Machine Learning for Text

Natural Language Processing

Discover The Power of Words!

BIRMINGHAM CITY
University

Amna Dridi, Ph.D

Amna.Dridi@bcu.ac.uk

Agenda

- Why We should care about it?
- What is NLP?
- NLP Tasks
- How to represent Text for ML?
- Classic vs Modern NLP Models
- **Demo**
- NLP Challenges
- NLP Domains
- Summary

Quiz

Why We Should Care About it?

Big Data (Text) Statistics

- In **2019,** internet users spent **1.2 billion** years online.
- **Google** gets over **3.5 billion** searches daily.
- **WhatsApp** users exchange up to **65 billion** messages daily.
- **Facebook** stores and processes more than **30 Petabytes** of user generated data.
- **Twitter** users send over **half a million** tweets **every minute**.
- **95% of businesses** cite the need to manage unstructured data.
- Using big data analytics, **Netflix** saves **$1 billion** per year on customer retention.
- Job listings for **data science** and analytics reached around **2.7 million** in **2020.**

Machine Learning for Text

↓

Natural Language Processing (NLP)

What is NLP?

- A sub-field of **Artificial Intelligence**.

- An **inter disciplinary** subject.

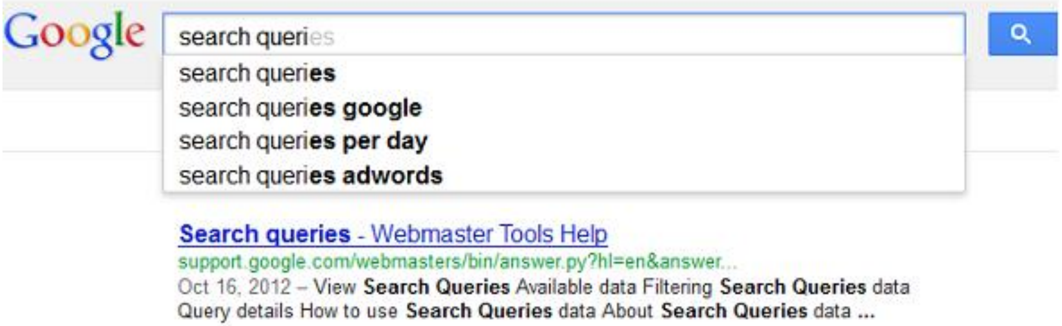
- **Aim:** To build intelligent systems that can **interact with human being like human being!**

- **Natural language**: refers to the **language spoken by people**, e.g. English, Japanese, Urdu, as opposed to artificial languages, like Java, Python, etc.

- **History:** 1950 -- **Alan Turing** published an article called "Machine and Intelligence."
  Started with **Machine Translation** Research

The
Alan Turing
Institute

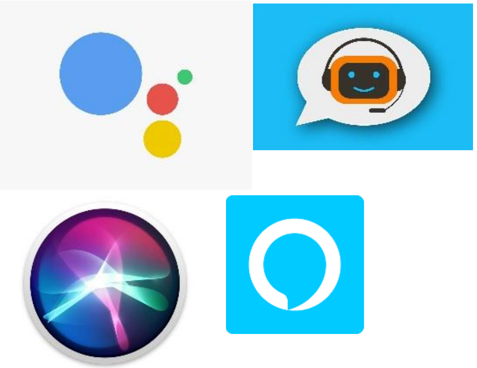
Have you ever used NLP products?

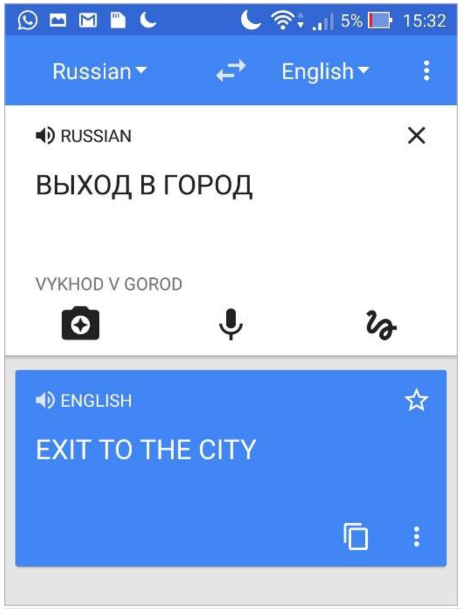


search queries

search queries
search queries **google**
search queries **per day**
search queries **adwords**

Search queries - Webmaster Tools Help
support.google.com/webmasters/bin/answer.py?hl=en&answer...
Oct 16, 2012 – View **Search Queries** Available data Filtering **Search Queries** data Query details How to use **Search Queries** data About **Search Queries** data ...





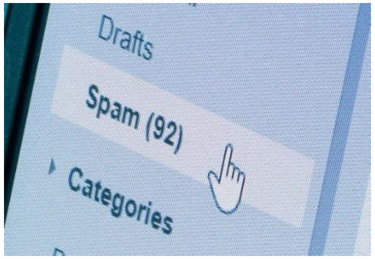
RUSSIAN

ВЫХОД В ГОРОД

VYKHOD V GOROD

ENGLISH

EXIT TO THE CITY



Boris Johnson
@BorisJohnson

I just voted Conservative in the local elections. Make sure you do too! You've got two hours left to get out and vote!

NLP Tasks

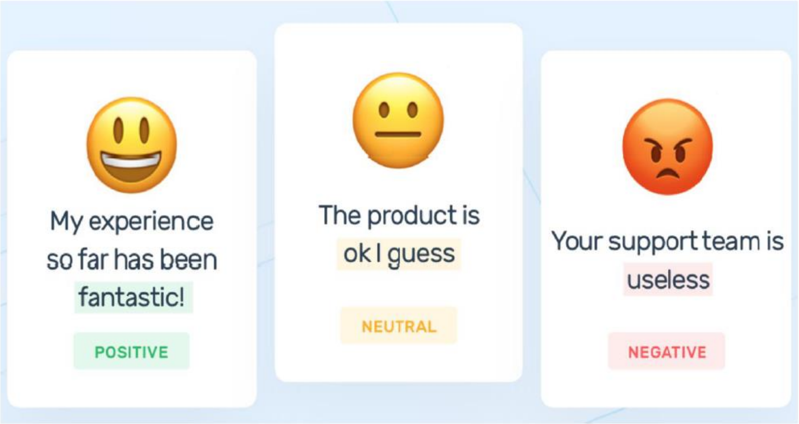
- Sequence Classification
- Sequence Labelling
- Sequence to Sequence

NLP Tasks

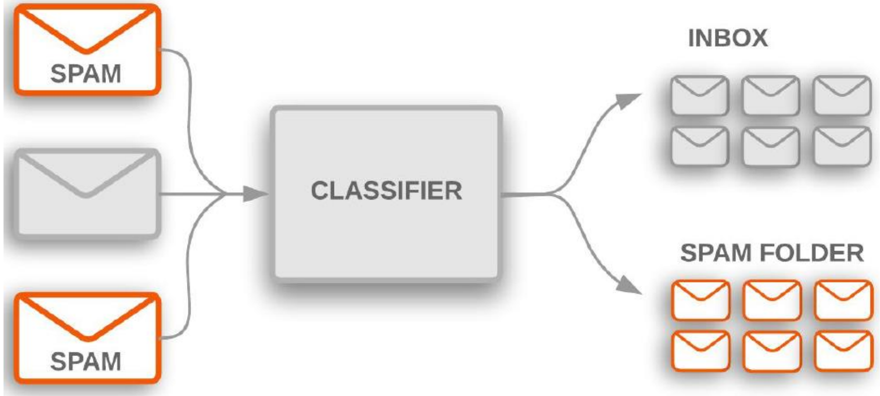
- **Sequence Classification**
- Sequence Labelling
- Sequence to Sequence

Input: Sequence of words

Output: Label/Class


My experience so far has been fantastic!
POSITIVE
The product is ok I guess
NEUTRAL
Your support team is useless
NEGATIVE

