MSCI 641: Text analytics

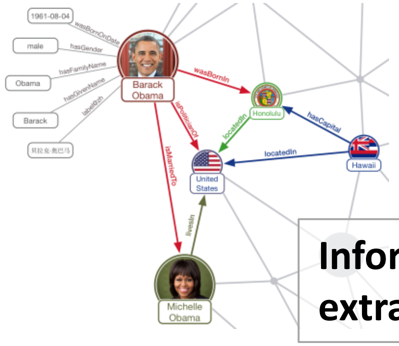
What is Text Analytics and Natural Language Processing (NLP)?

What is Natural Language Processing?

Information retrieval

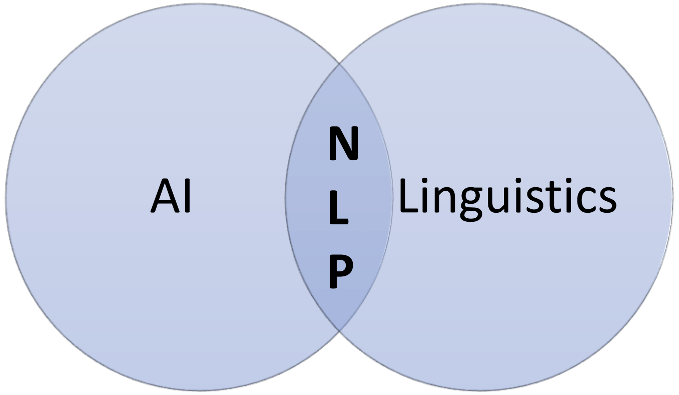
Acme Article
Technology
Sports
Entertainment

