



Università  
degli Studi  
di Palermo



# Introduzione al Natural Language Processing

CORSO DI NATURAL LANGUAGE PROCESSING (ELABORAZIONE DEL LINGUAGGIO NATURALE)

a.a. 2025/2026

Prof. Roberto Pirrone



# Cos'è il Natural Language Processing

- «L'elaborazione del linguaggio naturale (NLP) è un campo dell'intelligenza artificiale (AI) che consente ai computer di comprendere, interpretare e generare il linguaggio umano. Combina informatica, linguistica e apprendimento automatico per elaborare e analizzare testi e discorsi, consentendo alle macchine di interagire con gli esseri umani in modo naturale. Esempi di NLP includono chatbot, assistenti vocali, software di traduzione e motori di ricerca.»

*Generato in Inglese da Gemini, su richiesta Google,  
e tradotto con DeepL.com (versione gratuita)*

# Cos'è il Natural Language Processing

- «L'elaborazione del linguaggio naturale (NLP) è un campo dell'intelligenza artificiale (AI) che consente ai computer di comprendere, interpretare e generare il linguaggio umano. Combina informatica, linguistica e apprendimento automatico (**Machine Learning – ML**) per elaborare e analizzare testi e discorsi, consentendo alle macchine di interagire con gli esseri umani in modo naturale. Esempi di NLP includono chatbot, assistenti vocali, software di traduzione e motori di ricerca.»

*Generato in Inglese da Gemini, su richiesta Google,  
e tradotto con DeepL.com (versione gratuita).*

*Infine rivisto da me*

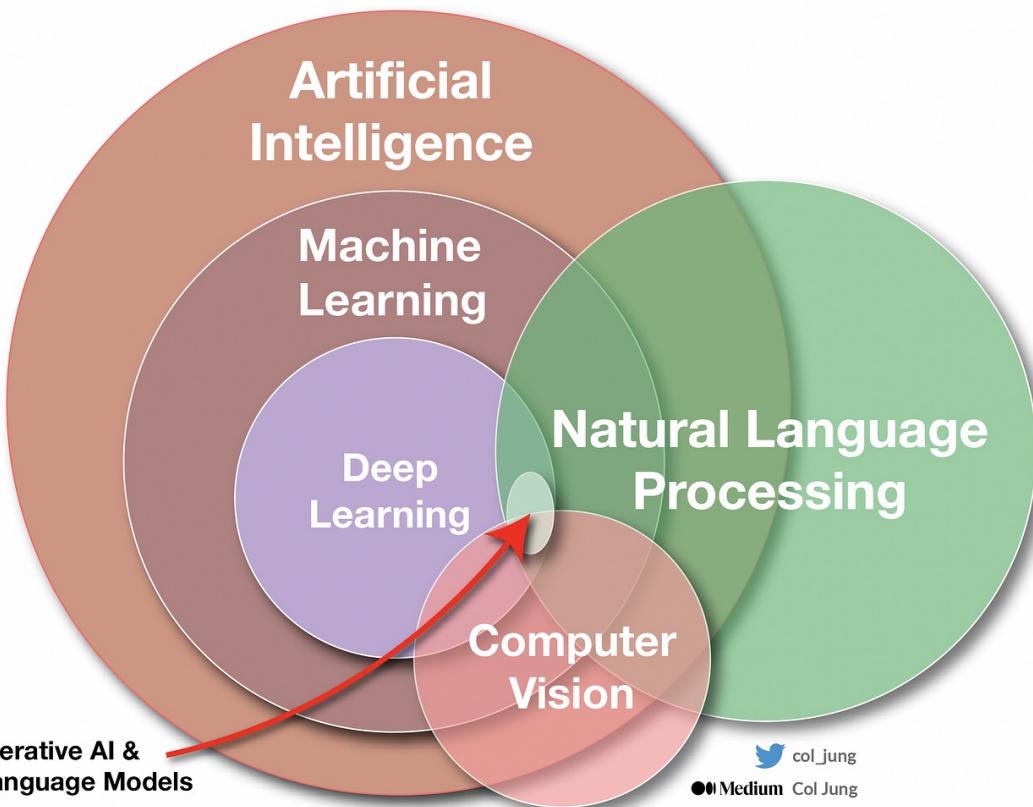
# Cos'è il Natural Language Processing

- L'insieme delle tecnologie di AI per consentire ad un robot o a un «agente virtuale» di comprendere il linguaggio umano per:
  - Rispondere a delle domande
  - Riconoscere la struttura di un testo
  - Riconoscere i tratti emotivi e di genere in un testo
  - Tradurre le diverse lingue e interagire in maniera fluente con gli utenti

