TRAINING ON AI AND MACHINE LEARNING WITH PYTHON

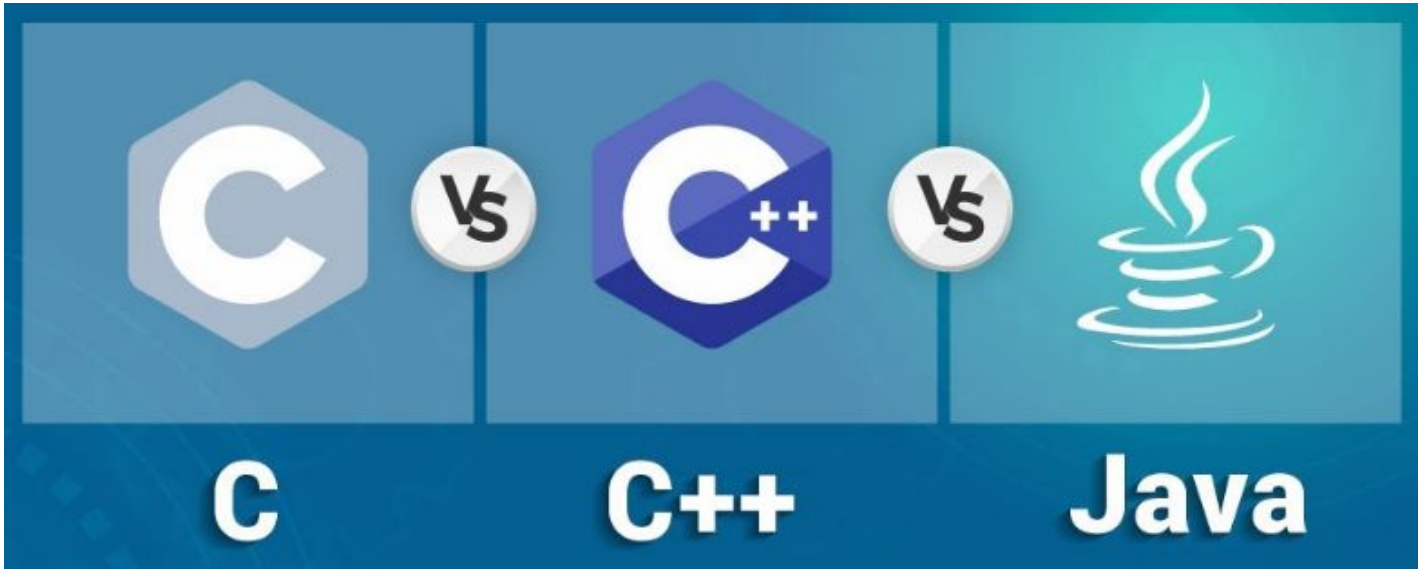
ORGANIZED BY
CUET IT BUSINESS INCUBATOR

python™

MACHINE
LEARNING

Natural Language

What is Natural Language?



- Referred to the language that is spoken by people; such as English, Chinese, Bengali, etc.

Natural Language Processing

What is Natural Language Processing?

- A subfield of artificial intelligence such as computer vision.
- Application of computational techniques to the analysis and synthesis of natural language and speech.

<Goal>

To build intelligent systems that can interact with a human being like a human being !!!

Have you ever used NLP-powered products?

What is NLP?

Question Answering

What is NLP?

Question Answering

Google

which countries border the black sea

All Maps Images News Shopping More Search tools

About 2,710,000 results (0.81 seconds)

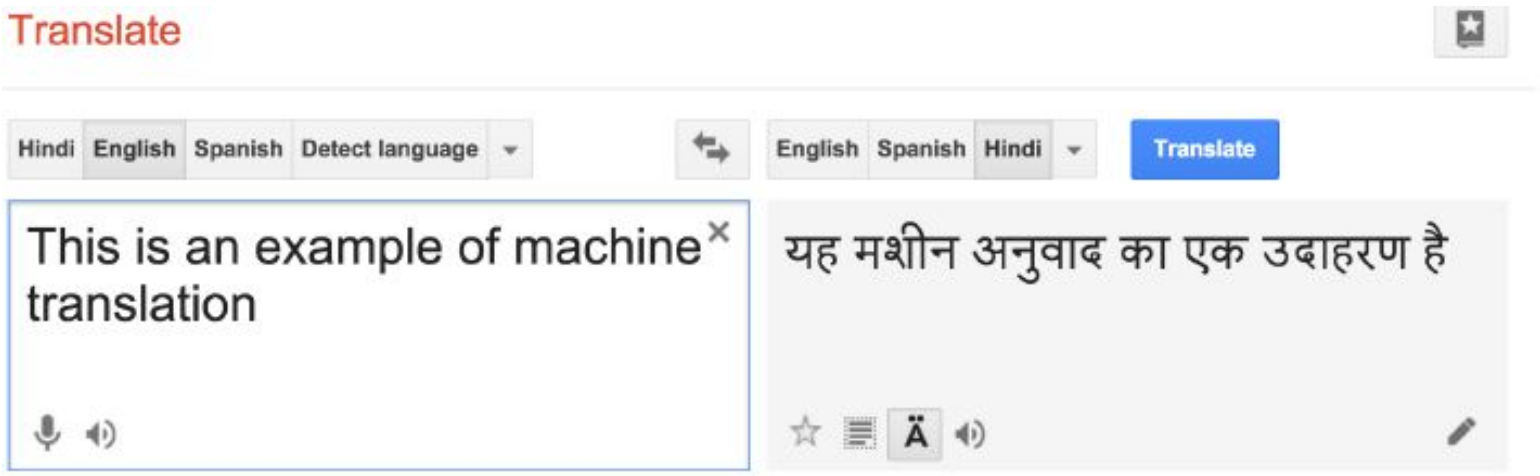
This major inland sea is bordered by six countries — **Romania** and **Bulgaria** to the west; **Ukraine**, **Russia**, and **Georgia** to the north and east; and **Turkey** to the south. Additionally, it is impacted by another 10 nations through the five major rivers that empty into the Black Sea, the largest of which is the Danube River.