Sentiment Analysis



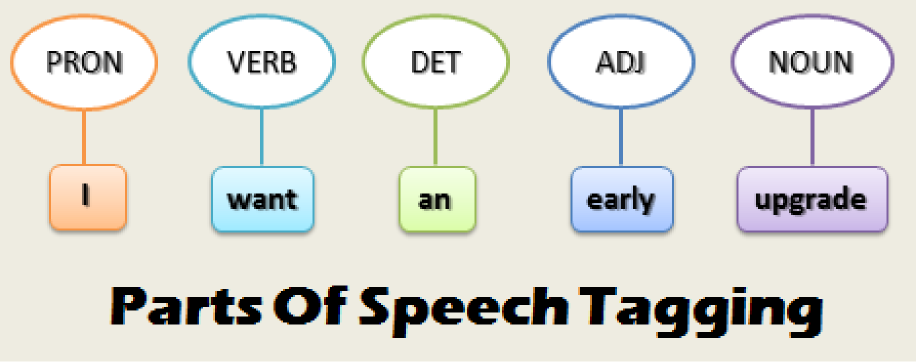
Spam Detector

NLP Tasks

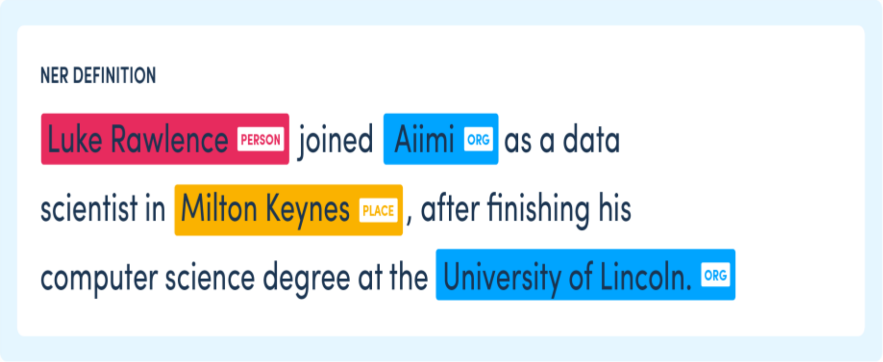
- Sequence Classification
- **Sequence Labelling**
- Sequence to Sequence

Input: Sequence of words

Output: Sequence of Labels/Classes



Part-of-Speech Tagging


NER DEFINITION

Luke Rawlence PERSON joined Aiimi ORG as a data scientist in Milton Keynes PLACE, after finishing his computer science degree at the University of Lincoln. ORG

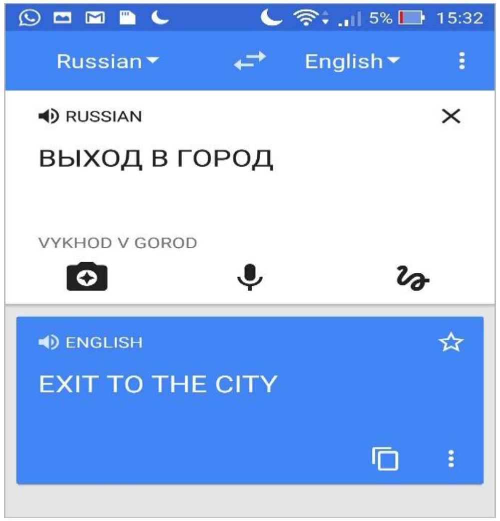
Named Entity Recognition

NLP Tasks

- Sequence Classification
- Sequence Labelling
- **Sequence to Sequence**

Input: Sequence of words

Output: Sequence of Words

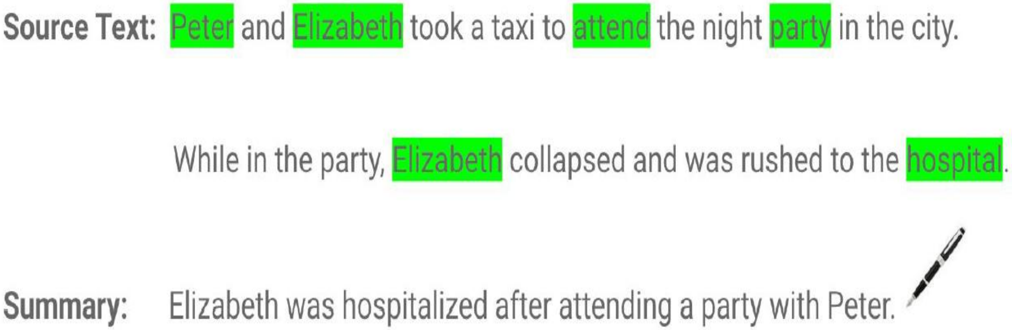


ВЫХОД В ГОРОД

VYKHOD V GOROD

EXIT TO THE CITY

Machine Translation



Source Text: Peter and Elizabeth took a taxi to attend the night party in the city. While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Elizabeth was hospitalized after attending a party with Peter.

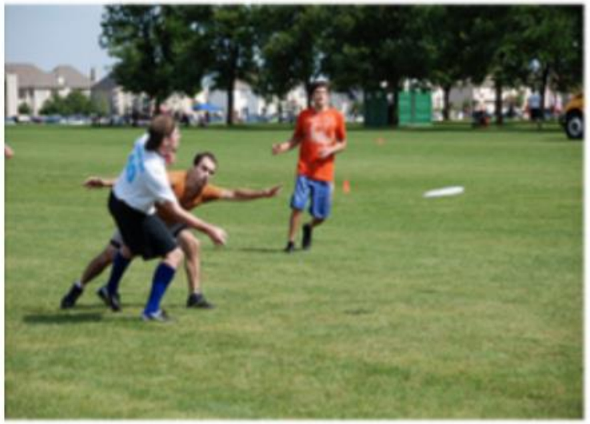
Summarisation

NLP Tasks with Other Media (Multimodal)

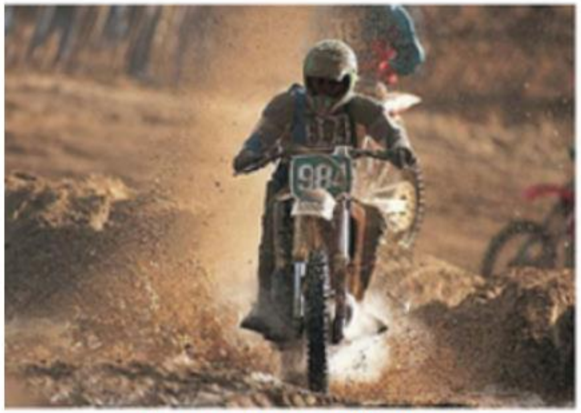
- **Image Captioning**

Input: Images

Output: Sequence of Words



A group of young people playing a game of frisbee.

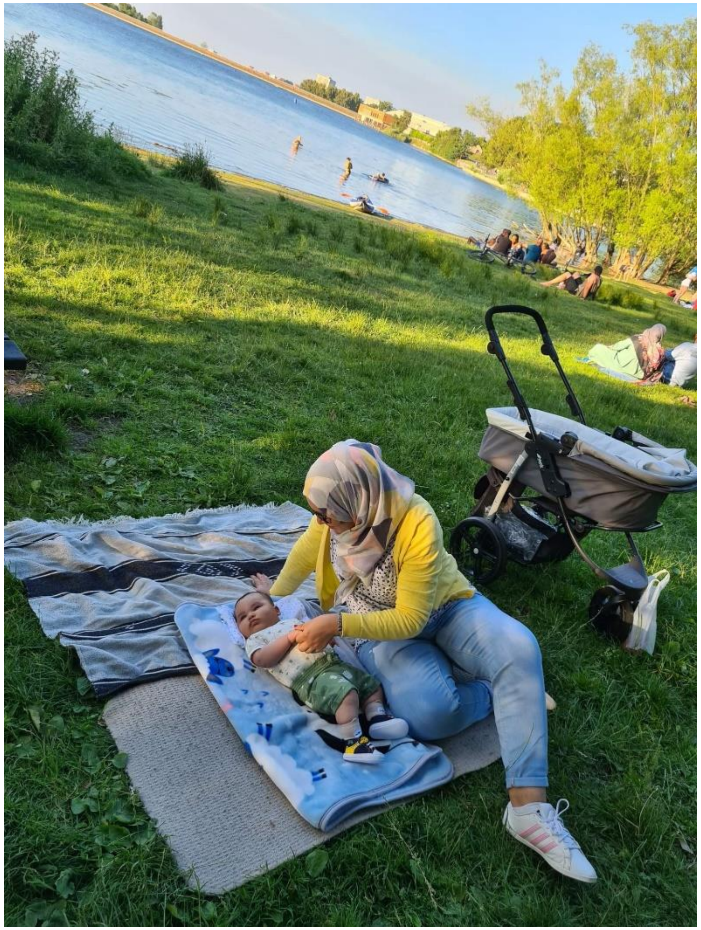


A person riding a motorcycle on a dirt road.