# Intelligenza Artificiale predittiva e generativa



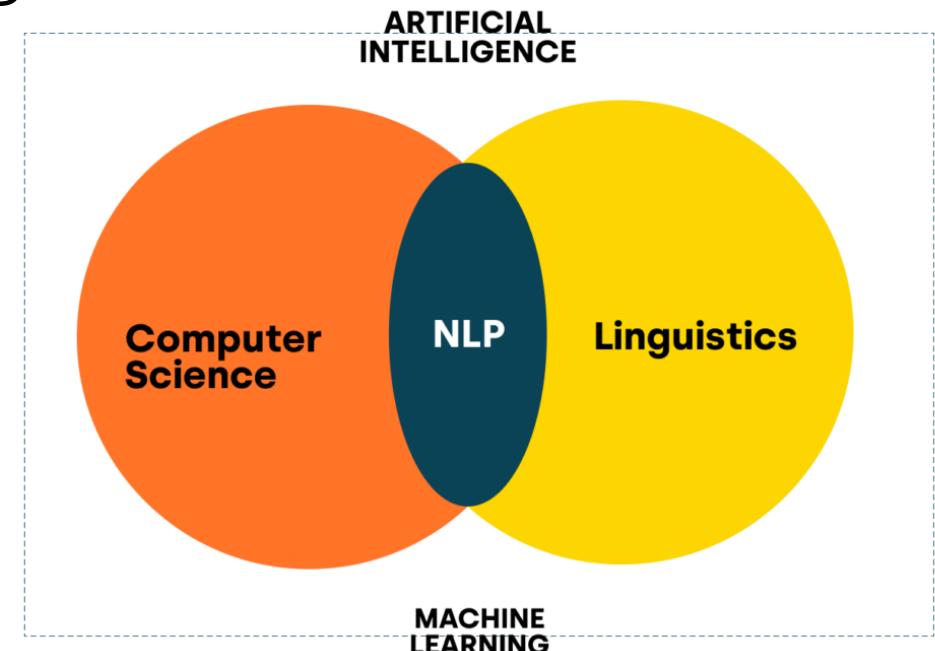
# Intelligenza Artificiale predittiva e generativa



# E dall'altra parte del diagramma cosa c'è?

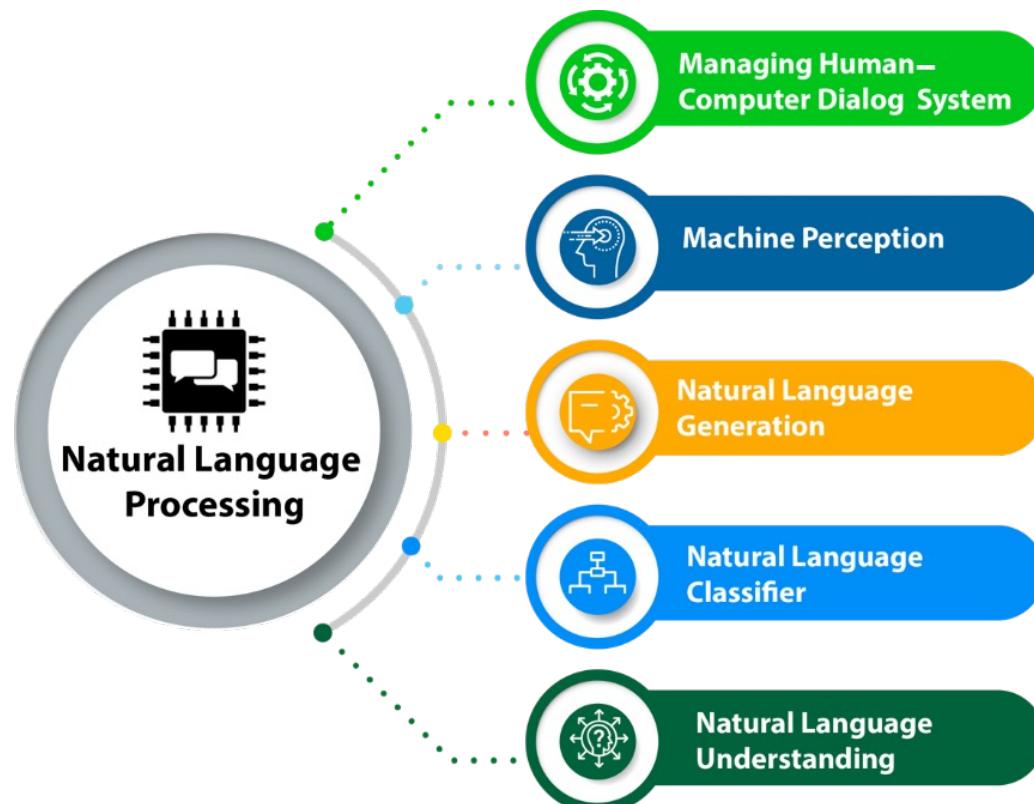
**Linguistica:** Scienza che studia sistematicamente il linguaggio umano nella totalità delle sue manifestazioni, e quindi le lingue come istituti storici e sociali, la loro ripartizione, i loro reciproci rapporti, nonché la funzionalità delle singole lingue sotto differenti aspetti (fonetico, sintattico, lessicale, semantico), sia nella struttura con cui si presentano in un determinato momento della loro storia sia nella loro evoluzione attraverso il tempo.