Black Sea Geography - College of Earth, Ocean, and Environment

https://www.ceoe.udel.edu/**blacksea**/geography/index.html University of Delaware

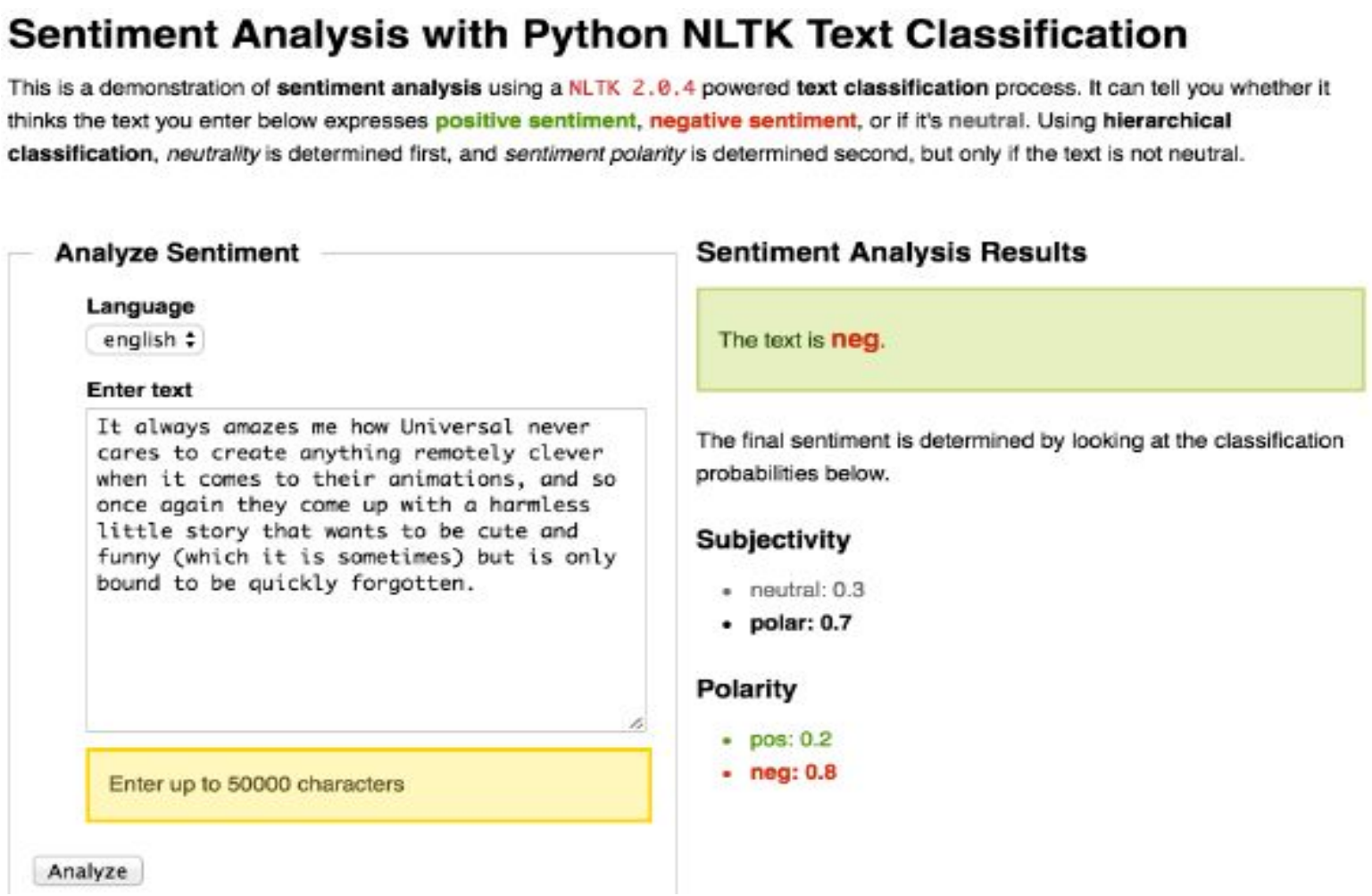
What is NLP?