Text classification



Information extraction

Machine translation



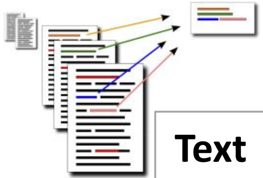
AI

NLP

Linguistics

Question answering

Sentiment analysis



Text summarization

NLP research trend: from feature engineering to deep learning

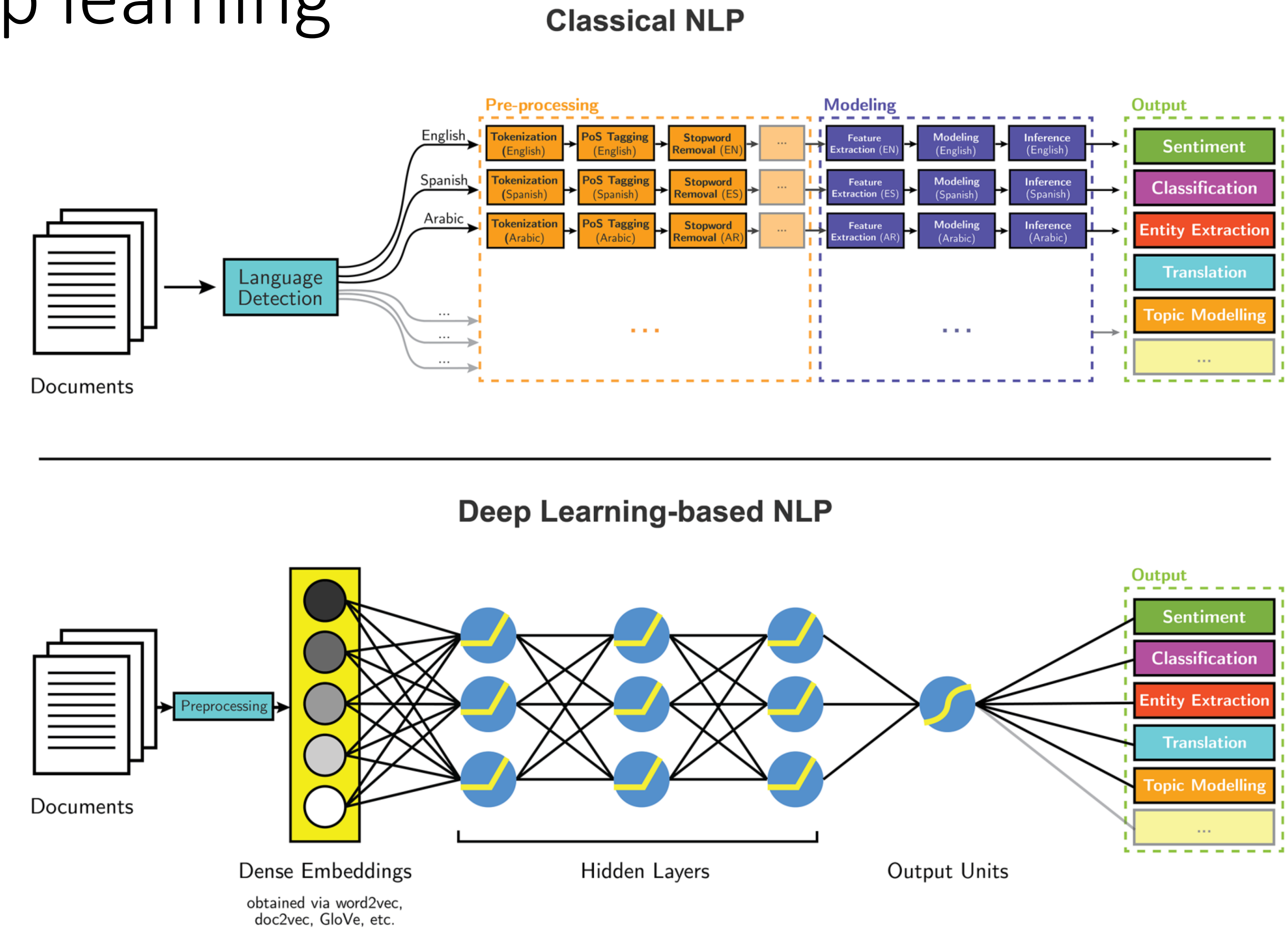
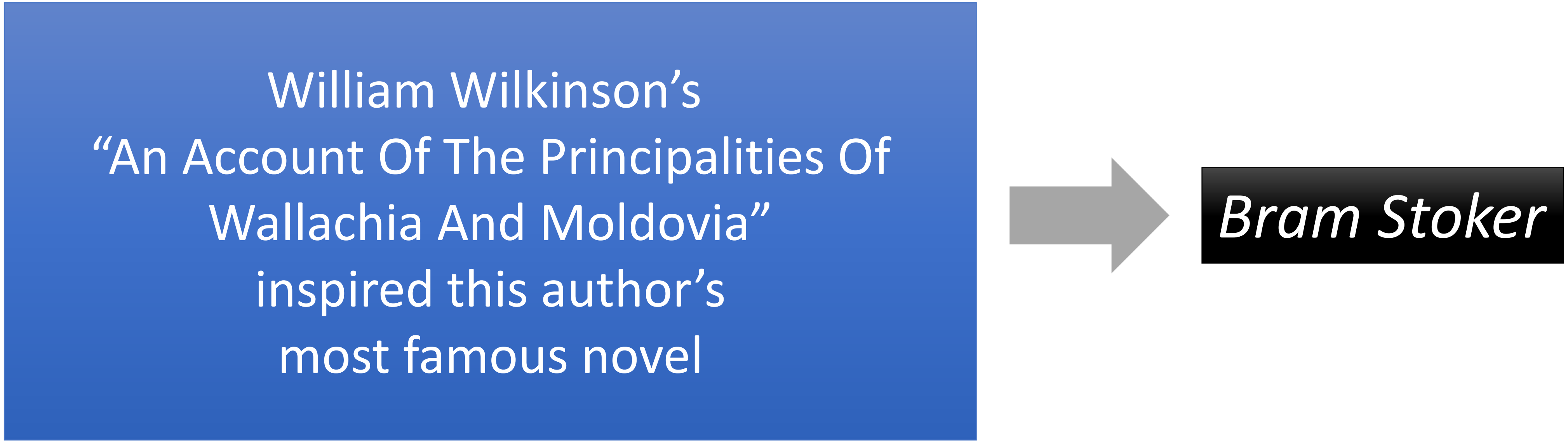


Image source: https://s3.amazonaws.com/aylien-main/misc/blog/images/nlp-language-dependence-small.png

