



**Università  
degli Studi  
di Palermo**



# Introduzione al Corso

CORSO DI NATURAL LANGUAGE PROCESSING (ELABORAZIONE DEL LINGUAGGIO NATURALE)

a.a. 2023/2024

Prof. Roberto Pirrone

# Sommario

- Il docente
- Perché «Natural Language Processing»
- Cosa non è «Natural Language Processing»
- Cosa è «Natural Language Processing»
- Il Syllabus
- Il materiale didattico
- Gli esami
- Le tesi di laurea

# Il Docente

- Roberto Pirrone
  - Studio: Edificio 6, terzo piano, stanza 3025
  - Email: [roberto.pirrone@unipa.it](mailto:roberto.pirrone@unipa.it),  
[roberto.pirrone@community.unipa.it](mailto:roberto.pirrone@community.unipa.it) (Google)
  - Telefono studio: 091238.62625, laboratorio: .62643
  - Ricevimento: ogni mercoledì dalle 11:30 alle 13 presso il proprio studio

# Perché «Natural Language Processing»

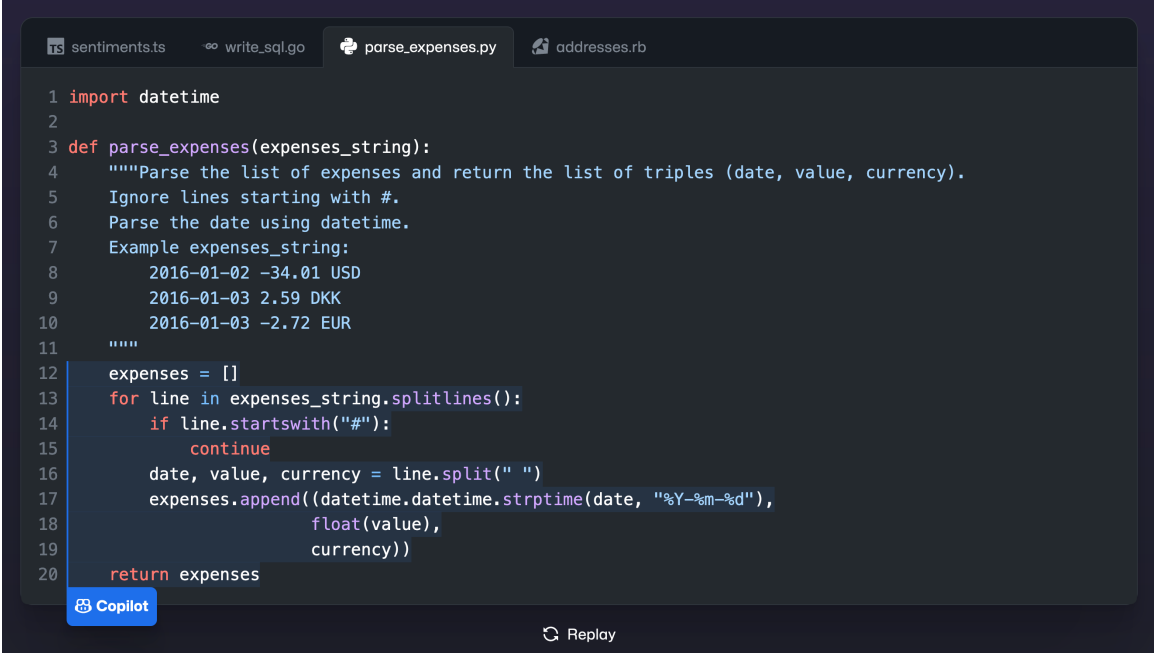
- Perché è ad oggi la tecnologia più «disruptiva» nel panorama dell'Intelligenza Artificiale
  - Assistenti virtuali avanzati eventualmente incorporati in un robot



<https://bit.ly/3PNdRd4>

# Perché «Natural Language Processing»

- Perché è ad oggi la tecnologia più «disruptiva» nel panorama dell'Intelligenza Artificiale
- Interazione fluente e semplice con l'uomo per compiti di analisi dei dati e/o programmazione