Machine Translation



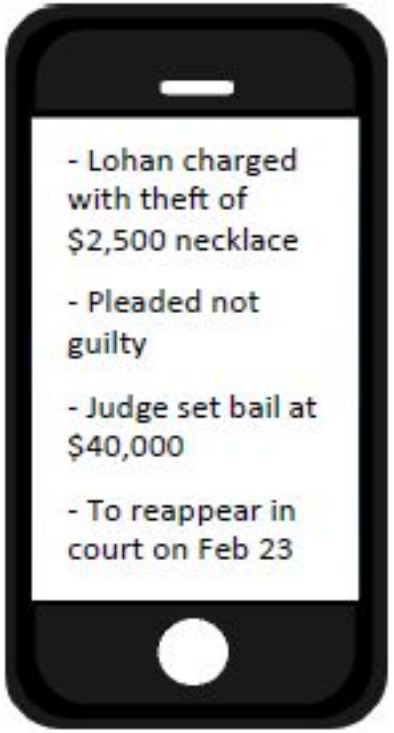
What is NLP?

Sentiment Analysis



What is NLP?

Natural Language Generation: Summarization



What is NLP?

Image Captioning

Tags

- authors
- scones
- luncheon
- breakfast
- seder

Nearest Caption in the Training Dataset

a man cuts a cake while children sit around at the table , looking on .

Generated Captions

- two people at a table with a cake .
- the two people are having a meal that is in a party .
- a man and two children in a blue table with a cake .
- a man sitting at a table with a bunch of cake on it .
- a man and woman sitting at a table with cake at a party .

back

What is NLP?

Video Captioning

Ground truth: A man is playing a violin.
A man is playing the violin on stage.
Baseline-XE: A man is playing the drums.
CIDEr-RL: A man is playing a guitar.
CIDEnt-RL: A man is playing a violin.

Ground truth: Two men are wrestling.
Two guys are wrestling in a competition.
Baseline-XE: A group of people are playing a game.
CIDEr-RL: A man is playing a wrestling.
CIDEnt-RL: Two men are wrestling.

What is NLP?

Visual Question Answering

What color are her eyes?
What is the mustache made of?

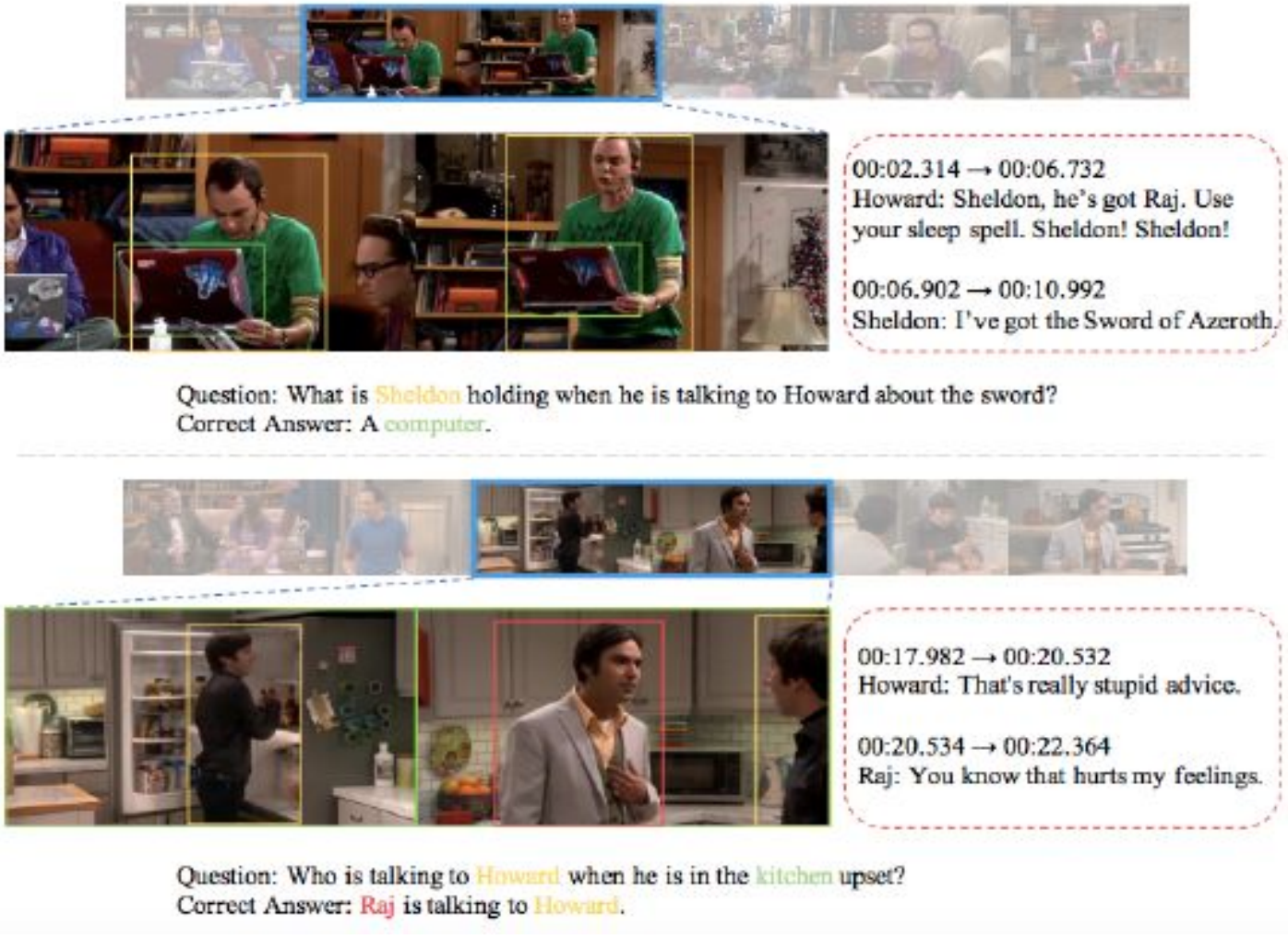
How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

What is NLP?

