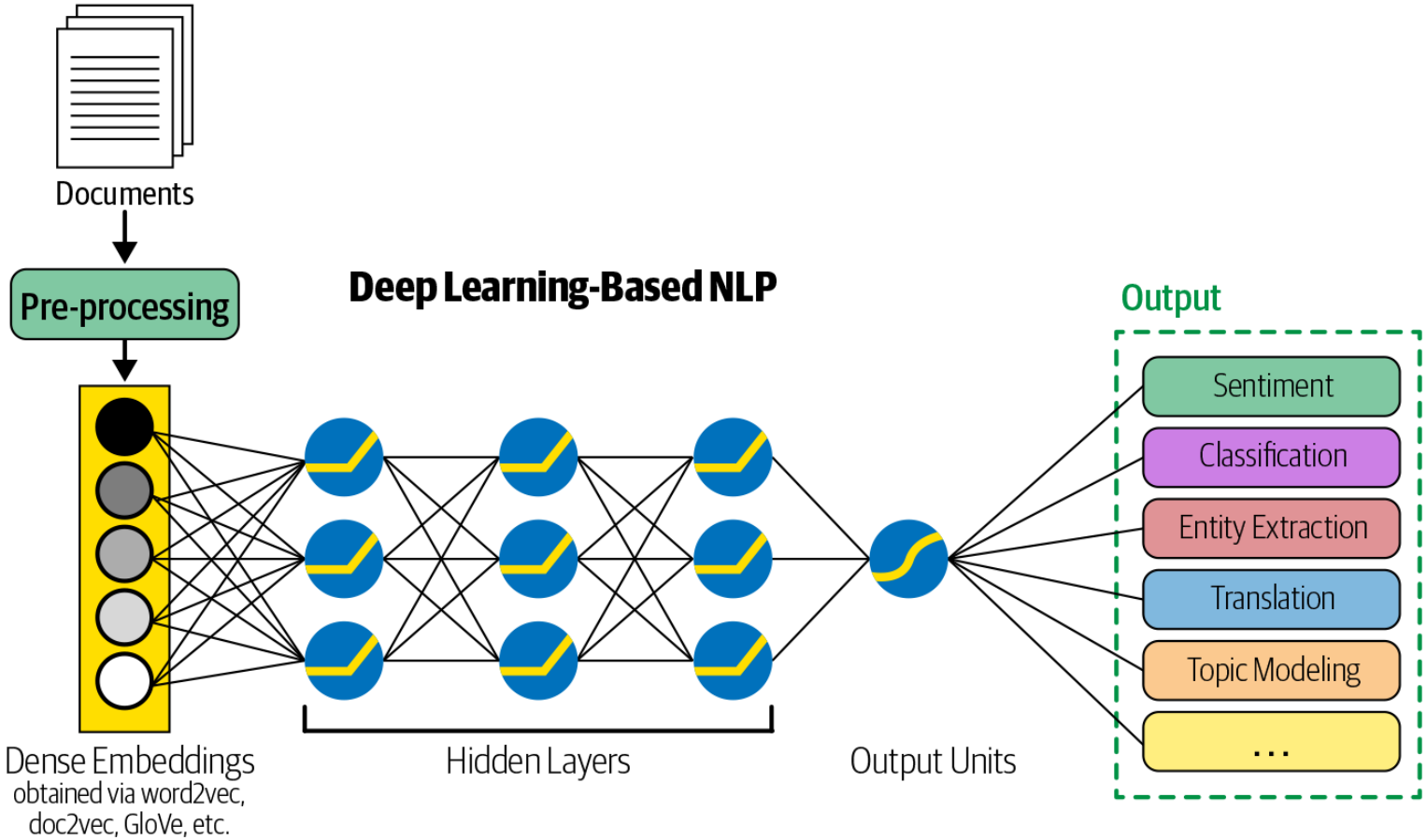
How well is my NLP approach doing?