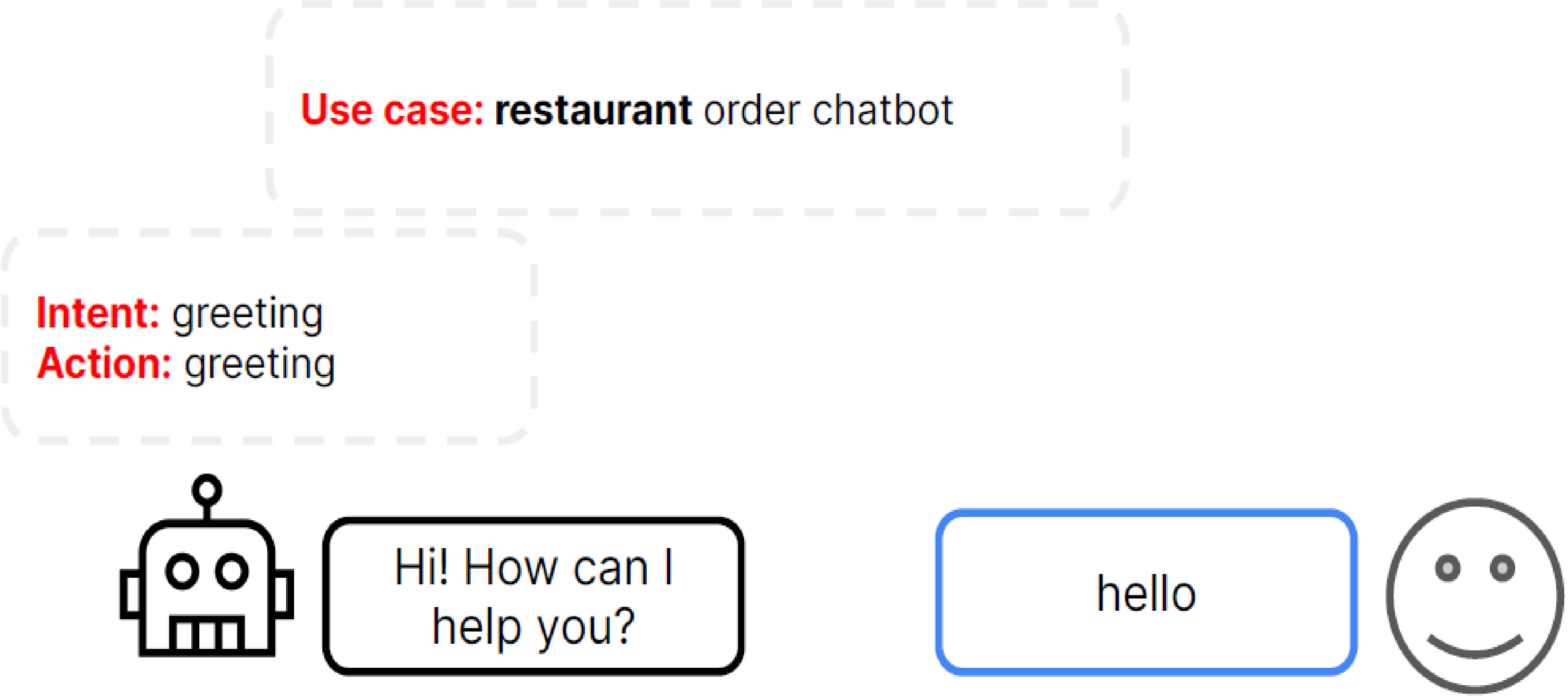
NLP Tasks with Other Media (Multimodal)

- Image Captioning

Amna and her baby boy “Mahdi”
at Edgbaston Reservoir ;)



Example of NLP Production: Restaurant Chatbot



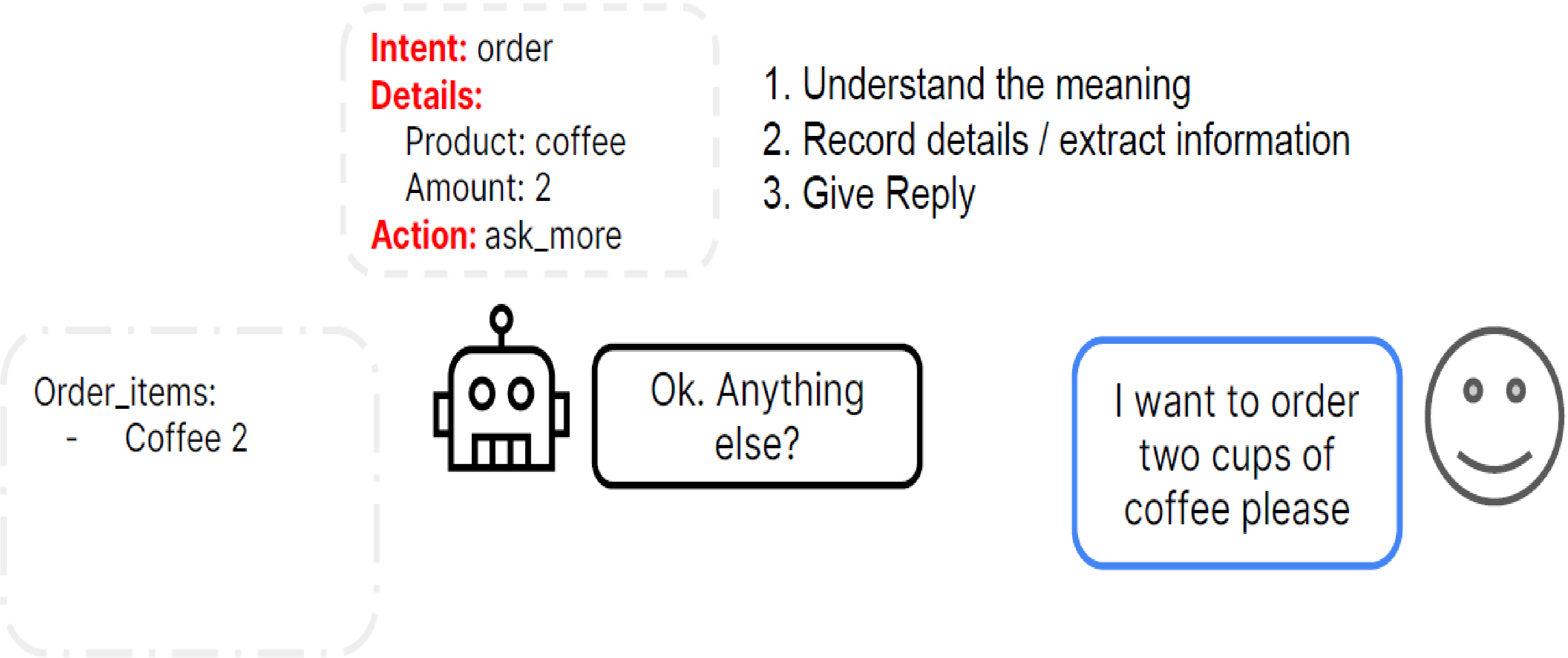
Use case: restaurant order chatbot

Intent: greeting
Action: greeting

Hi! How can I help you?

hello

Example of NLP Production: Restaurant Chatbot



Intent: order
Details:
Product: coffee
Amount: 2
Action: ask_more

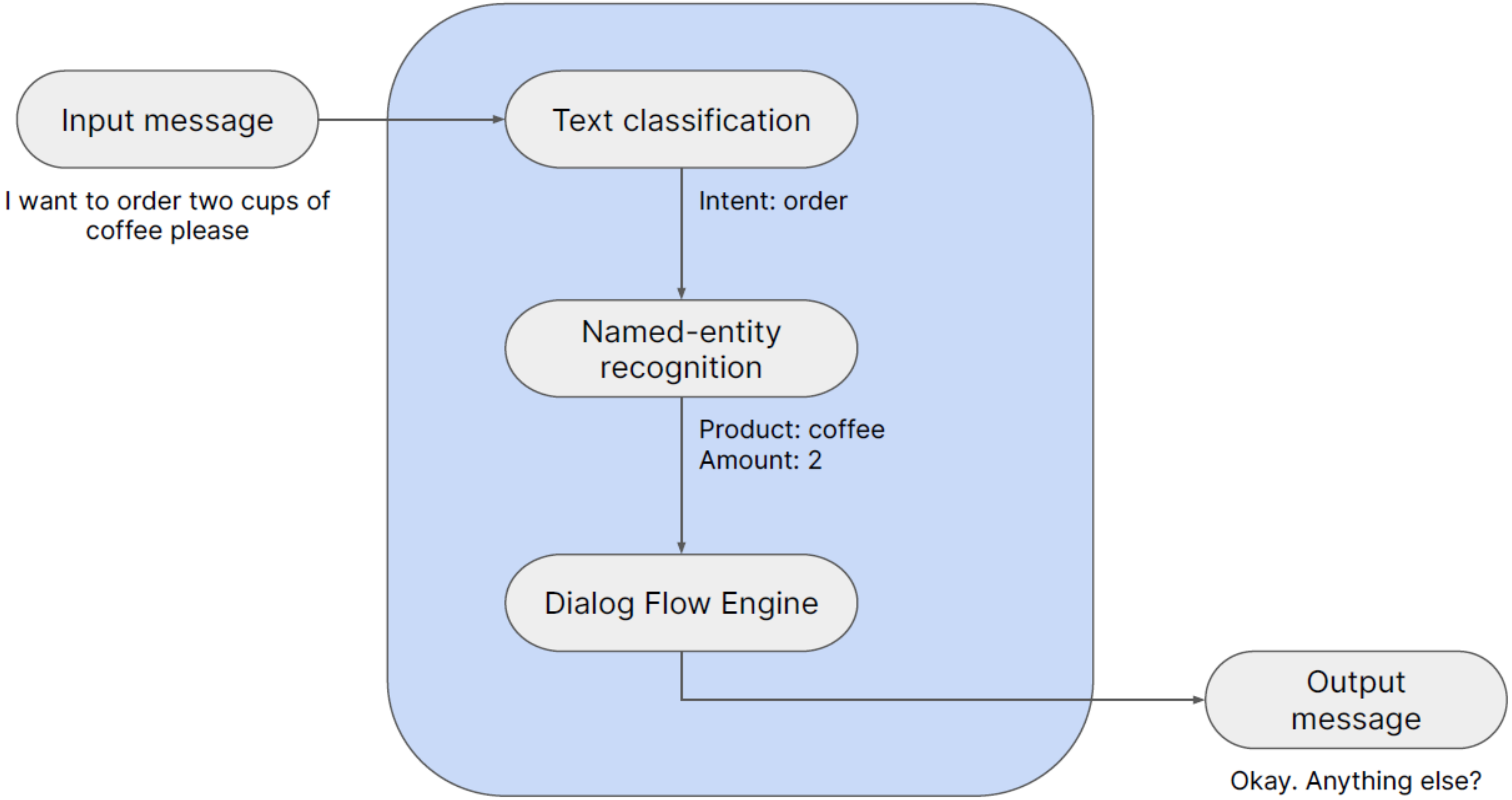
1. Understand the meaning
2. Record details / extract information
3. Give Reply