(vocab. Treccani)



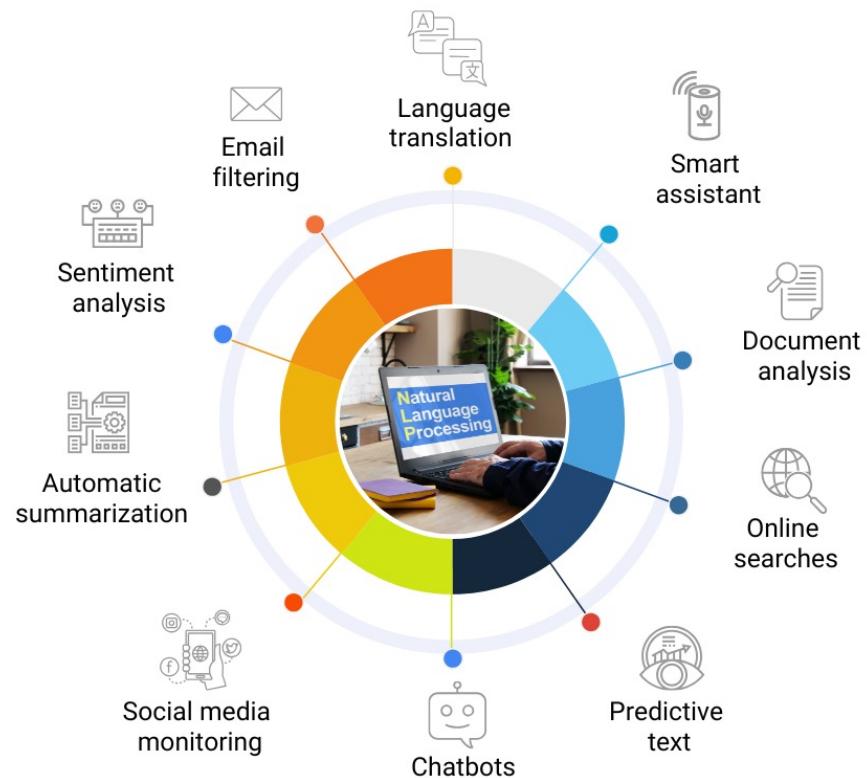
<https://bit.ly/3IEQ7Y8>

# Applicazioni del Natural Language Processing

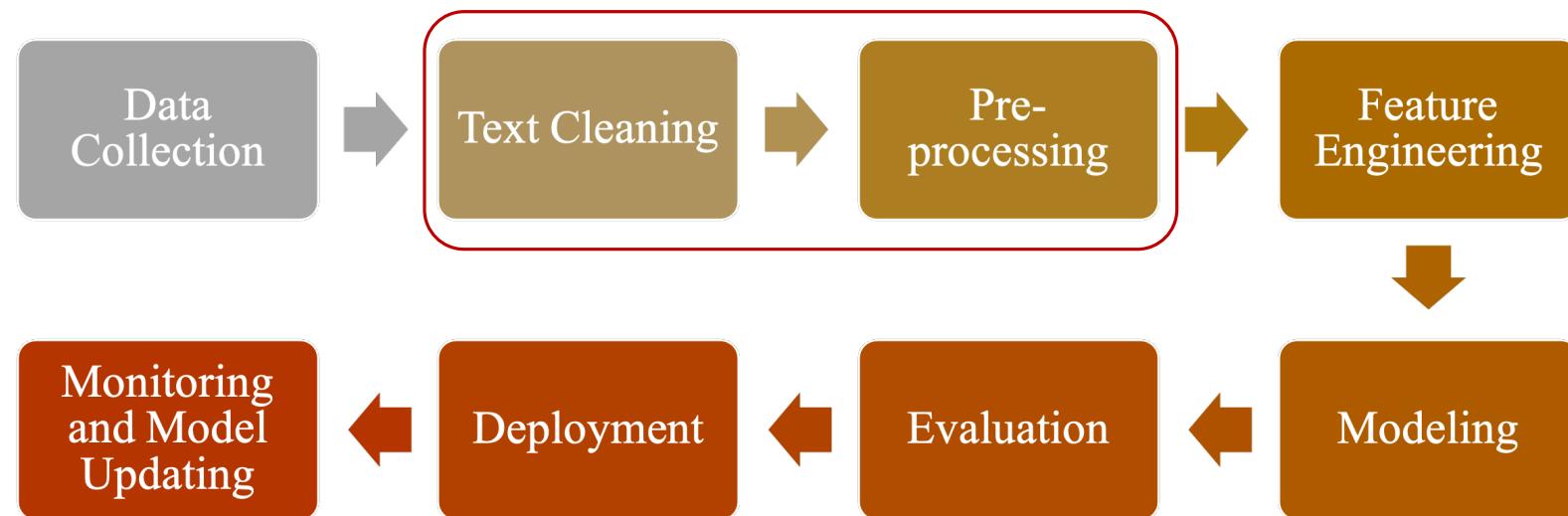


<https://bit.ly/4q7O7bu>

# Applicazioni del Natural Language Processing

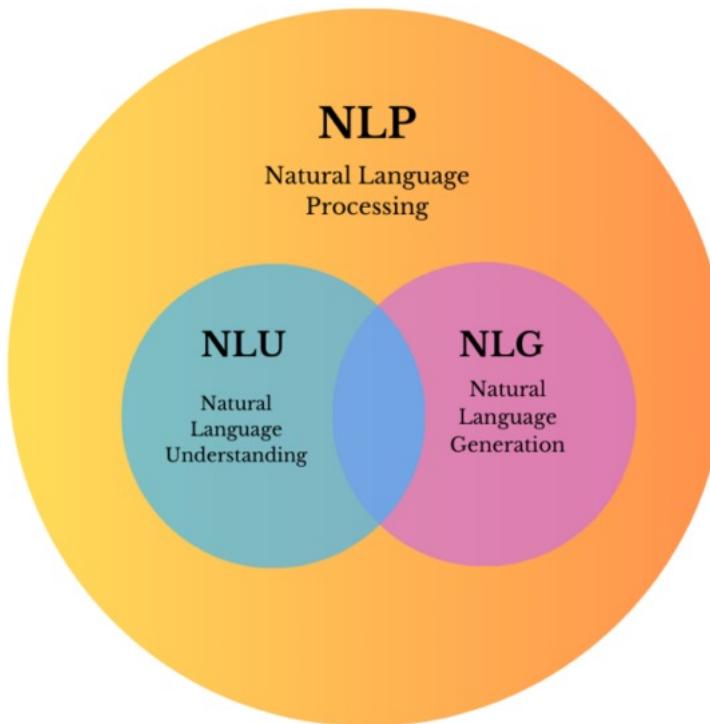


# Flusso di lavoro



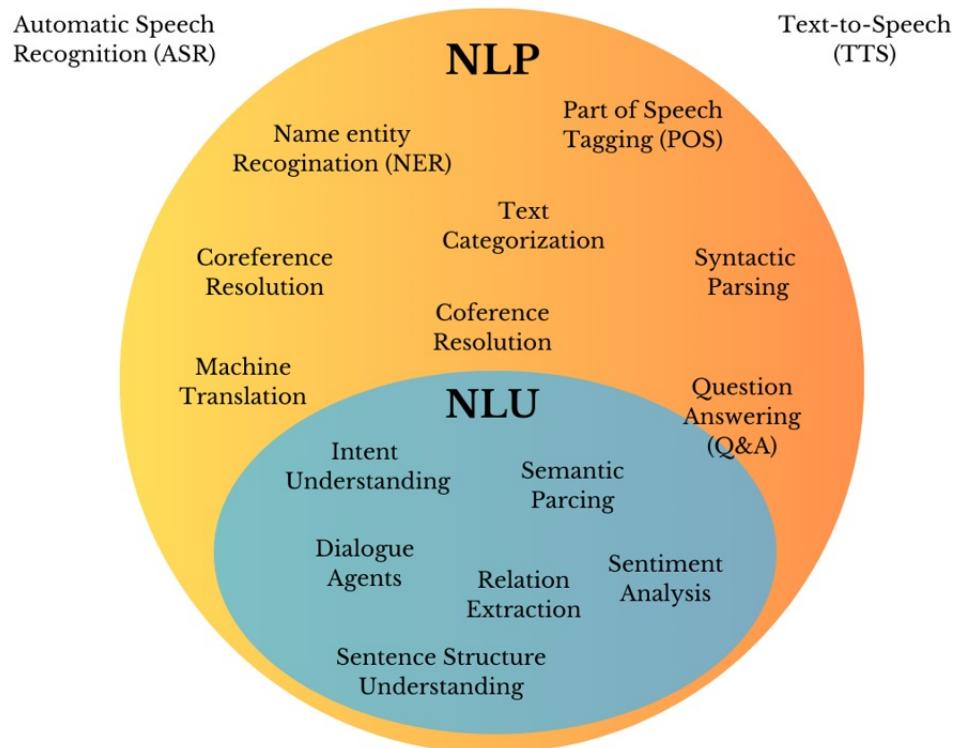
<https://bit.ly/4q6WC6L>

# NLP, NLU, NLG



- **NLP** Focuses on enabling computers to understand, interpret, and generate human language
- **NLG** Involves generating coherent and contextually relevant text
- **NLU** Enables machines to comprehend and interpret human language, extracting meaning, intent, and context.