- Intrinsic evaluation focuses on intermediary objectives, while extrinsic focuses on evaluating performance on the final objective.
- Consider a email spam classification system:
  1. Intrinsic evaluation will focus on measuring the system performance using precision and recall.
  2. Extrinsic evaluation will focus on measuring the time a user wasted because a spam email went to their inbox or a genuine email went to their spam folder.

Deployment, Monitoring and Model Updating

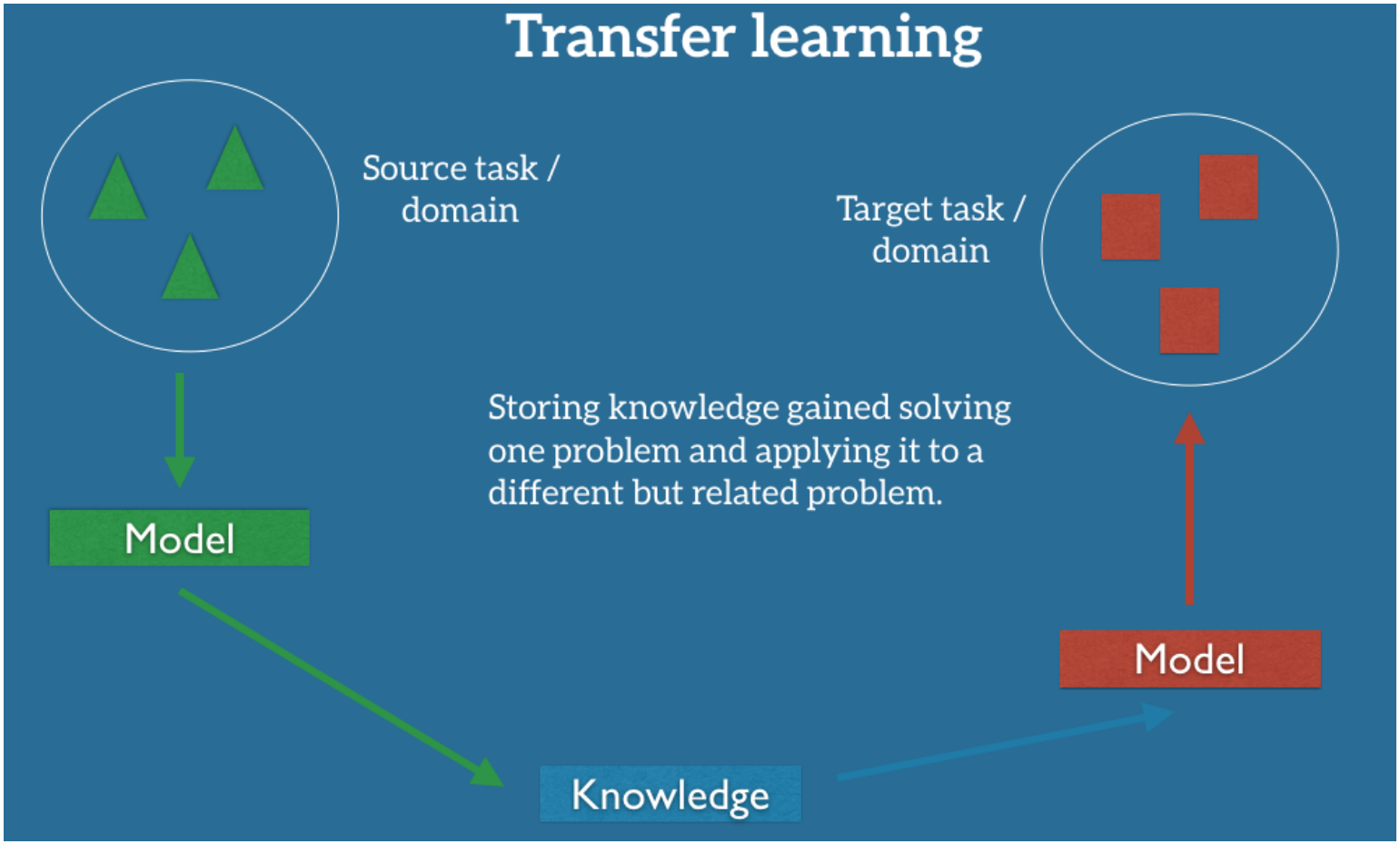
- Once we have a good model, we have to deploy it in the context of a larger system (e.g., spam classification is a part of an email software)
- A common approach: deploy the NLP system as a micro service/web service
- Model monitoring: e.g., using a performance dashboard showing the model parameters and key performance indicators
- Model updating: Model has to stay current, with changing data. So, we should have some way of regularly updating, evaluating and deploying a model.