Order_items:
- Coffee 2

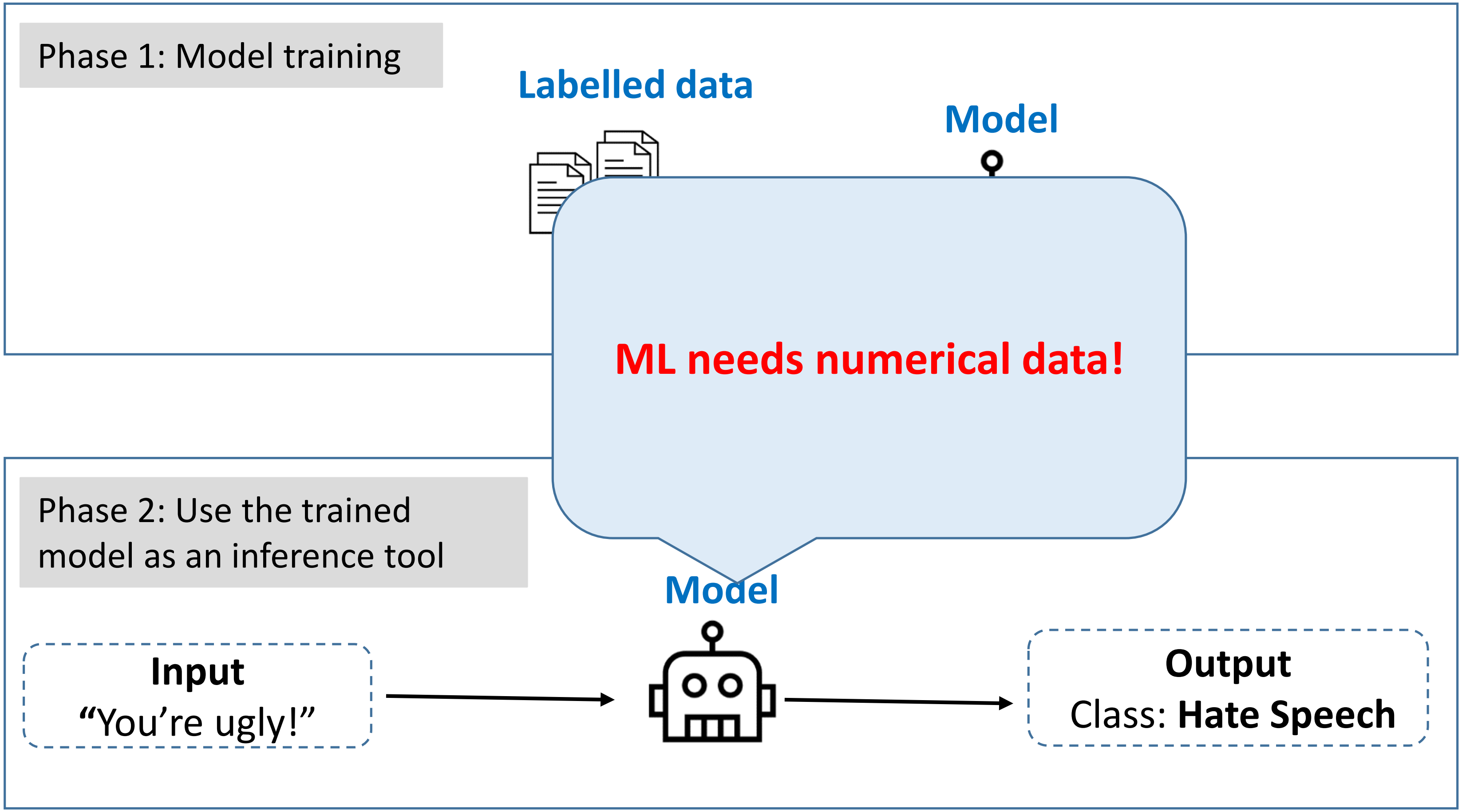
Ok. Anything else?

I want to order two cups of coffee please

Example of NLP Production: Restaurant Chatbot



Inside ML for NLP



Phase 1: Model training

Labelled data

Model

ML needs numerical data!

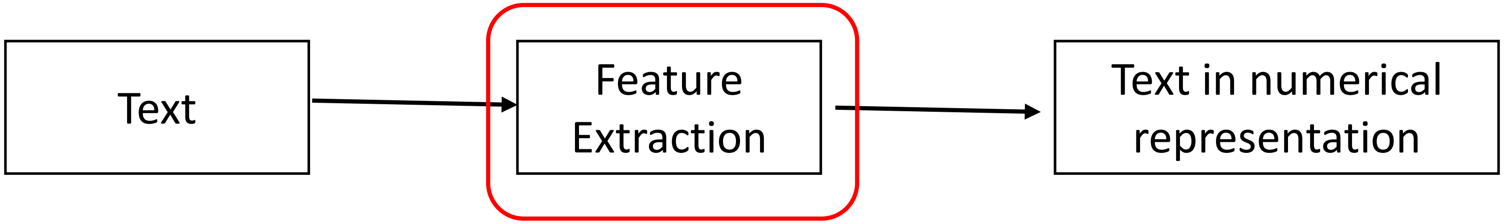
Phase 2: Use the trained model as an inference tool

Model

Input
“You’re ugly!”

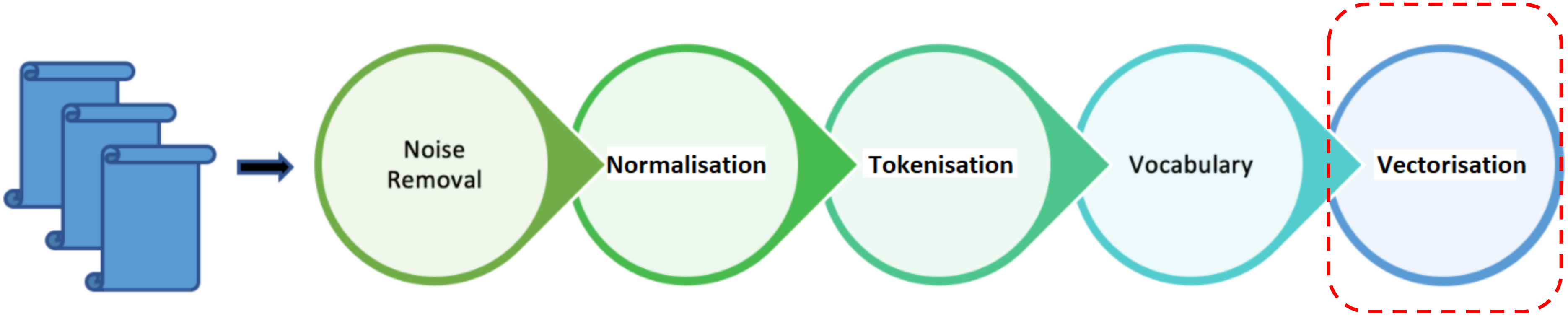
Output
Class: **Hate Speech**

How to represent Text for ML



How to do this?