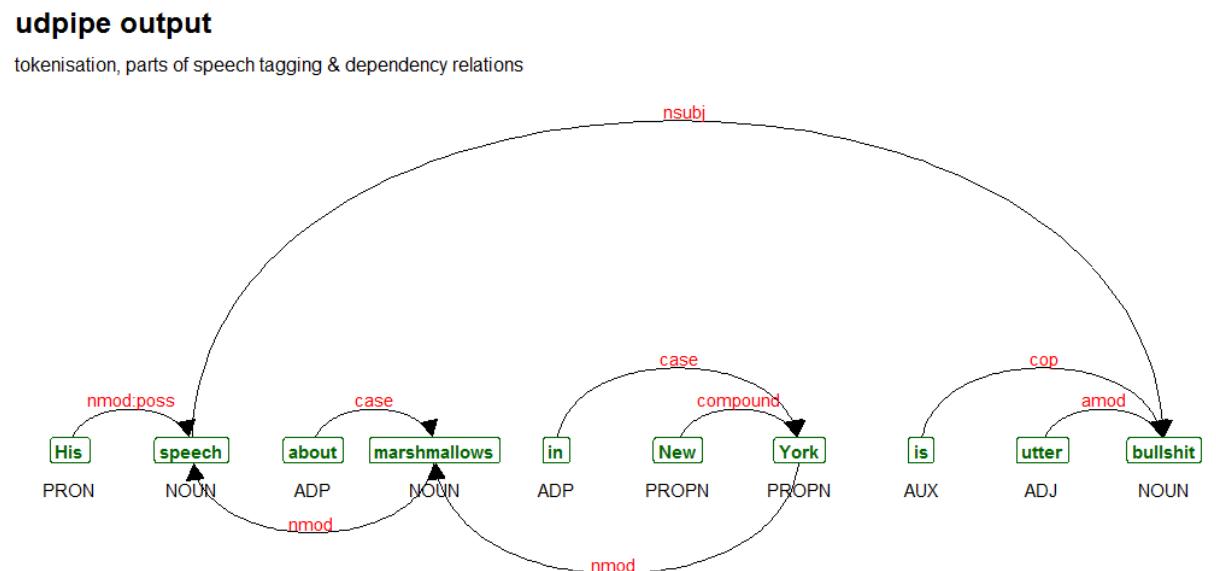
# NLP, NLU, NLG



<https://bit.ly/3IXDfwg>

# Tokenizzazione, POS Tagging e Dependency Parsing

- Consentono ad un robot o a un «agente virtuale» di elaborare il linguaggio umano al fine di *analizzare la struttura delle frasi*



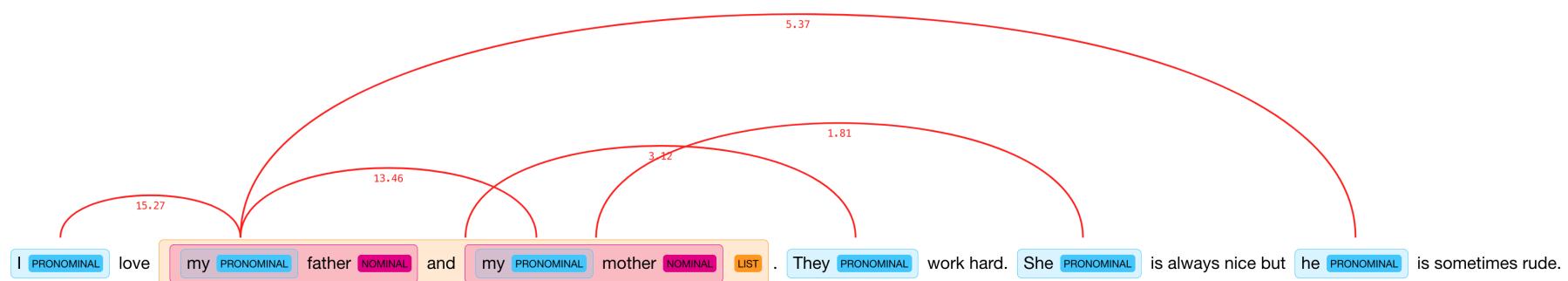
# Tokenizzazione, POS Tagging e Dependency Parsing