Video + Subtitle Question Answering



What is NLP?

Automatic Speech Recognition

Applications

- Key applications,
  - ✔ Conversational agents
  - ✔ Information extraction and question answering
  - ✔ Machine translation
  - ✔ Opinion and sentiment analysis
  - ✔ Social media analysis
  - ✔ Visual understanding
  - ✔ Essay evaluation
  - ✔ Mining legal, medical, or scholarly literature, etc.

Applications

- Fields with connections to NLP,
  - ✔ Machine learning, Deep Learning
  - ✔ Linguistics (including psycho-, socio-, and theoretical)
  - ✔ Cognitive science
  - ✔ Information theory
  - ✔ Data science
  - ✔ Political science
  - ✔ Psychology
  - ✔ Economics
  - ✔ Education.

Why Now?

Factors Changing NLP Landscape

1) Increases in computing power
2) The rise of the web, then the social web
3) Advances in machine learning
4) Advances in the understanding of language in the social context

Is NLP Hard?

Why NLP is Hard?

A ship-shipping ship, shipping shipping-ships

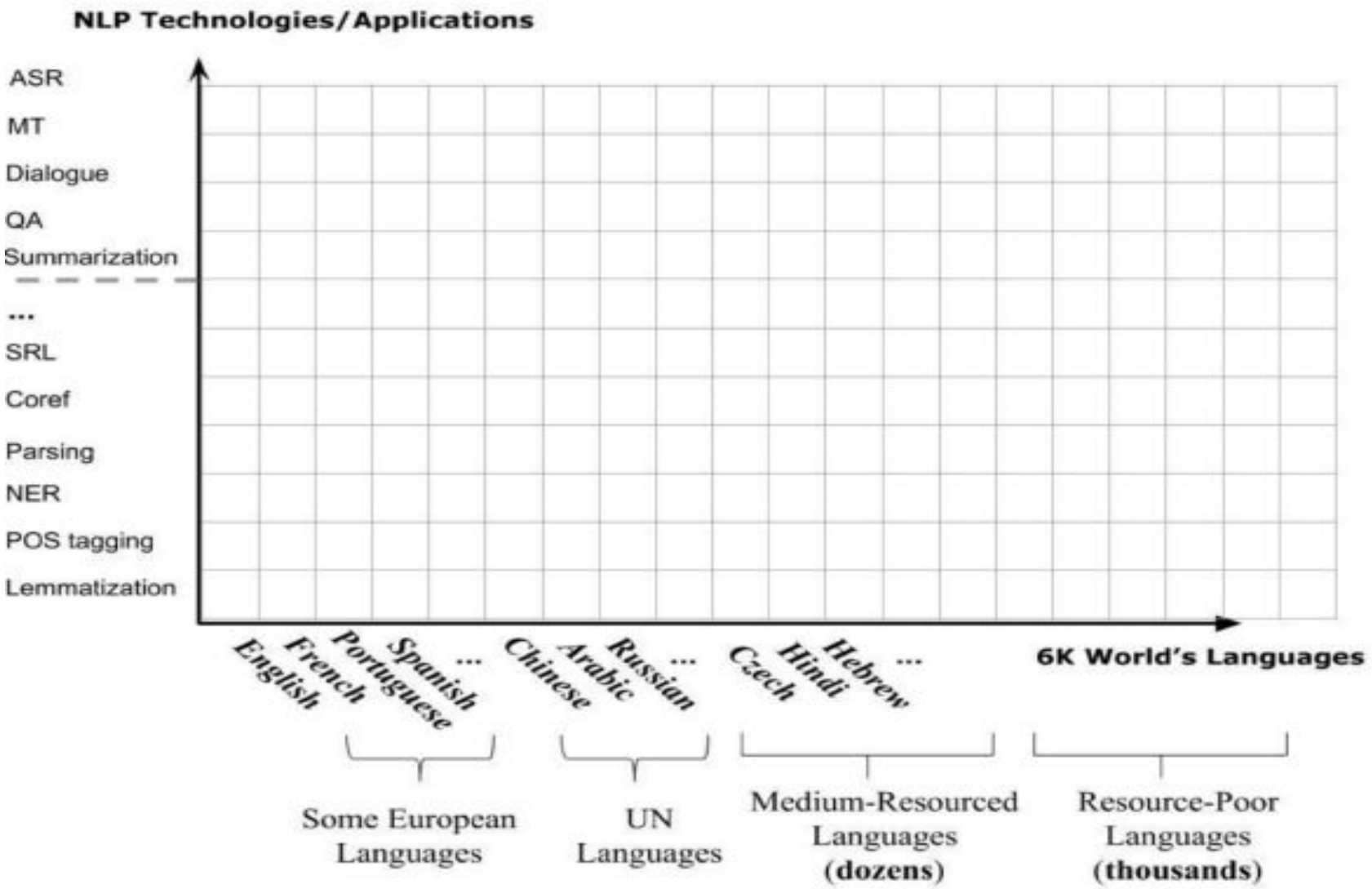
Why NLP is Hard?