Inside NLP

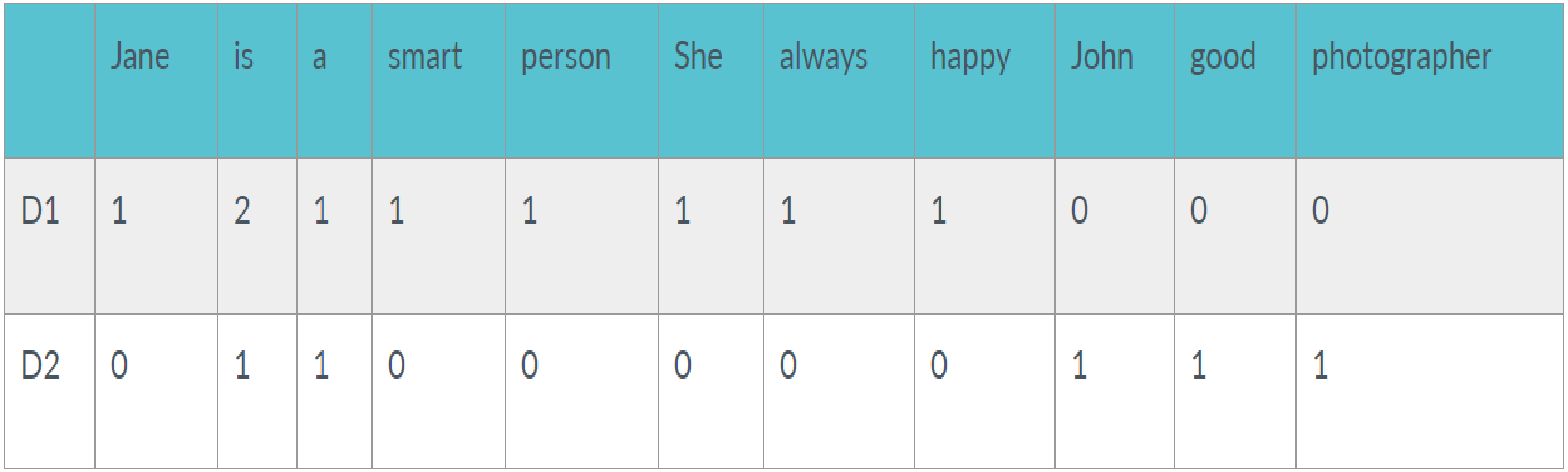


Source: https://www.freshgravity.com/

Text Feature Extraction: Classic NLP Models

1. Bag of Words (also known as Count Vectors)
Each feature represents the frequency of occurrence of each word.

Example
Doc1 (D1) -> Jane is a smart person. She is always happy.
Doc2 (D2) -> John is a good photographer.



	Jane	is	a	smart	person	She	always	happy	John	good	photographer
D1	1	2	1	1	1	1	1	1	0	0	0
D2	0	1	1	0	0	0	0	0	1	1	1

Text Feature Extraction: Classic

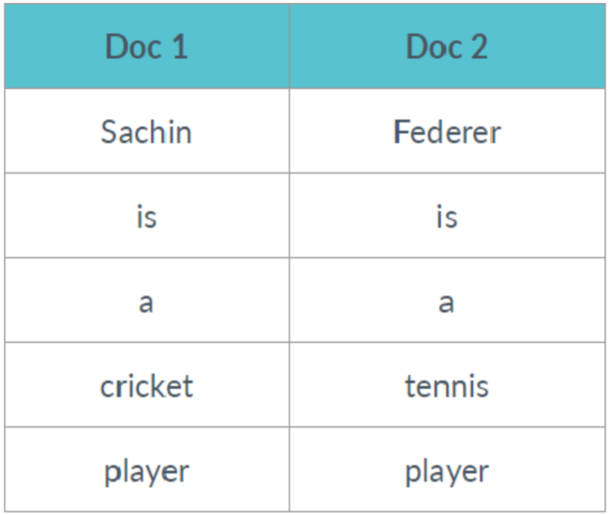
2. **TF-IDF** (Term Frequency-Inverse Document Frequency)

- TF is calculated as (number of times term t appears in the document) / (number of terms in the document). It denotes the contribution of words to the document.
- IDF is calculated as log(N/n), where N is the number of documents and n is the number of documents a term t has appeared in.

Example

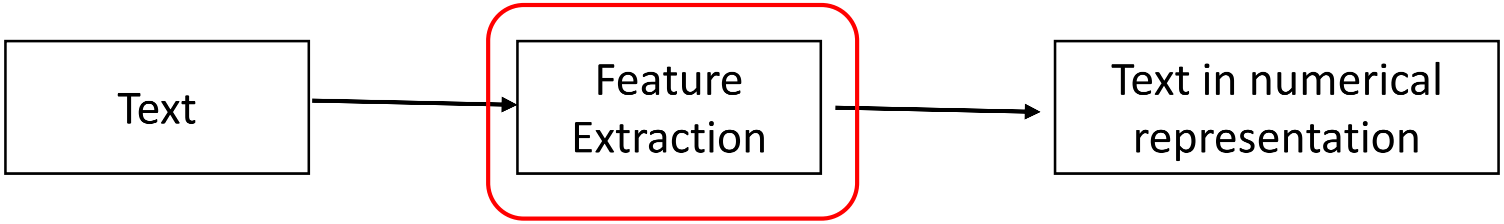
Doc1 (D1) -> Sachin is a cricket player.
Doc2 (D2) -> Federer is a tennis player.

- TF of D1 and D2 = 1/5
- IDF for ‘Sachin’ = log(2/1)= 0.301
- IDF for ‘a’ = log(2/2) = 0
- TF-IDF for ‘Sachin’ in D1 = (1/5) * 0.301 = 0.602
- TF-IDF for ‘a’ in D1 = (1/5) * 0 = 0



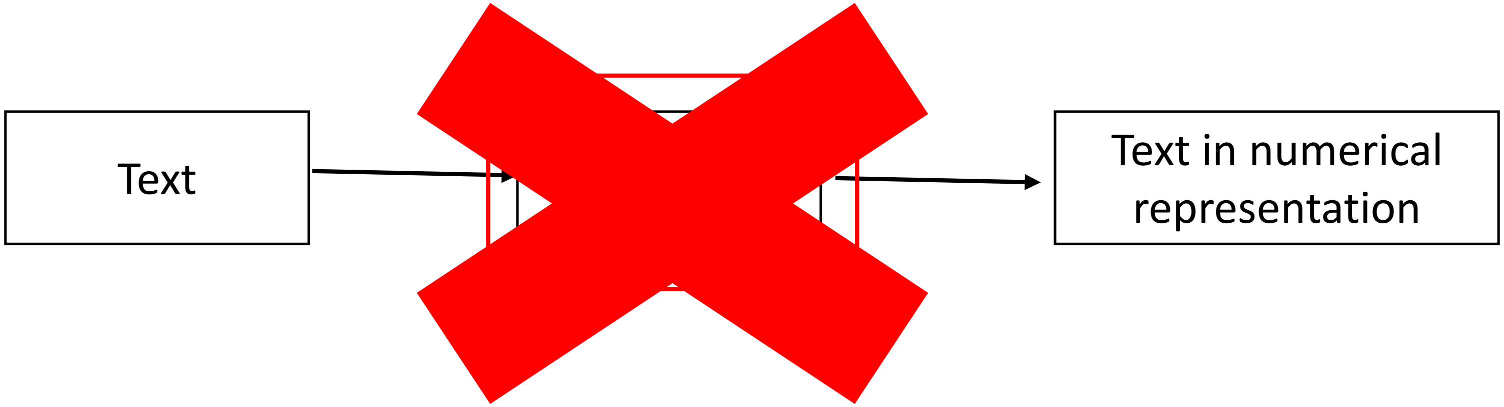
Doc 1	Doc 2
Sachin	Federer
is	is
a	a
cricket	tennis
player	player