Question Answering

- IBM Watson won Jeopardy on February 16, 2011

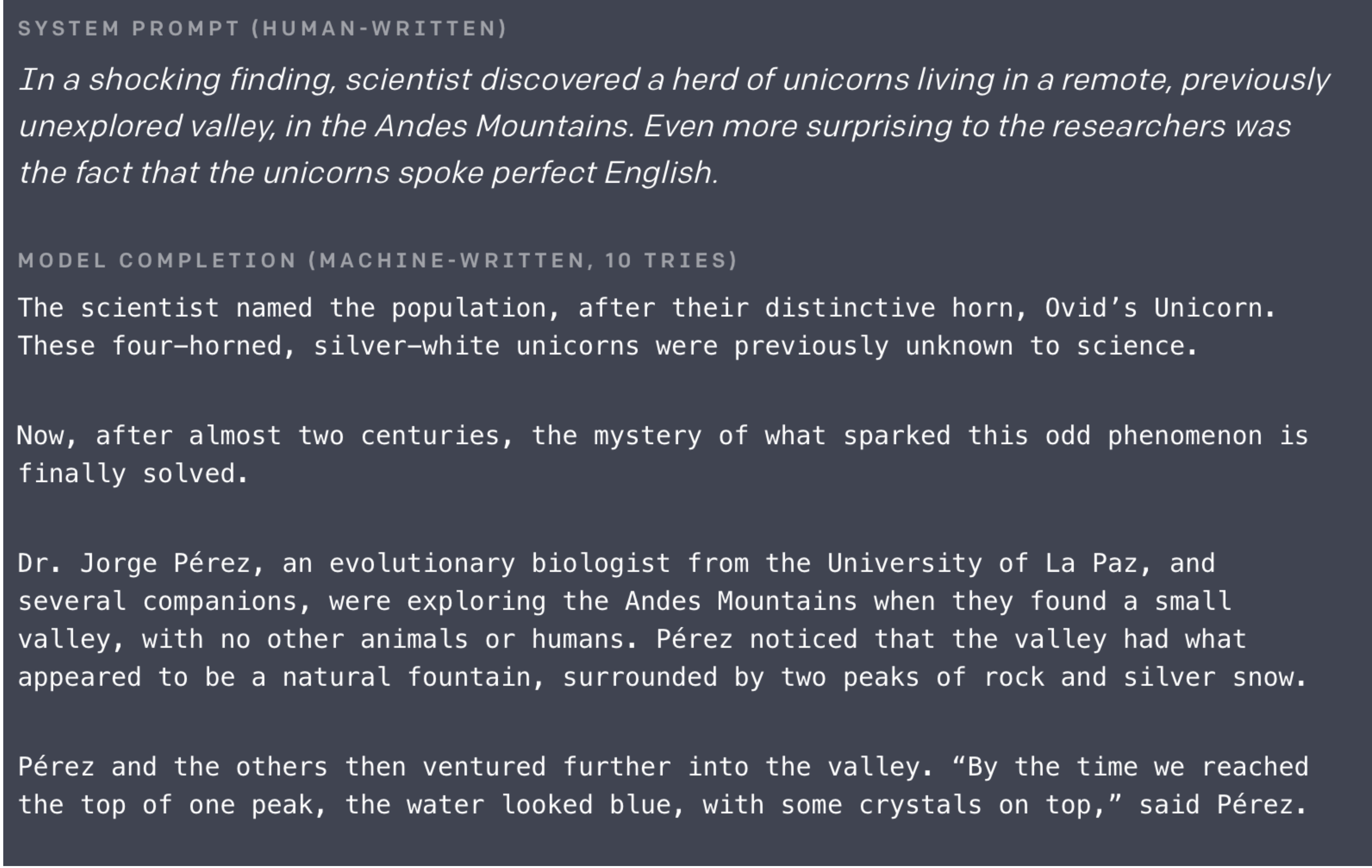


William Wilkinson's
"An Account Of The Principalities Of
Wallachia And Moldovia"
inspired this author's
most famous novel

Bram Stoker

Text generation:

OpenAI GPT-2 model



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

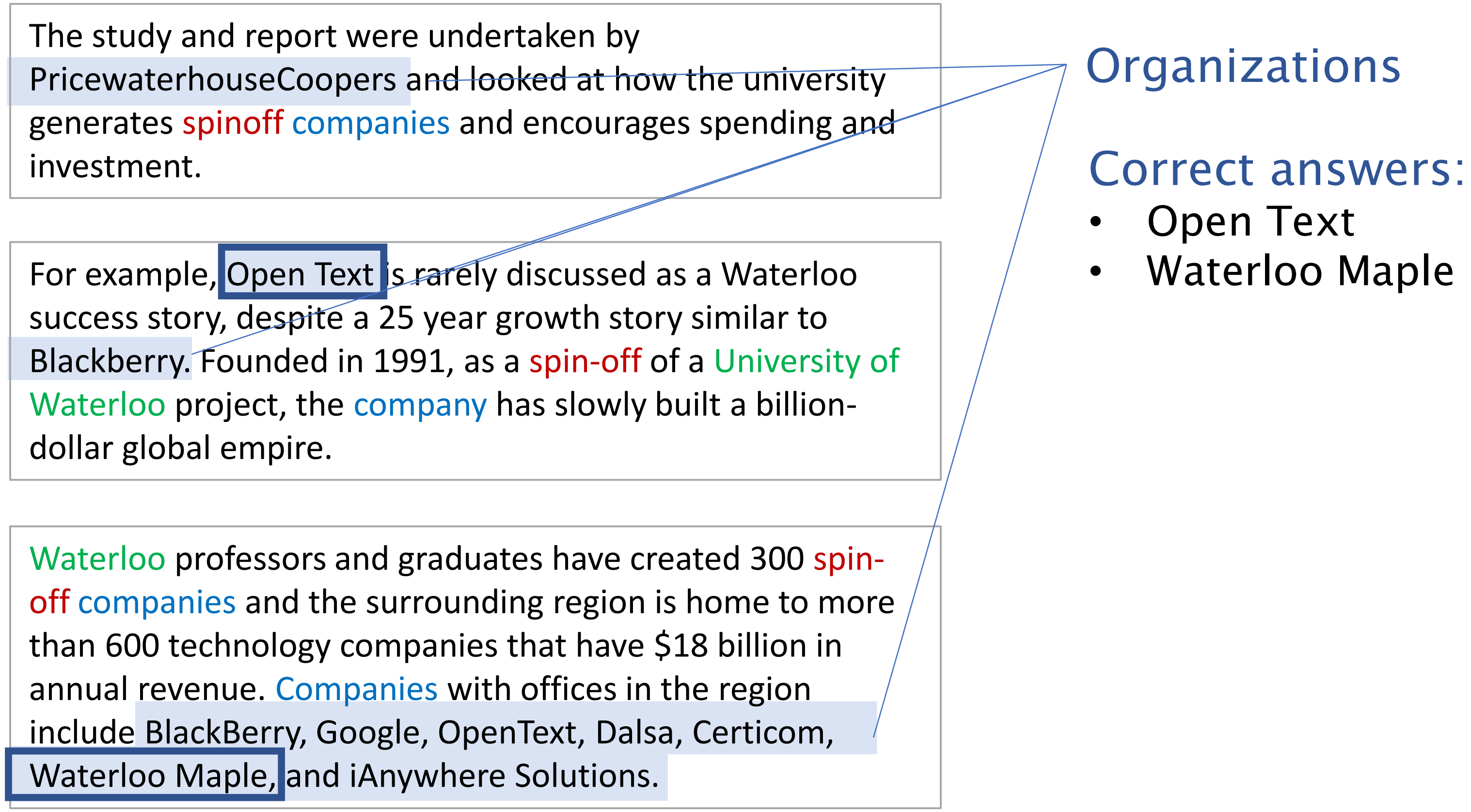
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

https://openai.com/blog/better-language-models/

Information Extraction and Question Answering

What are some of the spin-off companies from the University of Waterloo?



The study and report were undertaken by PricewaterhouseCoopers and looked at how the university generates spinoff companies and encourages spending and investment.

For example, Open Text is rarely discussed as a Waterloo success story, despite a 25 year growth story similar to Blackberry. Founded in 1991, as a spin-off of a University of Waterloo project, the company has slowly built a billion-dollar global empire.