1) Ambiguity
2) Scale
3) Sparsity
4) Variation
5) Expressivity
6) Unmodeled Variables
7) Unknown representations

Ambiguity

- Ambiguity at multiple levels
  - Word senses: bank (finance or river ?)
  - Part of speech: chair (noun or verb ?)
  - Syntactic structure: I can see a man with a telescope
  - Multiple: I made her duck

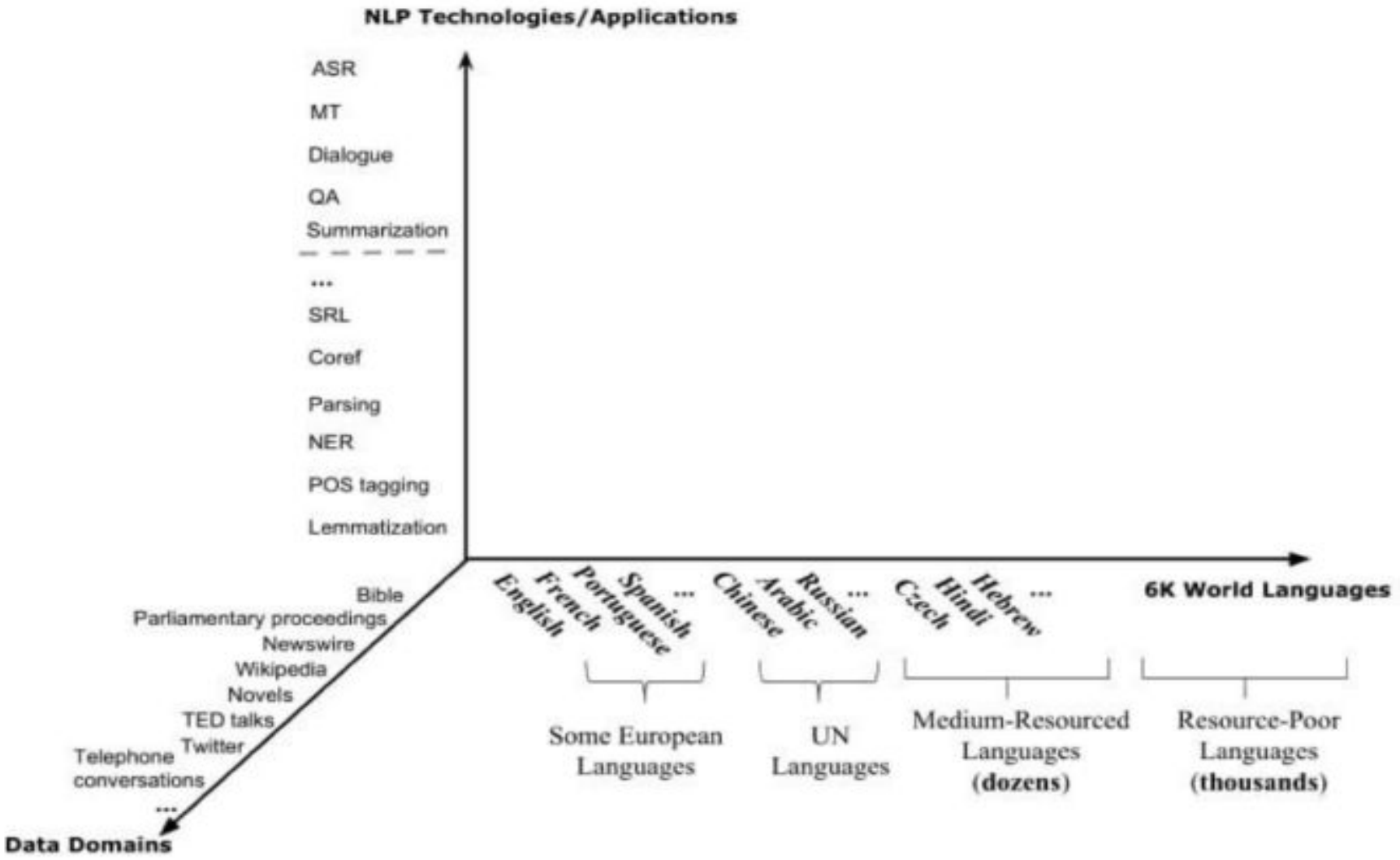
Scale


NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
6K World's Languages
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)