Pipeline Variation: Deep Learning Pipeline



source: https://blog.aylien.com/leveraging-deep-learning-for-multilingual/

Pipeline Variation: Transfer Learning Pipeline



source: https://ruder.io/transfer-learning/

NLP Pipeline: Some Questions

- What is the difference between traditional pipeline and deep learning pipeline?
- What are some advantages and disadvantages of deep learning?
- What are some advantages and disadvantages of transfer learning?
- If you already know some ML/deep learning usage, how many steps of this pipeline did you learn/think about so far?

Few More Questions

- What according to you is the most important part of the NLP Pipeline?

Few More Questions

- What according to you is the most important part of the NLP Pipeline?
- Should our data to train an NLP system come from only a single source?

Few More Questions

- What according to you is the most important part of the NLP Pipeline?
- Should our data to train an NLP system come from only a single source?
- When are rule based "models" relevant?
- What can be a rule based approach to developing a sentiment analyzer?

More observations

- There is no single model. We have to set up a model monitoring/updating pipeline using some evaluation criteria.
- There is no single training set or test set. They will also evolve with time.

More observations

- There is no single model. We have to set up a model monitoring/updating pipeline using some evaluation criteria.
- There is no single training set or test set. They will also evolve with time.
- NLP tools we use in our pipeline are not perfect. Even a simple thing as text extraction or tokenization can have many unresolved issues. While our models are all very valuable effort, these steps are, too.

More observations

- There is no single model. We have to set up a model monitoring/updating pipeline using some evaluation criteria.
- There is no single training set or test set. They will also evolve with time.
- NLP tools we use in our pipeline are not perfect. Even a simple thing as text extraction or tokenization can have many unresolved issues. While our models are all very valuable effort, these steps are, too.
- No model can solve the problem of data quality. So, focus on getting good quality data to solve your problem first.

More observations

- There is no single model. We have to set up a model monitoring/updating pipeline using some evaluation criteria.
- There is no single training set or test set. They will also evolve with time.
- NLP tools we use in our pipeline are not perfect. Even a simple thing as text extraction or tokenization can have many unresolved issues. While our models are all very valuable effort, these steps are, too.
- No model can solve the problem of data quality. So, focus on getting good quality data to solve your problem first.
- Build a solution incrementally. Don't jump into the most complex solution first. Eventually, you want your stuff to be reliable, and not too expensive to maintain in short/long term.

... ...

Summary

- When we read about NLP in the news/research updates, we usually see models everywhere.
- However, there is much more behind an NLP system
- I hope this session introduced some of the issues behind the scenes.

A (not so quick) detour: Text Representation

In "Feature Engineering", I said:

- There are primarily two ways of feature extraction in NLP:
  1. hand crafted features (e.g., number of words/sentences in a document, number of spelling errors/sentence etc.)
  2. automatically extracted features (e.g., Bag of words, Bag of N-grams etc.)

  - What exactly do these mean??

What is text representation?

- Any form of data text/image/video/audio should be converted into some form of numeric representation so that it can be processed by NLP/Machine Learning algorithms later
- In NLP, this conversion of raw text to a suitable numerical form is called text representation
- We will use text representation and feature extraction/representation as synonyms in this course.

How is text different from any other form of data?

- How do we do feature extraction with images? An image is stored in a computer as a matrix of pixels where each cell[i,j] in the matrix represents pixel i,j of the image. This matrix representation accurately represents the complete image.

How is text different from any other form of data?