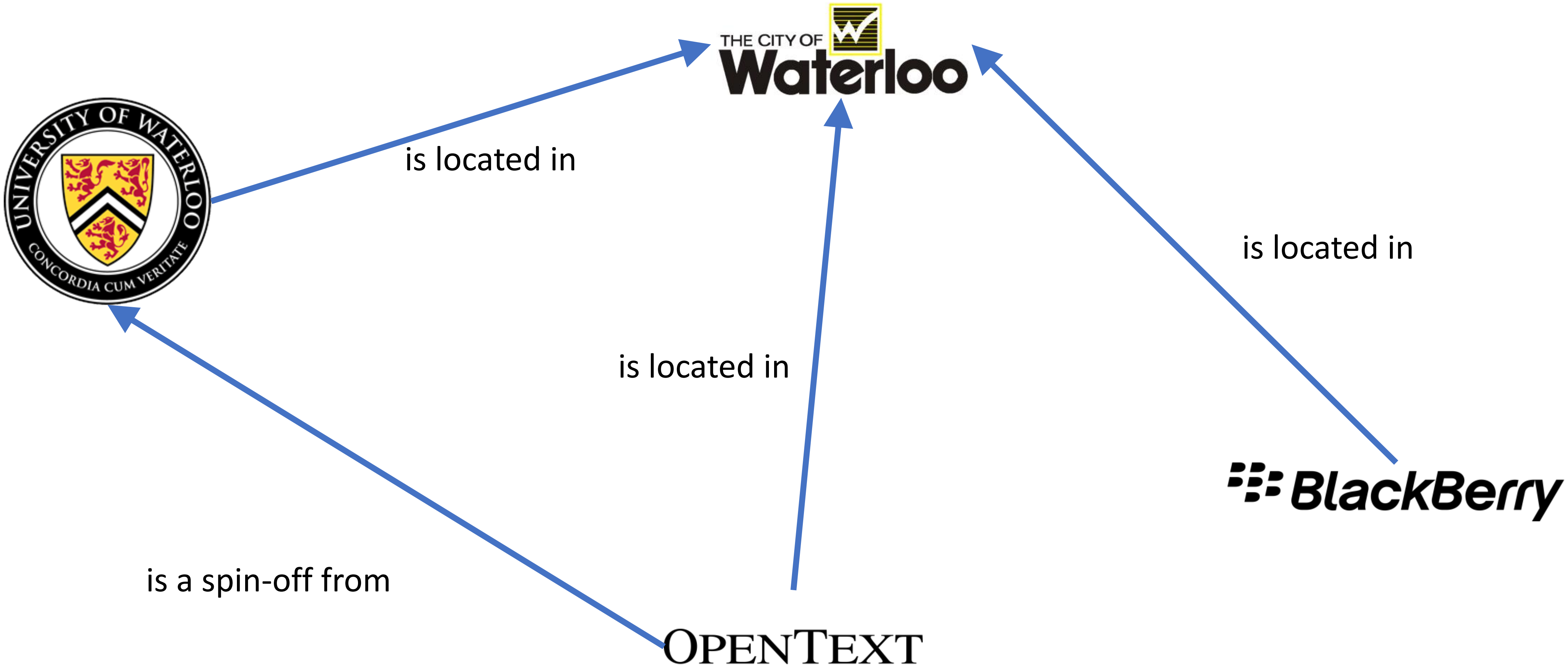
Waterloo professors and graduates have created 300 spin-off companies and the surrounding region is home to more than 600 technology companies that have $18 billion in annual revenue. Companies with offices in the region include BlackBerry, Google, OpenText, Dalsa, Certicom, Waterloo Maple, and iAnywhere Solutions.

Organizations

Correct answers:

- Open Text
- Waterloo Maple

IE and knowledge graph generation



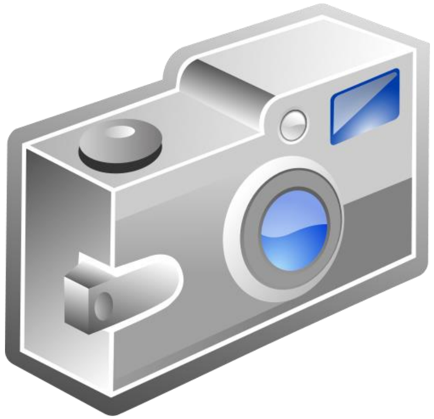
is located in

is located in

is located in

is a spin-off from

Information Extraction & Sentiment Analysis



Attributes:

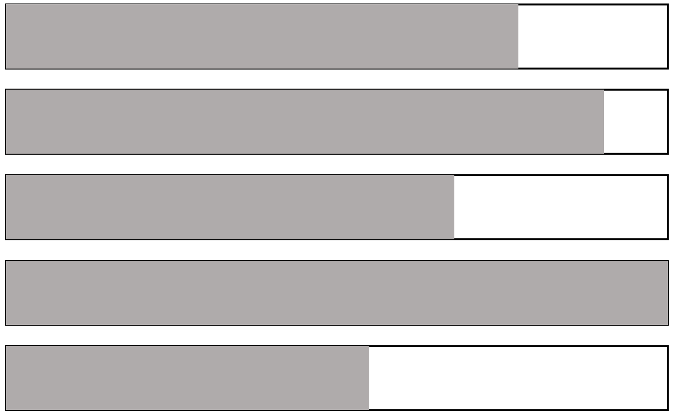
zoom

affordability

size and weight

flash

ease of use



Size and weight

- ✓ nice and compact to carry!
- ✓ since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!
- ✗ the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Machine Translation