Sparsity

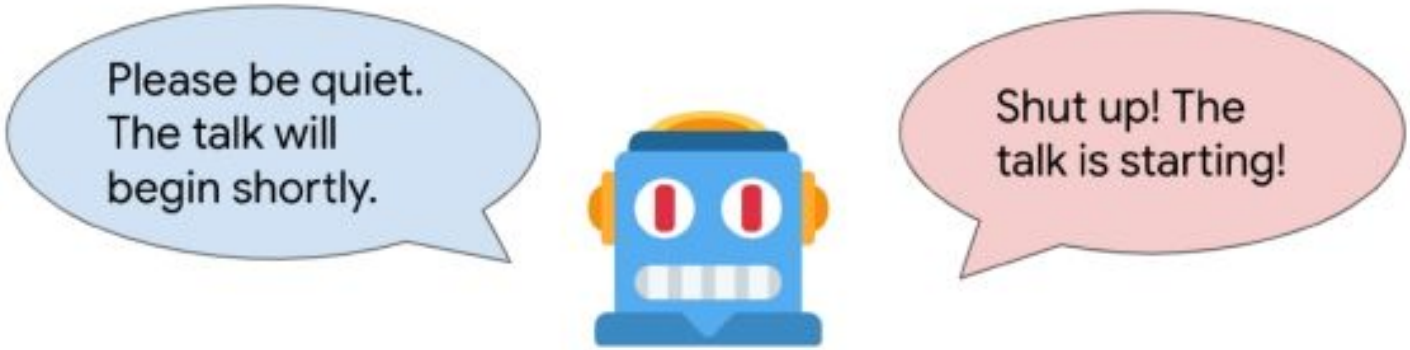
- Regardless of how large our corpus is, there will be a lot of infrequent words
- We need to find clever ways to estimate probabilities for things we have rarely or never seen

Variation


NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
Bible
Parliamentary proceedings
Newswire
Wikipedia
Novels
TED talks
Twitter
Telephone conversations
...
Data Domains
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
6K World Languages
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)

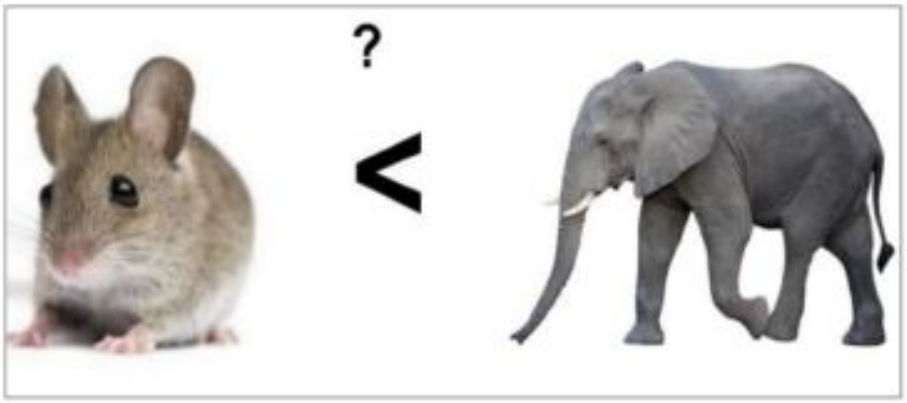
Expressivity

- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:
  - She gave the book to Tom vs. She gave Tom the book
  - Some kids popped by vs. A few children visited
  - Is that window still open? vs. Please close the window



Unmodeled Variables

"Drink this milk"



World knowledge

I dropped the glass on the floor and it broke

I dropped the hammer on the glass and it broke

Unmodeled Representation

- We don't even know how to represent the knowledge a human has/needs:
  - What is the "meaning" of a word or sentence?
  - How to model context?
  - Other general knowledge?

Expectations

What we should expect from NLP models

1) Sensitivity to a wide range of phenomena and constraints in human language

2) Generality across languages, modalities, genres, styles

3) Strong formal guarantees (e.g., convergence, statistical efficiency, consistency)

4) High accuracy when judged against expert annotations or test data

5) Ethical


time for
a break!

Let’s get our hands dirty!!!

Classification Task

- A mapping h from input data x (drawn from instance space X) to a label y from some enumerable output space Y.
  - X = set of all documents
  - Y = {English, Mandarin, Greek, …}
  - x = a single document
  - y = ancient Greek

Movie Ratings

positive

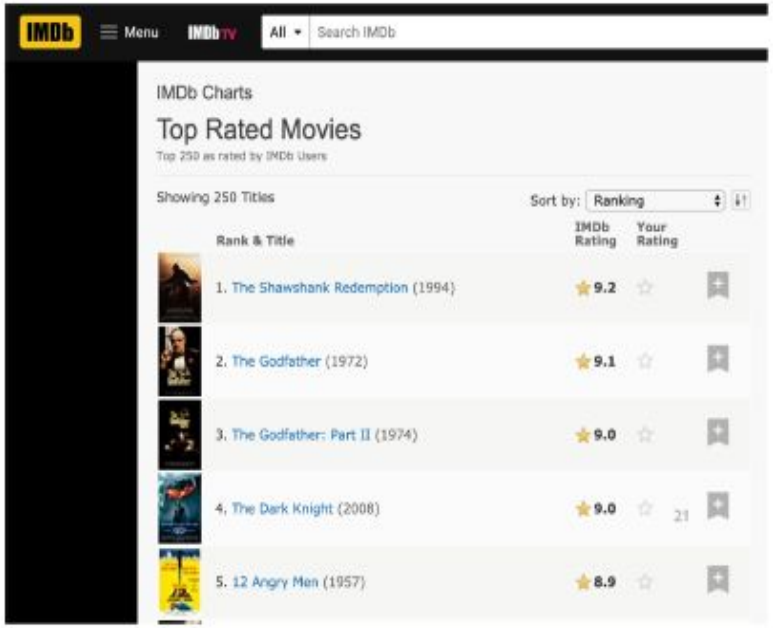
"… is a film which still causes real, not figurative, chills to run along my spine, and it is certainly the bravest and most ambitious fruit of Coppola's genius"

Roger Ebert, Apocalypse Now

- "I hated this movie. Hated hated hated hated hated this movie. Hated it. Hated every simpering stupid vacant audience-insulting moment of it. Hated the sensibility that thought anyone would like it."