<https://lindat.mff.cuni.cz/services/udpipe/>

The screenshot shows a web browser window with the URL <https://lindat.mff.cuni.cz/services/udpipe/>. The page is titled "UDPipe". It features a navigation bar with links to Catalog, Repository, Education, Projects, Tools, Services, and About. Below the navigation bar, there are logos for DARIAH-EU and CLARIN. The main content area is titled "UDPipe" and includes tabs for "About", "Run", and "REST API Documentation". A detailed description of UDPipe follows, mentioning its use for tokenization, tagging, lemmatization, and dependency parsing of CoNLL-U files. It is described as language-agnostic and trainable. The "Run" tab is selected, showing options to choose a model (e.g., UD 2.15, UD 2.12, UD 2.10, UD 2.6, PDT-C 1.0, EvaLatin) and perform actions like Tag and Lemmatize or Parse. An input text field contains the sentence "La mia mamma ha fatto una torta buonissima. Giorgio ne voleva una fetta.", and an "Input File" button is also present.

# Coreference resolution

- Consente ad un robot o a un «agente virtuale» di elaborare il linguaggio umano al fine di *trovare gli elementi sottintesi in più frasi dipendenti tra loro*



# Coreference resolution

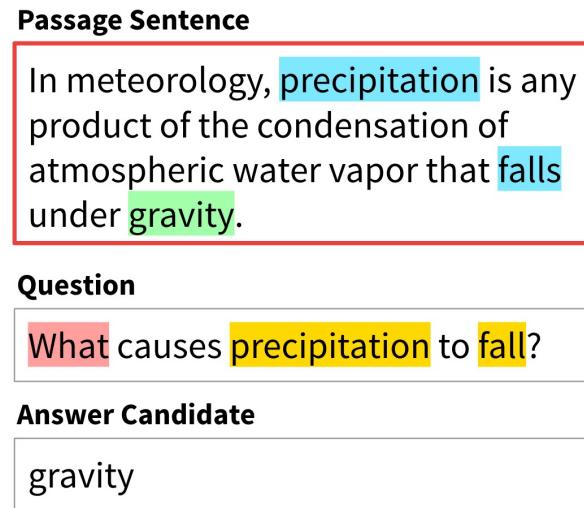
<https://huggingface.co/spaces/spark-nlp/coreference-resolution>

The screenshot shows a web browser window for the Hugging Face Spaces platform. The URL in the address bar is <https://huggingface.co/spaces/spark-nlp/coreference-resolution>. The page title is "State-of-the-Art Coreference Resolution in Spark NLP". On the left, there's a sidebar with a "Demo" section, a "Workflow & Model Overview" section where "spambert\_base\_coref" is selected as the pretrained model, and a "Reference notebook" section with a "Open in Colab" button. The main content area displays a sample sentence: "Alice went to the market. She bought some fresh vegetables there. The tomatoes she purchased were particularly ripe." Below this is a text input field labeled "Try it with your own Sentence!". Underneath, there's a section titled "Full example text" containing the same sentence again. At the bottom, there's a table titled "Processed output:" with three rows of data:

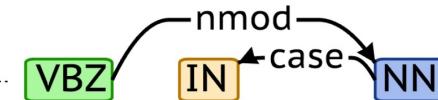
	tokens	corefs
1	alice	{"head:sentence": "1", "head": "ROOT", "head.begin": "1", "head.end": "1", "sentence": "0"}
2	she	{"head:sentence": "0", "head": "alice", "head.begin": "0", "head.end": "4", "sentence": "1"}
3	she	{"head:sentence": "0", "head": "alice", "head.begin": "0", "head.end": "4", "sentence": "2"}

# Question Answering

- Consente ad un robot o a un «agente virtuale» di elaborare il linguaggio umano al fine di *rispondere a delle domande*



- Path from **passage sentence words** (that also occur in **question**) to **answer**



- Combined with path from **wh-word** to **question word**.



# Question Answering

<https://rajpurkar.github.io/SQuAD-explorer/>

The screenshot shows a web browser displaying the SQuAD Explorer interface. The title bar reads "rajpurkar.github.io/SQuAD-explorer/". The main content area has a purple header with the text "SQuAD2.0" and "The Stanford Question Answering Dataset". Below the header, there are two main sections: "What is SQuAD?" and "Leaderboard".

**What is SQuAD?**

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

**SQuAD2.0** combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

[Explore SQuAD2.0 and model predictions](#)

[SQuAD2.0 paper \(Rajpurkar & Jia et al. '18\)](#)

**Leaderboard**

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1	IE-Net (ensemble) RICOH_SRCB_DML	90.939	93.214
2	FPNet (ensemble) Ant Service Intelligence Team	90.871	93.183
3	IE-NetV2 (ensemble) RICOH_SRCB_DML	90.860	93.100
4	SA-Net on Albert (ensemble) QIANXIN	90.724	93.011
5	SA-NetV2 (ensemble)	90.679	92.948

# Named Entity Recognition

- Consentire ad un robot o a un «agente virtuale» di comprendere il linguaggio umano per *analizzare la struttura delle frasi*

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE, Baidu ORG, and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space. The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the ‘future AI PERSON platforms’. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE. Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL, with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE.