- How do we do feature extraction with images? An image is stored in a computer as a matrix of pixels where each cell[i,j] in the matrix represents pixel i,j of the image. This matrix representation accurately represents the complete image.
- A video can be seen as a sequence of frames/images. So, it can be represented as a sequential collection of matrices.

How is text different from any other form of data?

- How do we do feature extraction with images? An image is stored in a computer as a matrix of pixels where each cell[i,j] in the matrix represents pixel i,j of the image. This matrix representation accurately represents the complete image.
- A video can be seen as a sequence of frames/images. So, it can be represented as a sequential collection of matrices.
- Consider speech—it is transmitted as a wave. To represent it mathematically, we sample the wave and record its amplitude (height).

What about text?

Hand crafted features to represent text

- When we know the domain well, it is possible to represent text as a collection of hand crafted features.
- e.g., when I am categorizing sentiment, I can create a list of positive words, and a list of negative words, and use the % of positive and negative words in a document as its two dimensional text representation.
- Assuming we come across a new problem/dataset, and we don't know where to begin, what can we do?

Hand crafted features to represent text

- When we know the domain well, it is possible to represent text as a collection of hand crafted features.
- e.g., when I am categorizing sentiment, I can create a list of positive words, and a list of negative words, and use the % of positive and negative words in a document as its two dimensional text representation.
- Assuming we come across a new problem/dataset, and we don't know where to begin, what can we do?
- in NLP, it is common to use what I will call today as **automatic text representations**, where we don't have to explicitly encode any information.

Consider a Toy corpus

Table 3-1. Our toy corpus

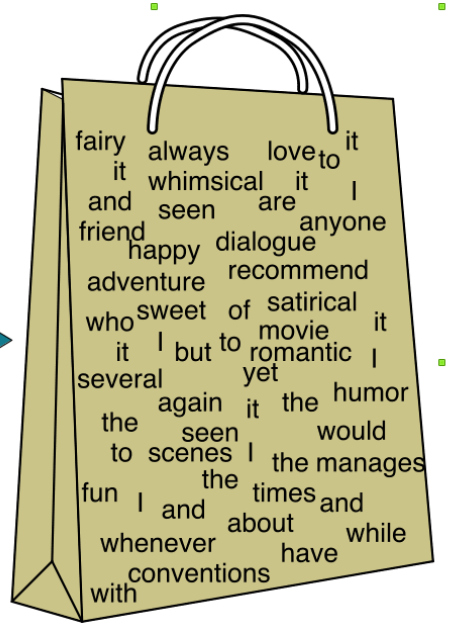
D1	Dog bites man.
D2	Man bites dog.
D3	Dog eats meat.
D4	Man eats food.

Let us start with some simple text representations

- The vocabulary in this corpus, ignoring case-differences and punctuation consists of 6 words: [dog, bites, man, eats, meat, food]
- We can organize the vocabulary in any order. In this example, we simply take the order in which the words appear in the corpus.
- Every document in this corpus can now be represented with a vector of size six.

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

source: https://web.stanford.edu/~jurafsky/slp3/4.pdf

Bag of Words

- for our toy corpus, where the word IDs are dog = 1, bites = 2, man = 3, meat = 4 , food = 5, eats = 6
- D1 ("Dog bites man") becomes [1 1 1 0 0 0]
- D4 ("Man eats food") becomes [0 0 1 0 1 1].

Advantages of BoW

- BoW is fairly simple to understand and implement.
- Documents sharing vocabulary have their vector representations closer to each other in Euclidean space. In our toy corpus, distance between D1 and D2 is 0 as compared to the distance between D1 and D4, which is 2.
- We have a fixed-length representation for any sentence of arbitrary length.

Disadvantages of BoW

- The size of the vector increases with the size of the vocabulary. One way to control this is to take the first N most frequent words in the entire corpus.
- t does not capture the similarity between different words that mean the same thing. Say we have three documents: "I run", "I ran", and "I ate". BoW vectors of all three documents will be equally apart.
- This representation does not have any way to handle out of vocabulary words (i.e., new words that were not seen in the corpus that was used to build the vectorizer).
- it is a "bag" of words—word order information is lost in this representation. Both D1 and D2 will have the same representation in this scheme.