Roger Ebert, North

negative



Customer Review

NOT DISHWASHER SAFE

Reviewed in the United States on April 5, 2019

Color: Blue | Verified Purchase

Used the bottle for one day. There was a slight lid leak, but I was willing to overlook that because I liked the other aspects of the product. Put it in the dishwasher with my other water bottles, air dry, and it melted. There is nothing in the product description that indicates it is not dishwasher safe, nor was there a product sheet included with the bottle indicating to hand wash only. I have a number of plastic water bottles that I routinely send through the dishwasher on this setting and have never had a problem. Extremely disappointed!

19 people found this helpful

Helpful | Comment | Report abuse

Makes Drinking Water Fun

Reviewed in the United States on March 31, 2019

Color: Transparent | Verified Purchase

It is always a challenge to drink the recommended amount of water each day, so important for health. This bottle makes it fun while serving as a reminder to keep drinking! Bottle is good quality, handle makes it easy to lift.

14 people found this helpful

Customer reviews

4.5 out of 5

451 customer ratings

5 star	78%
4 star	9%
3 star	5%
2 star	2%
1 star	6%

By feature

Sturdiness	4.5
Flavor	4.5
Durability	4.4

Political Opinion Mining

Is This Spam?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

What Is the Subject of This Article?

MEDLINE Article

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

Direct Text Classification Applications

Task	x	y
Language identification	text	{English, Mandarin, Greek, ...}
Spam classification	email	{spam, not spam}
Authorship attribution	text	{jk rowling, james joyce, ...}
Genre classification	novel	{detective, romance, gothic, ...}
Sentiment classification	text	{positive, negative, neutral, mixed}

Direct Text Classification Applications

Task	x	y
Language identification	text	{English, Mandarin, Greek, ...}
Spam classification	email	{spam, not spam}
Authorship attribution	text	{jk rowling, james joyce, ...}
Genre classification	novel	{detective, romance, gothic, ...}
Sentiment classification	text	{positive, negative, neutral, mixed}

Our Focus

Emotion Detection in Text

- Task of automatically attributing an emotion category to a textual document.
- Basic emotions: 6 types



Joy Sadness Anger Fear Disgust Surprise



Go to Google Colab!!!

Text preprocessing

What is Text?

- **You can think of text as a sequence of**

  - Characters
  - **Words**
  - Phrases and named entities
  - Sentences
  - Paragraphs
  - …

What is Word?

- **It seems natural to think of a text as a sequence of words**
  - A word is a meaningful sequence of characters
- **How to find the boundaries of words?**
  - In English we can split a sentence by spaces or punctuation

Input: Friends, Romans, Countrymen, lend me your ears;

Output:

Friends	Romans	Countrymen	lend	me	your	ears

- In German there are compound words which are written without spaces
  - "Rechtsschutzversicherungsgesellschaften" stands for "insurance companies which provide legal protection"
- In Japanese there are no spaces at all!
  - Butyoucanstillreaditright?

Tokenization

- **Tokenization is a process that splits an input sequence into so-called tokens**
  - You can think of a token as a useful unit for semantic processing
  - Can be a word, sentence, paragraph, etc.

- **An example of simple whitespace tokenizer**
  - nltk.tokenize.WhitespaceTokenizer

This	is	Andrew's	text,	isn't	it?

  - Problem: "it" and "it?" are different tokens with same meaning

Tokenization

- **Let's try to also split by punctuation**
  - nltk.tokenize.WordPunctTokenizer

This	is	Andrew	'	s	text	,	isn	'	t	it	?

  - Problem: "s", "isn", "t" are not very meaningful

- **We can come up with a set of rules**
  - nltk.tokenize.TreebankWordTokenizer

This	is	Andrew	's	text	,	is	n't	it	?

  - "'s" and "n't" are more meaningful for processing

Tokenization

```
import nltk
text = "This is Andrew's text, isn't it?"
```

```
tokenizer = nltk.tokenize.WhitespaceTokenizer()
tokenizer.tokenize(text)
```

```
['This', 'is', "Andrew's", 'text,', "isn't", 'it?']
```

```
tokenizer = nltk.tokenize.TreebankWordTokenizer()
tokenizer.tokenize(text)
```

```
['This', 'is', 'Andrew', "'s", 'text', ',', 'is', "n't",
'it', '?']
```

```
tokenizer = nltk.tokenize.WordPunctTokenizer()
tokenizer.tokenize(text)
```

```
['This', 'is', 'Andrew', "'", 's', 'text', ',', 'isn',
 "'", 't', 'it', '?']
```

Token Normalization

- **We may want the same token for different forms of the word**
  - wolf, wolves → wolf
  - talk, talks → talk
- **Stemming**
  - A process of removing and replacing suffixes to get to the root form of the word, which is called the **stem**
  - Usually refers to heuristics that chop off suffixes
- **Lemmatization**
  - Usually refers to doing things properly with the use of a vocabulary and morphological analysis
  - Returns the base or dictionary form of a word, which is known as the **lemma**

Stemming Example

Porter's stemmer

- 5 heuristic phases of word reductions, applied sequentially
- Example of phase 1 rules:

Rule	Example
SSES → SS	caresses → caress
IES → I	ponies → poni
SS → SS	caress → caress
S →	cats → cat

- nltk.stem.PorterStemmer
- Examples:
  - feet feet
  - wolves wolv

  cats cat
  talked talk
- Problem: fails on irregular forms, produces non-words

Lemmatization Example

- **WordNet lemmatizer**

  - Uses the WordNet Database to lookup lemmas
  - nltk.stem.WordNetLemmatizer
  - Examples:
    - feet foot
    - wolves wolf

    cats cat
    talked talked

- Problems: not all forms are reduced

- Takeaway: we need to try stemming or lemmatization and choose best for our task

Stemming Example

```
import nltk
text = "feet cats wolves talked"
tokenizer = nltk.tokenize.TreebankWordTokenizer()
tokens = tokenizer.tokenize(text)
```

```
stemmer = nltk.stem.PorterStemmer()
" ".join(stemmer.stem(token) for token in tokens)
```

```
u'feet cat wolv talk'
```

```
stemmer = nltk.stem.WordNetLemmatizer()
" ".join(stemmer.lemmatize(token) for token in tokens)
```

```
u'foot cat wolf talked'
```

Transforming tokens into features

Bag of Words (BOWs)

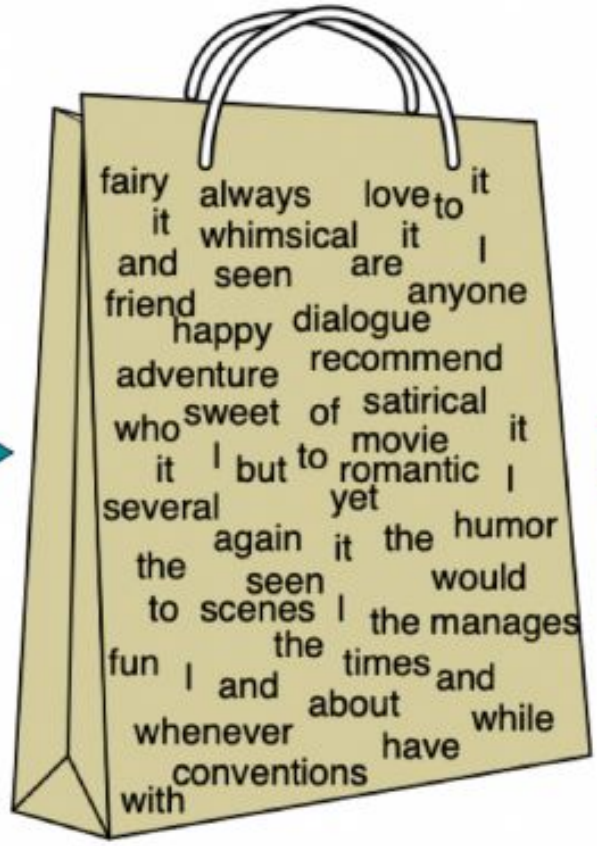
- **Let's count occurrences of a particular token in our text**
  - For each token we will have a feature column, this is called **text vectorization**.

	good	movie	not	a	did	like
good movie	**1**	**1**	0	0	0	0
not a good movie	**1**	**1**	**1**	**1**	0	0
did not like	0	0	**1**	0	**1**	**1**

- Problems:
  - we lose word order, hence the name "bag of words"
  - counters are not normalized

Bag of Words (BOWs)

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!



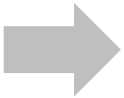
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Bag of Words (BOWs)

Let's preserve some ordering

- **We can count token pairs, triplets, etc.**
  - Also known as n-grams
    - 1-grams for tokens
    - 2-grams for token pairs
  - …

good movie
not a good movie
did not like



good movie	movie	did not	a	…
1	**1**	0	0	…
1	**1**	0	**1**	…
0	0	**1**	0	…

- Problems:
  - too many features

Bag of Words (BOWs)

- **Let's remove some n-grams from features based on their occurrence frequency in documents of our corpus**
  - **High-frequency n-grams:**
    - Articles, prepositions, etc. (example: and, a, the)
    - They are called **stop-words**, they won't help us to discriminate texts remove them
  - **Low-frequency n-grams:**
    - Typos, rare n-grams
    - We don't need them either, otherwise we will likely overfit
  - **Medium frequency n-grams:**
    - Those are good n-grams

Bag of Words (BOWs)

- **Idea**: the n-gram with smaller frequency can be more discriminating because it can capture a specific issue in the review

TF-IDF

Term frequency (TF)

- $\text{tf}(t, d)$ – frequency for term (or n-gram) t in document d
- Variants:

weighting scheme	TF weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} / \sum_{t' \in d} f_{t',d}$
log normalization	$1 + \log(f_{t,d})$

TF-IDF

Inverse document frequency (IDF)

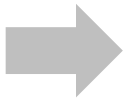
- $N = |D|$ – total number of documents in corpus
- $|\{d \in D: t \in d\}|$ – number of documents where the term t appears
- $\text{idf}(t, D) = \log \frac{N}{|\{d \in D: t \in d\}|}$

TF-IDF

- $\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$
- A high weight in TF-IDF is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents

TF-IDF

good movie
not a good movie
did not like



good movie	movie	did not	…
0.17	**0.17**	0	…
0.17	**0.17**	0	…
0	0	**0.47**	…