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG, IBM ORG, and Microsoft ORG.

# Named Entity Recognition

- Consentire ad un robot o a un «agente virtuale» di comprendere il linguaggio umano per *analizzare la struttura delle frasi*

NER    Annotation    Recheck    Back

Dog Bite

CHIEF COMPLAINT: Dog bite to his right lower leg.

HISTORY OF PRESENT ILLNESS: This 50-year-old white male earlier this afternoon was attempting to adjust a cable that a dog was tied to. Dog was a German shepherd, it belonged to his brother, and the dog spontaneously attacked him. He sustained a bite to his right lower leg. Apparently, according to the patient, the dog is well known and is up-to-date on his shots and they wanted to confirm that. The dog has given no prior history of any reason to believe he is not a healthy dog. The patient himself developed a puncture wound with a flap injury. The patient has a flap **wound SYMPTOM** also below the puncture wound, a V-shaped flap, which is pointing towards the foot. It appears to be viable. The **wound SYMPTOM** is open about may be roughly a centimeter in the inside of the flap. He was seen by his medical primary care physician and was given a tetanus shot and the **wound SYMPTOM** was cleaned and wrapped, and then he was referred to us for further assessment.

PAST MEDICAL HISTORY: Significant for history of **pulmonary fibrosis DISEASE** and **atrial fibrillation DISEASE**. He is status post bilateral lung transplant back in 2004 because of the **pulmonary fibrosis DISEASE**.

ALLERGIES: There are no known allergies.

MEDICATIONS: Include multiple medications that are significant for his lung transplant including Prograf, **CellCept CHEMICAL**, **prednisone CHEMICAL**, **omeprazole CHEMICAL**, **Bactrim CHEMICAL** which he is on chronically, **folic acid CHEMICAL**, **vitamin D CHEMICAL**, Mag-Ox, Toprol-XL, **calcium CHEMICAL**, **500 mg DOSAGE**, vitamin B1, Centrum Silver, **verapamil CHEMICAL**, and **digoxin CHEMICAL**.

FAMILY HISTORY: Consistent with a sister of his has **ovarian cancer DISEASE** and his father had **liver cancer DISEASE**, **Heart disease DISEASE** in the patient's mother and father, and father also has **diabetes DISEASE**.

SOCIAL HISTORY: He is a non-cigarette smoker. He has occasional glass of wine. He is married. He has one biological child and three stepchildren. He works for

# Named Entity Recognition

<https://demos.explosion.ai/displacy-ent>

The screenshot shows a web browser window with the URL <https://demos.explosion.ai/displacy-ent>. The page has a dark blue header with the title "displaCy Named Entity Visualizer". Below the header, there is a search bar containing the text: "A little less than a decade later, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation." To the right of the search bar is a section titled "Entity labels (select all)" with several checkboxes. Some checkboxes are checked, including PERSON, ORG, GPE, DATE, and TIME. Below this is a "Model" dropdown set to "English - en\_core\_web\_sm (v3.5.0)". The main content area displays the same text with entities highlighted by colored boxes and labeled with their types: Sebastian Thrun (PERSON), Google (ORG), 2007 (DATE), American (NORP), Thrun (PERSON), Recode (ORG), and earlier this week (DATE). At the bottom of the page, there are two sections: "Using and customizing NER models" and "displaCy Named Entity Visualizer".

# Sentiment Analysis

- Consentire ad un robot o a un «agente virtuale» di comprendere il linguaggio umano per *rilevare le emozioni, la tipologia del discorso o il genere dell'autore*