Bag of N-grams

- BoN breaking text into chunks of n contiguous words (or tokens) called n-grams, instead of individual words
- This can help us capture some context,
- The corpus vocabulary, V, is then nothing but a collection of all unique n-grams across the text corpus.
- Then, each document in the corpus is represented by a vector of length —V—. This vector simply contains the frequency counts of n-grams present in the document and zero for the n-grams that are not present.

Bag of N-grams with the Toy Corpus

- The set of all bigrams in the corpus is as follows: dog bites, bites man, man bites, bites dog, dog eats, eats meat, man eats, eats food
- The bigram representation for the first two documents is as follows: D1 : [1,1,0,0,0,0,0,0], D2 : [0,0,1,1,0,0,0,0]

Bag of N-grams with the Toy Corpus

- The set of all bigrams in the corpus is as follows: dog bites, bites man, man bites, bites dog, dog eats, eats meat, man eats, eats food
- The bigram representation for the first two documents is as follows: D1 : [1,1,0,0,0,0,0,0], D2 : [0,0,1,1,0,0,0,0]
- Note that the BoW scheme is a special case of the BoN scheme, with n=1. n=2 is called a "bigram model," and n=3 is called a "trigram model." and so on.

Pros and Cons of Bag of N-grams

- It captures some context and word-order information in the form of n-grams.
- Thus, resulting vector space is able to capture some semantic similarity. Documents having the same n-grams will have their vectors closer to each other in Euclidean space as compared to documents with completely different n-grams.

Pros and Cons of Bag of N-grams

- It captures some context and word-order information in the form of n-grams.
- Thus, resulting vector space is able to capture some semantic similarity. Documents having the same n-grams will have their vectors closer to each other in Euclidean space as compared to documents with completely different n-grams.
- We still have the out of vocabulary issue, and large feature vector issue.

TF-IDF representation

- In BoW/BoNgrams, all the words in the text are treated as equally important—there's no notion of some words in the document being more important than others.
- TF-IDF, or term frequency–inverse document frequency, aims to quantify the importance of a given word relative to other words in the document and in the corpus.

TF-IDF representation

- In BoW/BoNgrams, all the words in the text are treated as equally important—there's no notion of some words in the document being more important than others.
- TF-IDF, or term frequency–inverse document frequency, aims to quantify the importance of a given word relative to other words in the document and in the corpus.

TF(t,d)=(Number of occurrences of term t in document d)/(Total number of terms in the document d)
IDF(t)=$\log_e$(Total number of documents in the corpus)/(Number of documents with term t in them)

Pros and Cons of TFIDF

- We can use the TF-IDF vectors to calculate similarity between two texts using a similarity measure like Euclidean distance or cosine similarity.
- TF-IDF is a commonly used representation in application scenarios such as information retrieval and text classification.

Pros and Cons of TFIDF

- We can use the TF-IDF vectors to calculate similarity between two texts using a similarity measure like Euclidean distance or cosine similarity.
- TF-IDF is a commonly used representation in application scenarios such as information retrieval and text classification.
- However, despite the fact that TF-IDF is better than the vectorization methods we saw earlier in terms of capturing similarities between words, it still suffers from the curse of high dimensionality.

Embeddings

Distributionally simliar words are words that are likely to occur in similar contexts.

- If we're given the word "USA," distributionally similar words could be other countries (e.g., Canada, Germany, India, etc.) or cities in the USA.
- If we're given the word "beautiful," words that share some relationship with this word (e.g., synonyms, antonyms) could be considered distributionally similar words.

Embeddings

Distributionally simliar words are words that are likely to occur in similar contexts.

- If we're given the word "USA," distributionally similar words could be other countries (e.g., Canada, Germany, India, etc.) or cities in the USA.
- If we're given the word "beautiful," words that share some relationship with this word (e.g., synonyms, antonyms) could be considered distributionally similar words.