How to represent Text for ML

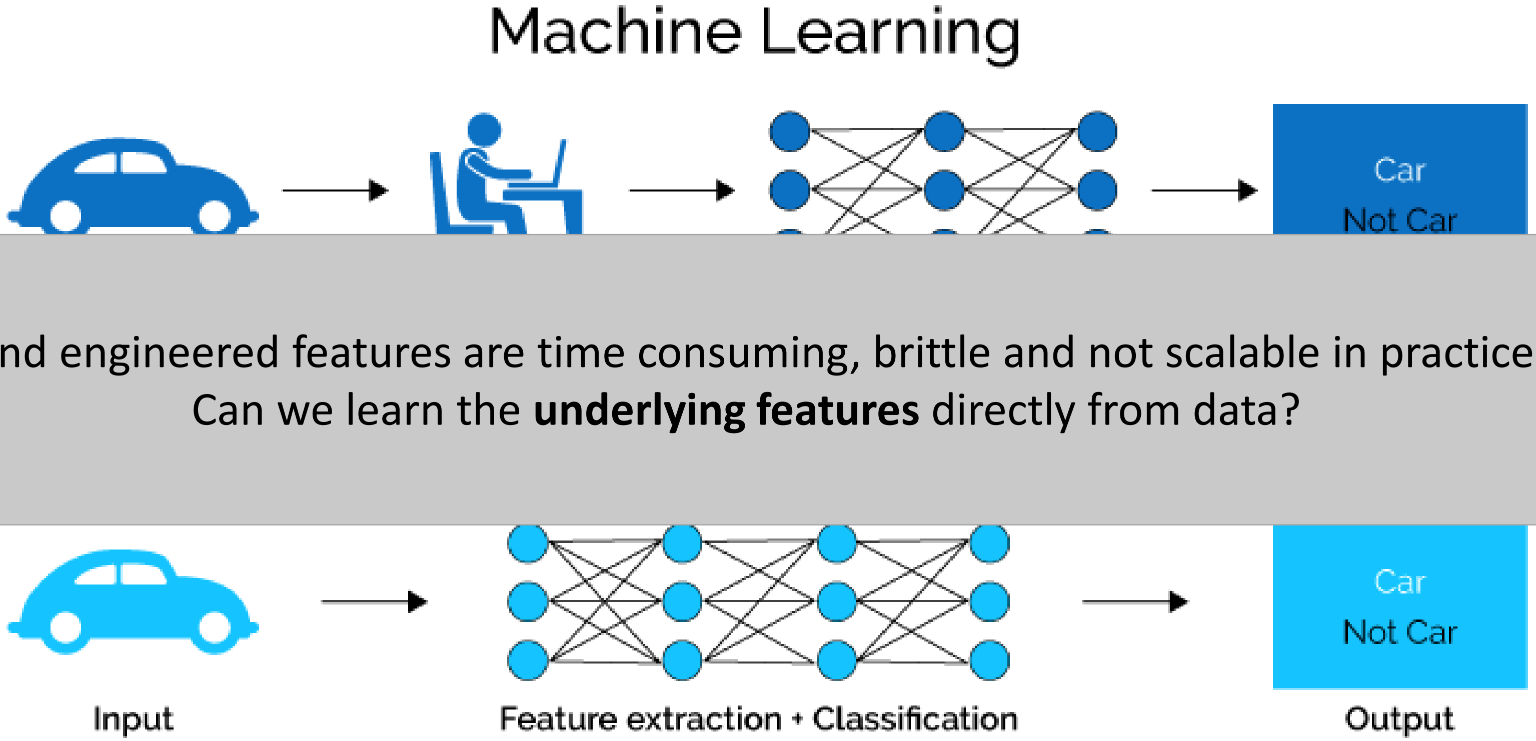


We select how to represent the text!

Text Feature Extraction: Modern



Text Feature Extraction: Modern NLP Models – Towards Deep Learning!

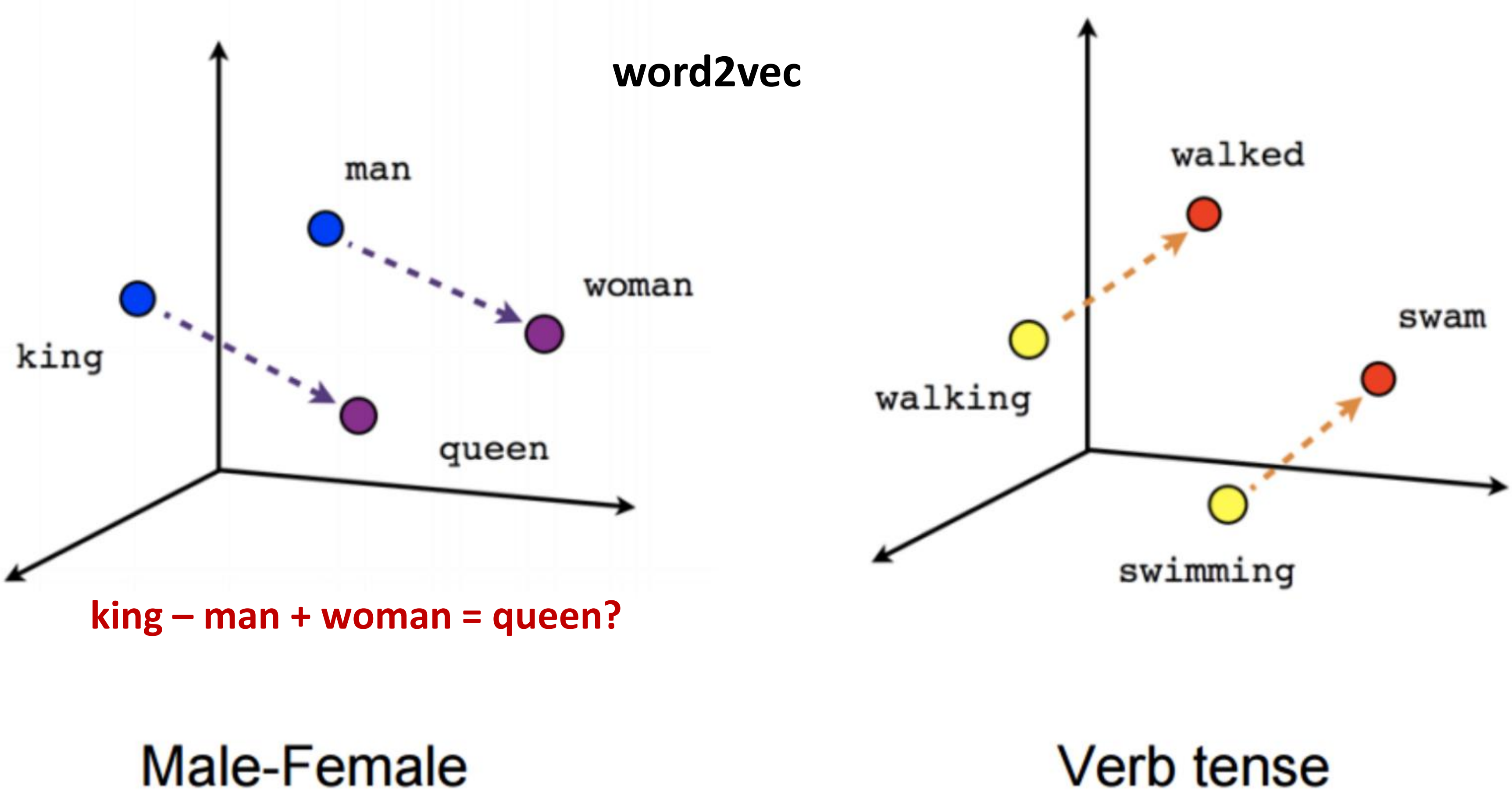


Hand engineered features are time consuming, brittle and not scalable in practice. Can we learn the **underlying features** directly from data?

Source: https://www.xenonstack.com/

Text Feature Extraction: Modern – Word Embedding

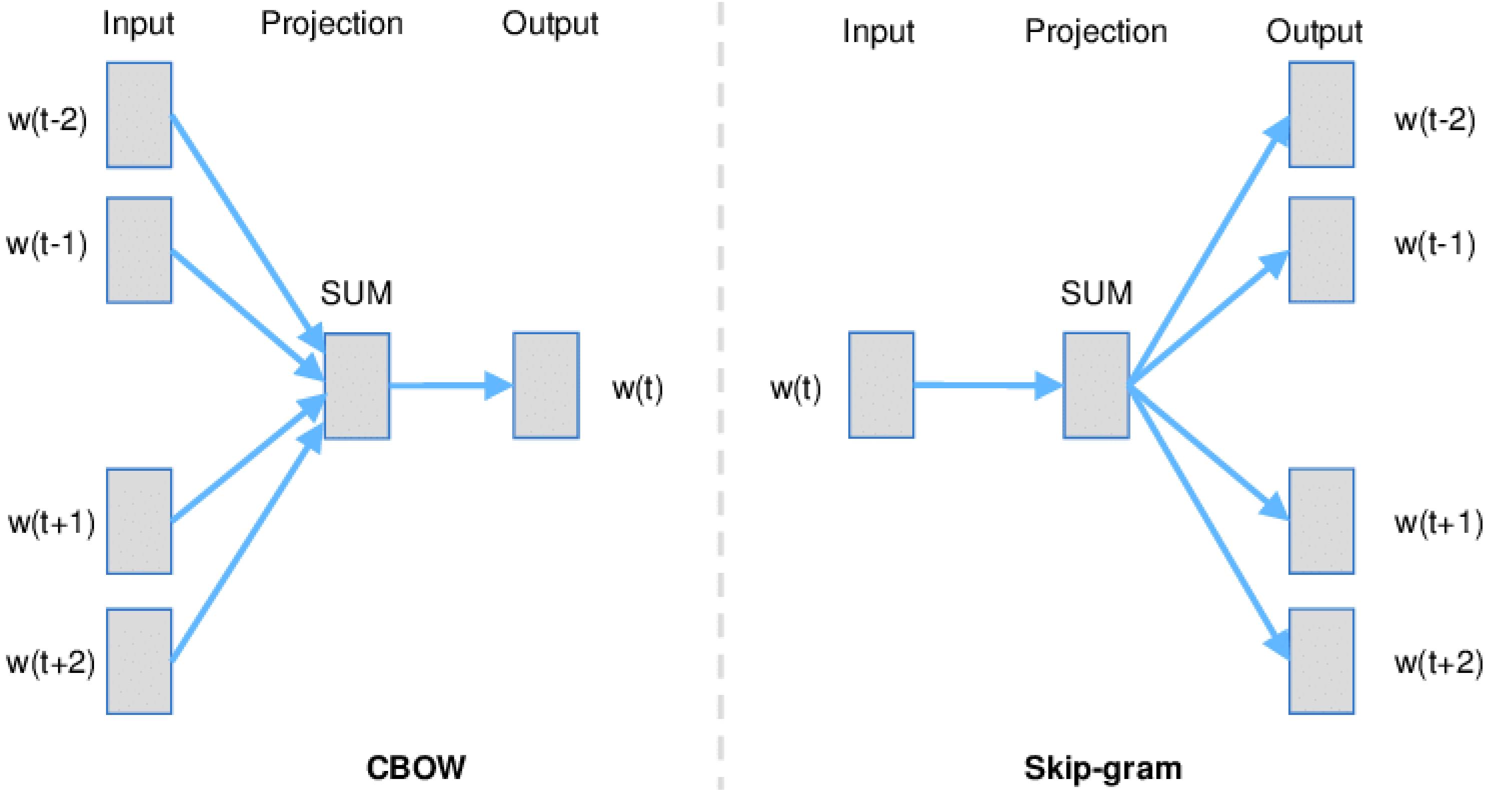
word2vec



king – man + woman = queen?