- Fully automatic

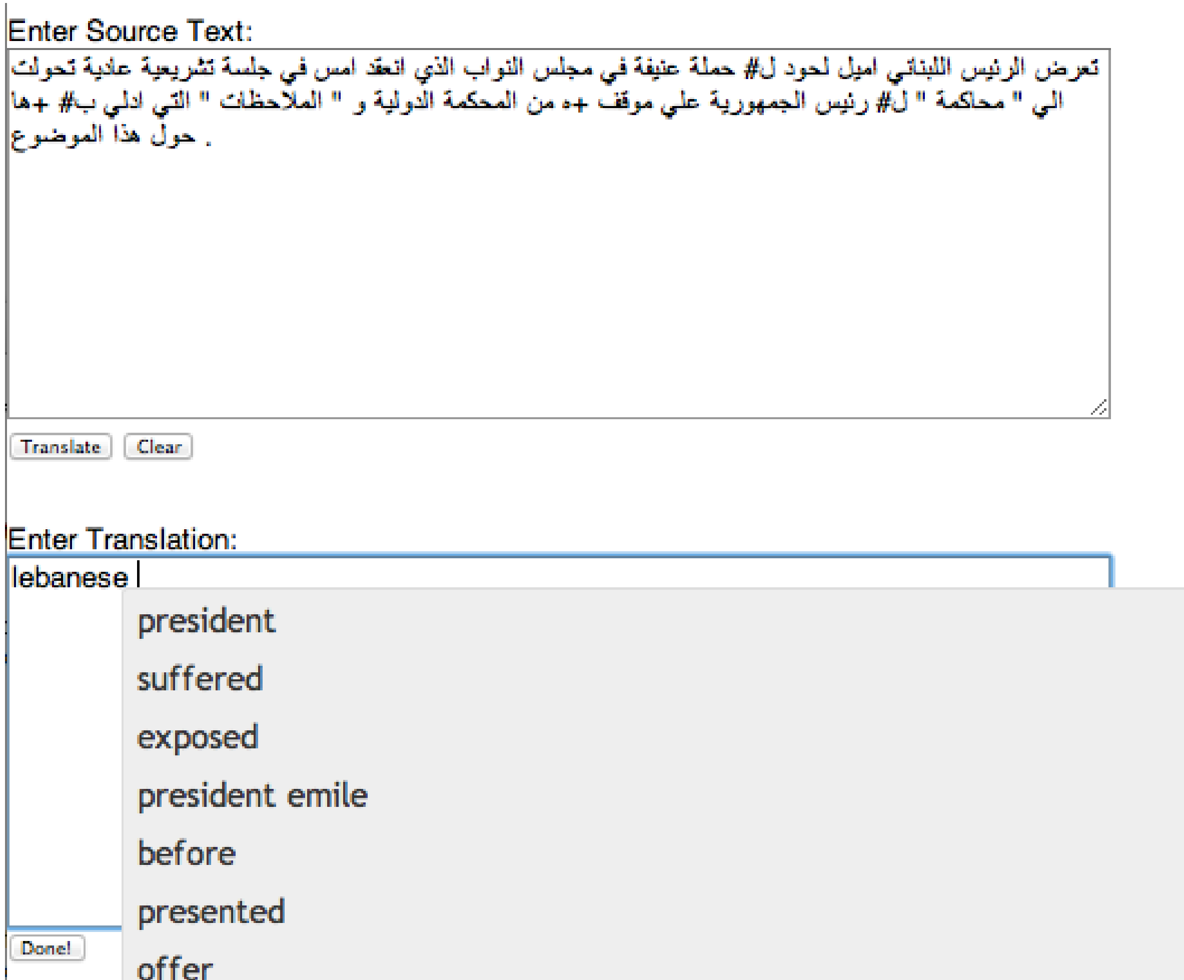
Enter Source Text:

这 不过 是 一 个 时间 的 问题 .

Translation from Stanford’s *Phrasal*:

This is only a matter of time.

- Helping human translators



Enter Source Text:

تعرض الرئيس اللبناني اميل لحود ل# حملة عنيفة في مجلس النواب الذي انعقد امس في جلسة تشريعية عادية تحولت
الي " محاكمة " ل# رئيس الجمهورية علي موقف +ه من المحكمة الدولية و " الملاحظات " التي ادلي ب# +ها
حول هذا الموضوع .