SENTIMENT ANALYSIS



Discovering people opinions, emotions and feelings about a product or service

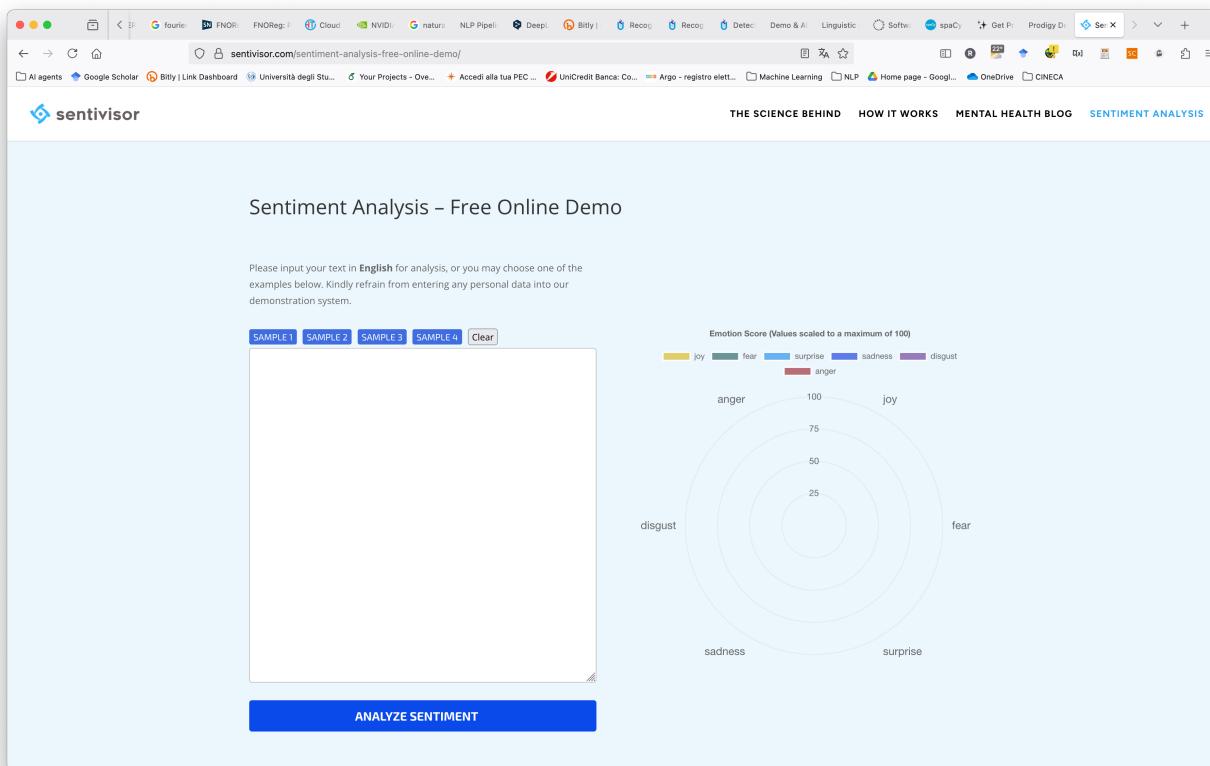
# Sentiment Analysis

<https://text2data.com/Demo>

The screenshot shows a web browser window for the URL <https://text2data.com/Demo>. The page title is "Free sentiment analysis demo". The main content area contains a text input box with the text: "When I woke up this morning I was like a cloud fluctuating in the sky. Then George entered my room and his face told me all in a moment: my world collapsed in a minute." Below the text input are two buttons: "Twitter-like content" (unchecked) and "SHARE THIS ANALYSIS". To the right of these buttons is a yellow "RUN ANALYSIS" button. The analysis results are displayed below the input text. It states: "This document is: negative (-0.64)" and "Magnitude: 1.03". A color scale bar indicates the score range from -1 (negative) to +1 (positive), with 0 being neutral. The subjectivity is labeled as "subjective". At the bottom, there are three call-to-action boxes: "Try sentiment analysis in Excel or Google Sheets (no programming required)", "Train your own document classification or extraction model", and "Train your own sentiment model using our SMTT Tool".

# Sentiment Analysis

<https://sentivisor.com/sentiment-analysis-free-online-demo/>



# Hate Speech Detection

- Consentire ad un robot o a un «agente virtuale» di comprendere il linguaggio umano per *rilevare le emozioni, la tipologia del discorso o il genere dell'autore*



# Hate Speech Detection

- Andiamo su Hugging Face

<https://huggingface.co/>



# Author Gender Detection

- Consentire ad un robot o a un «agente virtuale» di comprendere il linguaggio umano per *rilevare le emozioni, la tipologia del discorso o il genere dell'autore*



# Author Gender Detection

<https://app.readable.com/text/gender/>

The screenshot shows a web browser window for the URL <https://app.readable.com/text/gender/>. The interface is titled "Gender Analyzer". On the left, there is a text input area with the placeholder "Type or paste your text here to analyze its gender balance." Above this input area are navigation tabs: "Select or Add Website", "Go Prof", "Text", "Files", "URLs", and "Emails". To the right of the input area is a large pink button labeled "Analyze Gender". To the right of the button, the text "Gender Analyzer" is displayed, followed by a description: "Gender analysis identifies whether your text looks like it was written by a man or a woman. Our gender analysis tool looks at your text and compares it with a corpus of data with a known origin, looking at specific word frequencies to estimate the gender of the author. Gender analysis currently has an accuracy of about 70%." Below this section is a heading "Need More Power?" with a descriptive text about the Text Readability tool. Further down are sections for "ReadablePro" (with a "Comprehensive Readability Analysis" button) and "Free Tools" (listing "Keyword Density Analysis", "Gender Analysis", and "Profanity Detector").