Modern day NLP is based text representations which **learn** such semantic relationships in a dense, low dimensional space (compared to sparse, high dimensional space we saw earlier)

The OOV problem

- We can also train our own word embeddings, but both pre-trained and self-trained word embeddings depend on the vocabulary they see in the training data.
- What should we do when we encounter a new word in a document?

The OOV problem

- We can also train our own word embeddings, but both pre-trained and self-trained word embeddings depend on the vocabulary they see in the training data.
- What should we do when we encounter a new word in a document?
- A simple approach that often works is to exclude those words from the feature extraction process so we don't have to worry about how to get their representations.
- Another way to deal with the OOV problem for word embeddings is to create vectors that are initialized randomly, where each component is between –0.25 to +0.25, and continue to use these vectors.
- There are also other approaches that handle the OOV problem by modifying the training process by bringing in characters and other subword-level linguistic components.

Beyond Words: Contextual representations

- In all the representations we've seen so far, we notice that one word gets one fixed representation. Can this be a problem?
- Well, to some extent, yes. Words can mean different things in different contexts.
- For example, the sentences "I went to a bank to withdraw money" and "I sat by the river bank and pondered about text representations" both use the word "bank." However, they mean different things in each sentence.
- Well, to some extent, yes. Words can mean different things in different contexts. For example, the sentences "I went to a bank to withdraw money" and "I sat by the river bank and pondered about text representations" both use the word "bank." However, they mean different things in each sentence.

SOTA: Universal Text Representations

- Neural architectures such as recurrent neural networks (RNNs) and transformers were used to develop large-scale models of language (BERT, and beyond), which can be used as pre-trained models to get text representations.
- Process: Take a large pre-trained model, and "fine-tune" it to a given task/dataset.
- These models have shown significant improvements on some fundamental NLP tasks, such as question answering, semantic role labeling, named entity recognition, and coreference resolution, to name a few.
- There are a lot of custom models, especially for BERT, such as: SciBERT, LAWBert, FinBERT, PatentBERT, BioBERT etc.

NLP in Economics

Based on: Gentzkow, M., Kelly, B., Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535-74.

Example Use Cases

- In Financial economics, "text from financial news, social media, and company filings is used to predict asset price movements and study the causal impact of new information."
- "In macroeconomics, text is used to forecast variation in inflation and unemployment and estimate the effects of policy uncertainty"
- "In media economics, text from news and social media is used to study the drivers and effects of political slant."
- "In industrial organization and marketing, text from advertisements and product reviews is used to study the drivers of consumer decision making."
- "In political economy, text from politicians' speeches is used to study the dynamics of political agendas and debate."

Example Use Cases with Causal Inference goals

- using "Google search data to estimate local areas' racial animus, then studies the causal elect of racial animus on votes for Barack Obama in the 2008 election."
- using "congressional and news text to estimate each news outlet's political slant, then study the supply and demand forces that determine slant in equilibrium."
- "measure local news coverage of earnings announcements, then use the relationship between coverage and trading by local investors to separate the causal elect of news from other sources of correlation between news and stock prices."

Examples from a methods perspective

- authorship attribution approaches from NLP "to answer an authorship question of more direct interest to economists: who invented instrumental variables?"
- word frequencies are used for "a regression approach to construct an index of news-implied market volatility based on text from the Wall Street Journal from 1890-2009."
- topic models on diaries of CEOs to identify the behavioral patterns of leaders/managers.
- sentiment analysis to study financial stability

.... and so on.

NLP in Economics - summary

- Clearly, there are many use cases for NLP methods in various branches of economics.
- All sorts of texts from tweets to bank documents are used in these analyses.
- A range of NLP approaches, coupled with machine learning methods are commonly used for predictive analyses.
- Although not as common, using these methods to understand causal inferences is also not uncommon.

Summary of today's class

- Course overview
- NLP overview
- Representing text as a vector
- NLP in Economics

Tomorrow

- Python overview
- Working with text in Python - overview

Note: Next class is a hands on one. So, please have laptops ready. You can work with Google Colab Notebooks, and you don't have to install anything locally. If you want to, then, install Python, and Jupyter Notebook.