Enter Translation:

lebanese

Language Technology

mostly solved

Spam detection

Let’s go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON			ORG		LOC
Einstein	met	with	UN	officials in	Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco! 👍

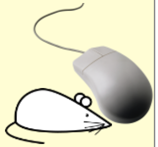
The waiter ignored us for 20 minutes. 👎

Coreference resolution

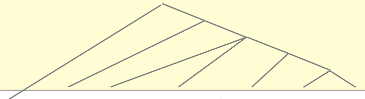
Carter told Mubarak he shouldn’t run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.



Parsing



I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You’re invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Fake news detection

Woolly mammoths found living in Siberia ✗

Woolly mammoths on verge of resurrection ✓

Style transfer

Hey, how’s it going? → Good morning, how are you?

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Ambiguity makes NLP hard:

100% REAL

Violinist Linked to JAL Crash Blossoms

Teacher Strikes Idle Kids

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either ❤

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

Where is *A Bug’s Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene …

But that’s what makes it fun!

Linguistics Levels of Analysis

- Speech
- Written language
    - Phonology: sounds / letters / pronunciation
    - Morphology: the structure of words
    - Syntax: how these sequences are structured
    - Semantics: meaning of the strings
- Interaction between levels

Issues in Syntax

“the dog ate my homework” - Who did what?

1. Identify the part of speech (POS)

   Dog = noun ; ate = verb ; homework = noun

   English POS tagging: 95%

2. Identify collocations

   mother in law, hot dog

   Compositional versus non-compositional collocates

Issues in Syntax

- Shallow parsing:

  “the dog chased the bear”

  “the dog” “chased the bear”

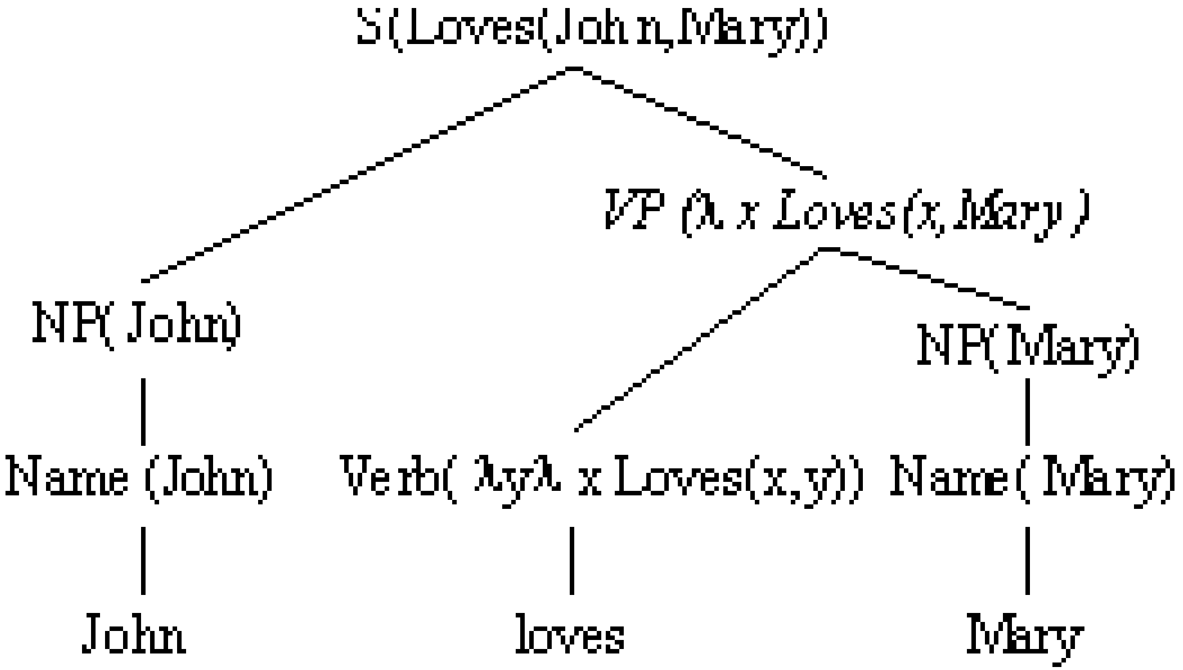
  subject   - predicate

  Identify basic structures

  NP-[the dog] VP-[chased the bear]

Issues in Syntax

- Full parsing: John loves Mary



Help figuring out (automatically) questions like: Who did what and when?