Text Feature Extraction: Modern – Word Embedding

Word2vec Models



Text Feature Extraction: Modern – Word Embedding

Skip-Gram Model

Example: Covid-19 is short form for coronavirus disease 2019. (window size=2)

Covid-19	is	short	form	for	coronavirus	disease	2019

- Training samples: (form, is), (form, short), (form, for), (form, coronavirus)

- We are going to represent an **input** word like "short" as a **one-hot vector**. This vector will have as many components as in our vocabulary and we will place "1" in the position corresponding to the words "short". Example: "short" will be **00100000**.
- The **hidden layer** is going to be represented by a weight matrix with the dimension (**vocabulary size x number of hidden neurons**). Example- (**8×300**).
- **The output** of the network is a **single vector** containing for every word in our vocabulary, the **probability** that a randomly selected nearby word is that vocabulary word.

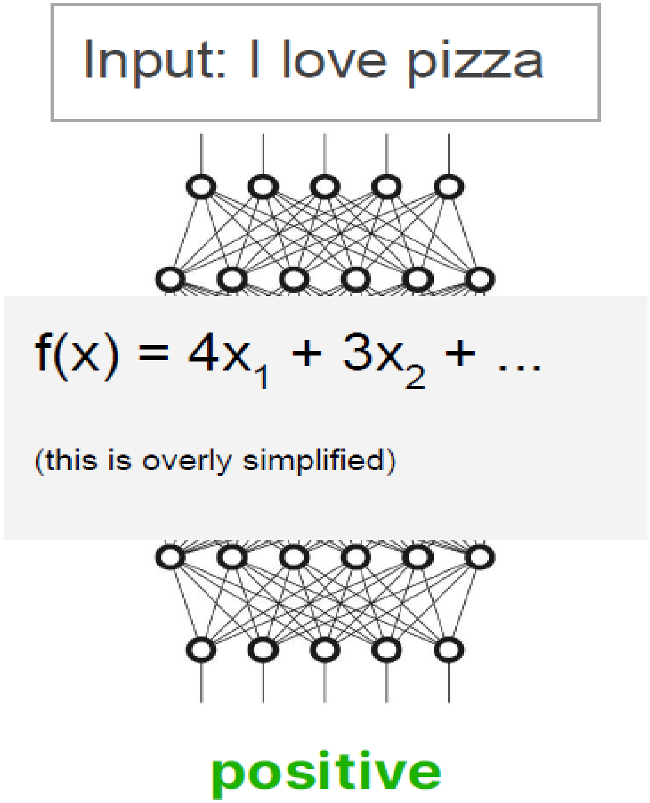
Demo: Word2Vec Training and Visualisation

- **Dataset:** Wiki pages of “**coronavirus**”
- **Programming language:** Python
- **Library:** **Gensim**

Modern NLP: Deep Learning

Deep Neural Network →

(a **complex math**)



Input: I love pizza

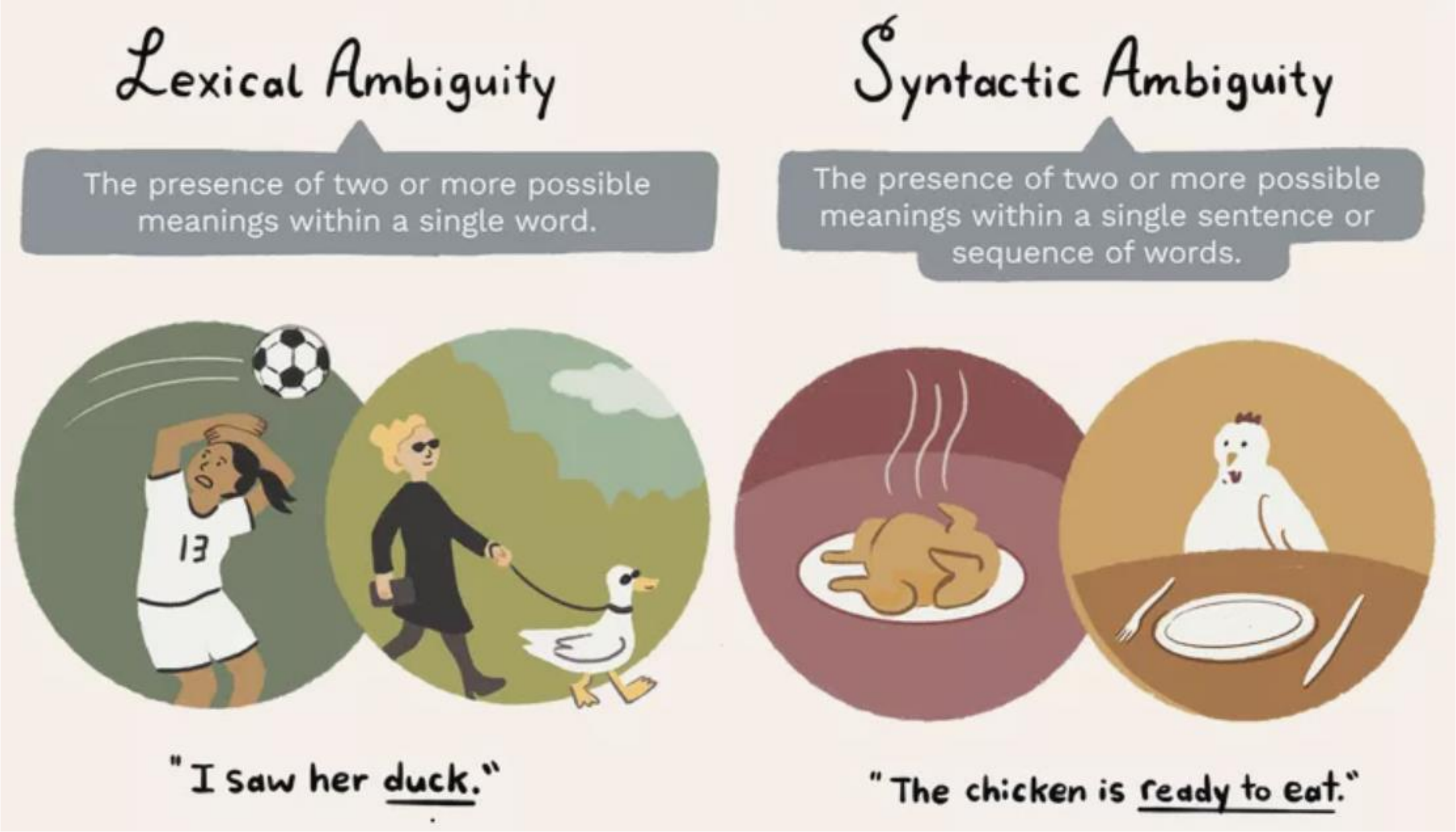
$f(x) = 4x_1 + 3x_2 + \ldots$

(this is overly simplified)

positive

Challenges in NLP

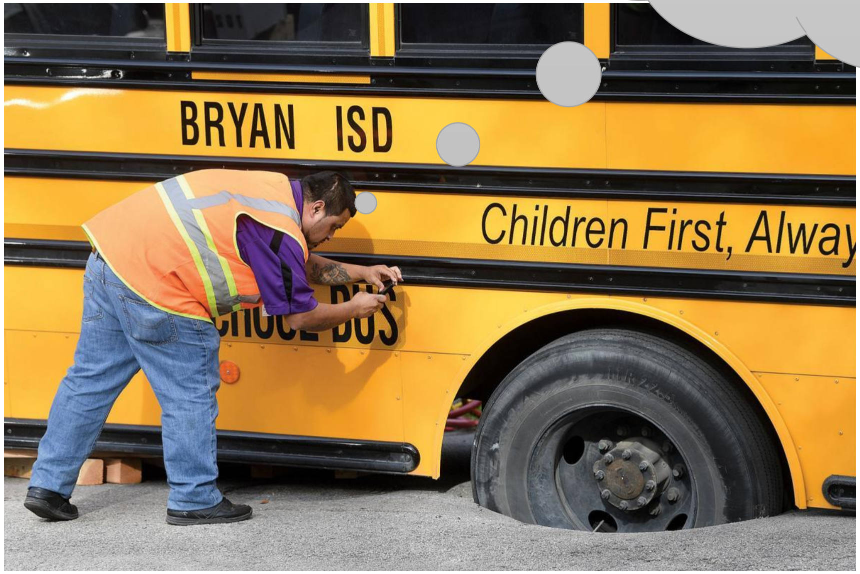
NLP is hard!
Language is ambiguous!