More Issues in Syntax

- Anaphora Resolution:

“The dog entered my room. It scared me ”

- Preposition Attachment

“I saw the man in the park with a telescope”

Issues in Semantics

- Understand language! How?
- *“plant ” = industrial plant*
- *“plant ” = living organism*
- Words are ambiguous
- Importance of semantics?
    - Machine Translation: wrong translations
    - Information Retrieval: wrong information
    - Anaphora Resolution: wrong referents

Why Semantics?

The sea is at the home for billions factories and animals

The sea is home to million of plants and animals

English → French [commercial MT system]

Le mer est a la maison de billion des usines et des animaux

French → English

Issues in Semantics

- How to learn the meaning of words?
- From dictionaries:

  plant, works, industrial plant -- (buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles")

  plant, flora, plant life -- (a living organism lacking the power of locomotion)

They are producing about 1,000 automobiles in the new plant

The sea flora consists of 1,000 different plant species

The plant was close to the farm of animals.

Issues in Semantics

- Learn from annotated examples:
  - Assume 100 examples containing “plant” previously tagged by a human
  - Train a learning algorithm
  - How to choose the learning algorithm?
  - How to obtain the 100 tagged examples?

Issues in Information Extraction

- “There was a group of about 8-9 people close to the entrance on Highway 75”
- Who? “8-9 people”
- Where? “highway 75”

- Extract information
- Detect new patterns:
  - Detect hacking / hidden information / etc.
- Gov./mil. puts lots of money into IE research

Issues in Information Retrieval

- General model:
  - A huge collection of texts
  - A query
- Task: find documents that are relevant to the given query
- How? Create an index, like the index in a book
- More …
  - Vector-space models
  - Boolean models
- Examples: Google, Bing, etc.

Issues in Information Retrieval

- Retrieve specific information
- Question Answering
- “What is the height of mount Everest?”
- 11,000 feet

Issues in Information Retrieval

- Find information across languages!
- Cross Language Information Retrieval
- “What is the minimum age requirement for car rental in Italy?”
- Search also Italian texts for “eta minima per noleggio macchine”
- Integrate large number of languages
- Integrate into performant IR engines

Issues in Machine Translations

- Text to Text Machine Translations
- Speech to Speech Machine Translations

- Most of the work has addressed pairs of widely spread languages like English-French, English-Chinese

Issues in Machine Translations

- How to translate text?
  - Learn from previously translated data
- → Need parallel corpora
- French-English, Chinese-English have the Hansards
- Reasonable translations
- Chinese-Hindi – no such tools available today!

This course: Methods

- Fundamentals: tokenization, stemming, stopwords, lemmatization, part-of-speech tagging, syntactic parsing.
- Word association measures
- N-gram language modeling
- Naïve Bayes
- TF.IDF
- Neural networks and deep learning: Convolutional neural networks (CNNs), Recurrent Neural Networks (RNNs), Autoencoders, Sequence-to-Sequence models

This course: Applications

- Opinion mining
- Text classification
- Information Extraction
- Machine Translation
- Dialogue systems
- Paraphrase detection
- Text Summarization
- Text Style transfer
- And more…

Week 1	Introduction to natural language processing (NLP). Linguistics Essentials.
Week 2	Foundations of text processing: tokenization, stemming, stopwords, lemmatization, part-of-speech tagging, syntactic parsing.
Week 3	Word association measures. Distributional word similarity.
Week 4	Probabilistic language modelling. N-grams, perplexity, smoothing. Text classification. Naïve Bayes. Evaluation of text classification systems.
Week 5	Introduction to Neural Networks (NNs) for NLP. Feed-forward NNs, activation functions, cross-entropy loss, Mean Squared Error (MSE) loss, word embeddings.
Week 6	Recurrent neural networks (RNNs) for text classification. Attention mechanisms.
Week 7	Convolutional neural networks (CNNs). Sentence classification with CNNs.
Week 8	Autoencoders and their NLP applications.
Week 9	Sequence-to-sequence models. Machine translation. Dialogue systems.
Week 10	Adversarial and multi-task learning for NLP.
Week 11	Other current research topics in NLP (TBD)
Week 12	Other current research topics in NLP (TBD)

Skills you’ll need for this course

- Linear algebra (vectors, matrices)
- Basic probability theory
- Python proficiency

Credits

- Some slides have been adapted from:

https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html