Lexical Ambiguity
The presence of two or more possible meanings within a single word.
13
"I saw her duck."
Syntactic Ambiguity
The presence of two or more possible meanings within a single sentence or sequence of words.
"The chicken is ready to eat."

Source: https://www.thoughtco.com/

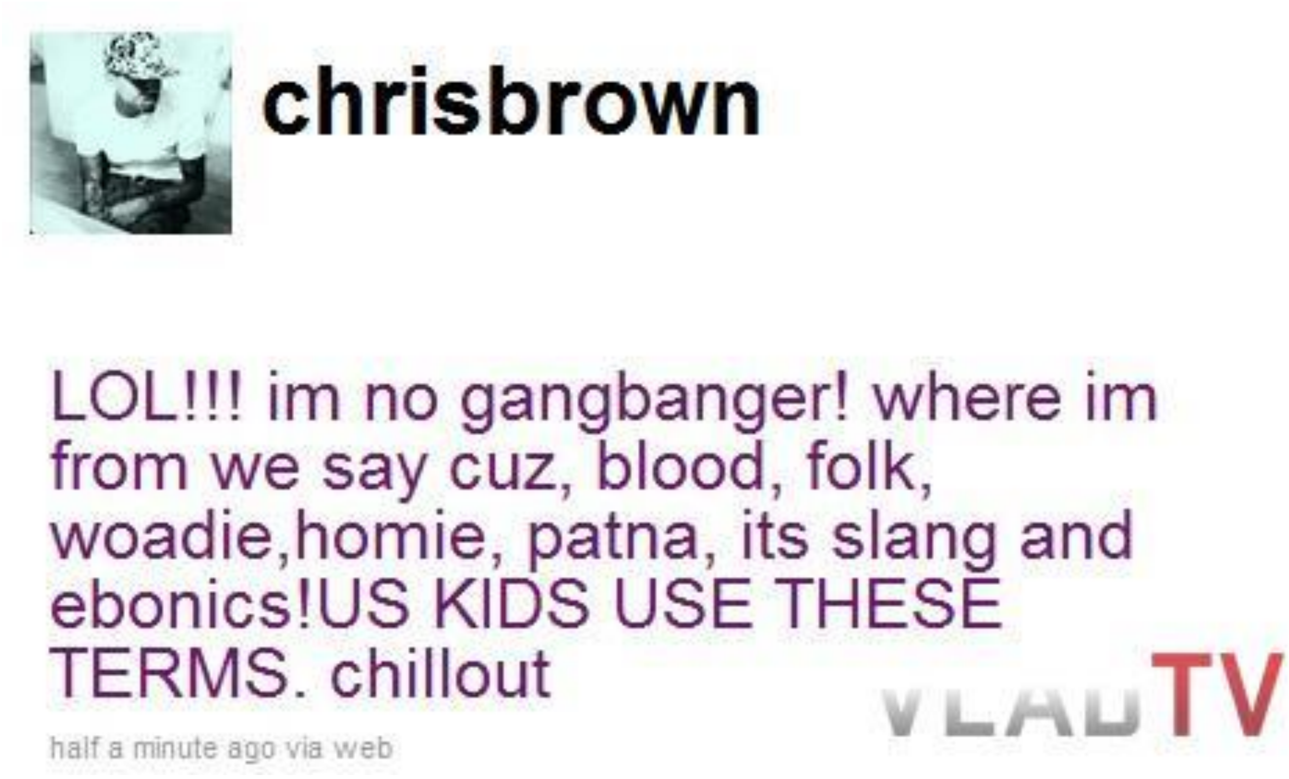
NLP is hard!

Irony!



"Excellent! This day couldn't start off any better!"

NLP is hard!
Slang & Non-standard words!

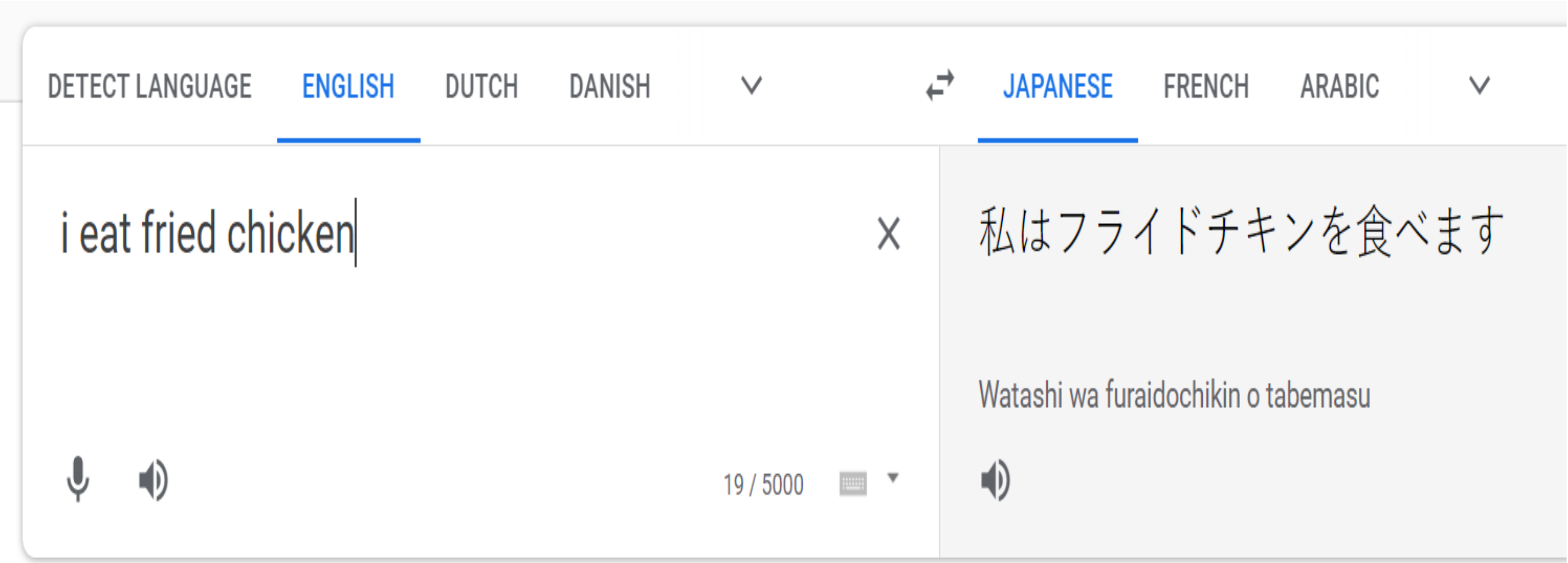


chrisbrown

LOL!!! im no gangbanger! where im from we say cuz, blood, folk, woadie,homie, patna, its slang and ebonics!US KIDS USE THESE TERMS. chillout

half a minute ago via web

NLP is hard!
Text Localisation (E.g. English vs Japanese)



i eat fried chicken

私はフライドチキンを食べます

Watashi wa furaidochikin o tabemasu

NLP in Other Domains

- Bio-medical
- Forensic science
- Advertisement
- Education
- Politics
- E-governance
- Business Development
- … and wherever we use language!

Summary

- **NLP** is a branch of **AI** which helps computers to understand, interpret and manipulate **human language**.

- **NLP** started when **Alan Turing** published an article called "***Machine and Intelligence***".

- The main **NLP tasks** are **sequence classification**, **sequence labelling** and **sequence to sequence**.

- **Classic NLP models** are based on **feature extraction** and **statistical models**.

- **Modern NLP models** are based on **neural networks**.

- Essential **Applications of NLP** are Information retrieval & Web Search, Grammar Correction, Question Answering, , Text Summarization, Machine Translation, etc

NLP Tools

python™
spaCy
scikit learn
TensorFlow
GENSIM
topic modelling for humans
NLTK
PyTorch
HUGGING FACE

… and many more!

NLP Resources Tips

Medium

deeplearning.ai

Stanford | NLP

Thank You :)

Amna Dridi, Ph.D
Amna.Dridi@bcu.ac.uk

Quiz