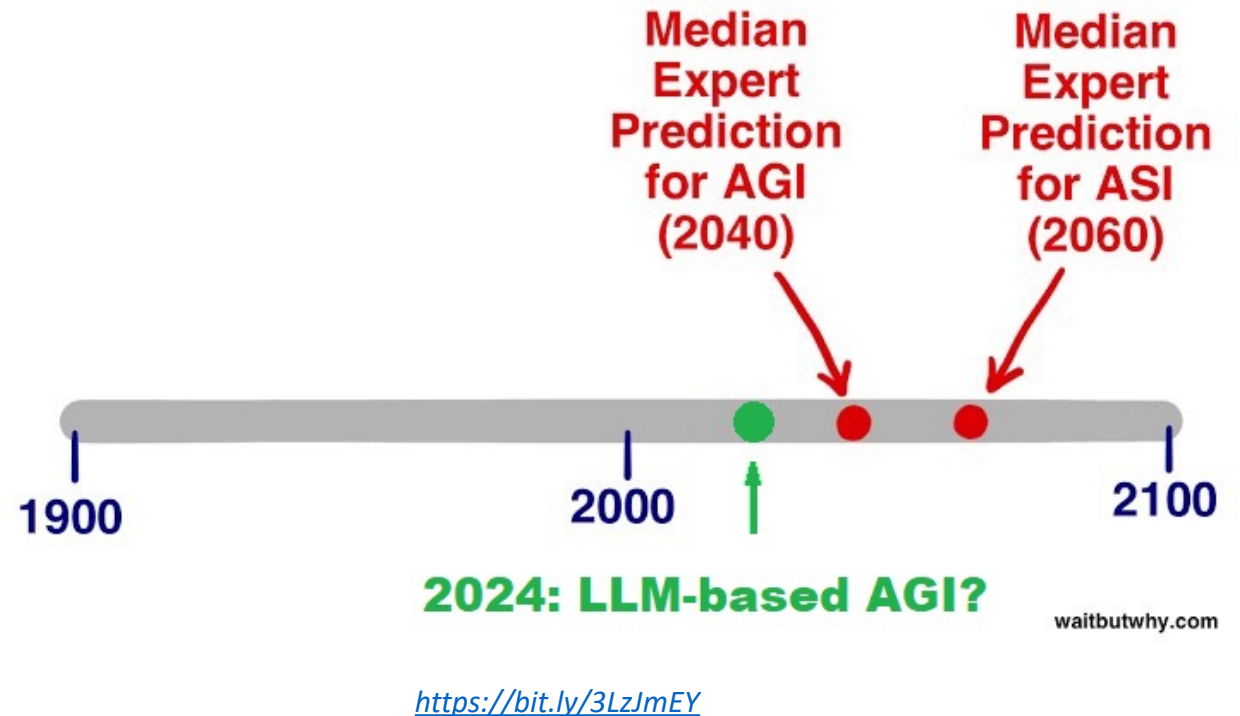
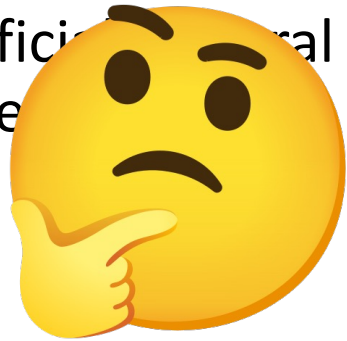
The screenshot shows a code editor with a dark theme. At the top, there are tabs for 'sentiments.ts', 'write\_sql.go', 'parse\_expenses.py', and 'addresses.rb'. The active tab is 'parse\_expenses.py'. The code is a Python function named 'parse\_expenses' that takes an 'expenses\_string' as input. It parses the string into a list of tuples, each containing a date, a value, and a currency. The code includes comments and an example input string. At the bottom left, there is a 'Copilot' button, and at the bottom right, there is a 'Replay' button.

```
1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, value, currency).
5     Ignore lines starting with #.
6     Parse the date using datetime.
7     Example expenses_string:
8         2016-01-02 -34.01 USD
9         2016-01-03 2.59 DKK
10        2016-01-03 -2.72 EUR
11    """
12    expenses = []
13    for line in expenses_string.splitlines():
14        if line.startswith("#"):
15            continue
16        date, value, currency = line.split(" ")
17        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
18                        float(value),
19                        currency))
20    return expenses
```

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

# Perché «Natural Language Processing»

- Perché è ad oggi la tecnologia più «disruptiva» nel panorama dell'Intelligenza Artificiale
- La nuova frontiera della «Artificial Consciousness» o della «Artificial General Intelligence»



# Cosa non è «Natural Language Processing»

- Il corso di «Natural Language Processing» *non è*:
  - Un corso di PyTorch (anche se impareremo a conoscerlo e ad usarlo)
  - Una serie di tutorial su framework esoterici: NLTK, Hugging Face, ... (ma anche questi studieremo)
  - Un tutorial per sviluppare applicazioni per vincere competizioni Kaggle o simili (anche se faremo un po' anche di questo)

# Cos'è «Natural Language Processing»

- Il corso di «Natural Language Processing» ci fornirà una panoramica delle più recenti 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



# Cos'è «Natural Language Processing»

- Consentire ad un robot o a un «agente virtuale» di comprendere il linguaggio umano per *rispondere a delle domande*

## Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

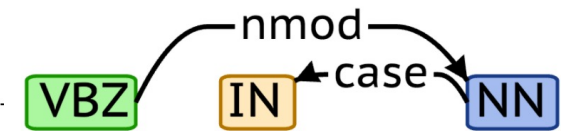
## Question

What causes precipitation to fall?

## Answer Candidate

gravity

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



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



# Cos'è «Natural Language Processing»

- 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**.

# Cos'è «Natural Language Processing»

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

The screenshot displays a web-based NER interface with a dark header containing tabs for 'NER', 'Annotation', and 'Recheck', and a 'Back' button on the right. The main content area shows a medical record for a 'Dog Bite' case. The text is annotated with colored boxes representing different entity types: purple for 'wound SYMPTOM', green for 'DISEASE', orange for 'CHEMICAL', and blue for 'DOSAGE'. The text includes sections for Chief Complaint, History of Present Illness, Past Medical History, Allergies, Medications, Family History, and Social History.

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.

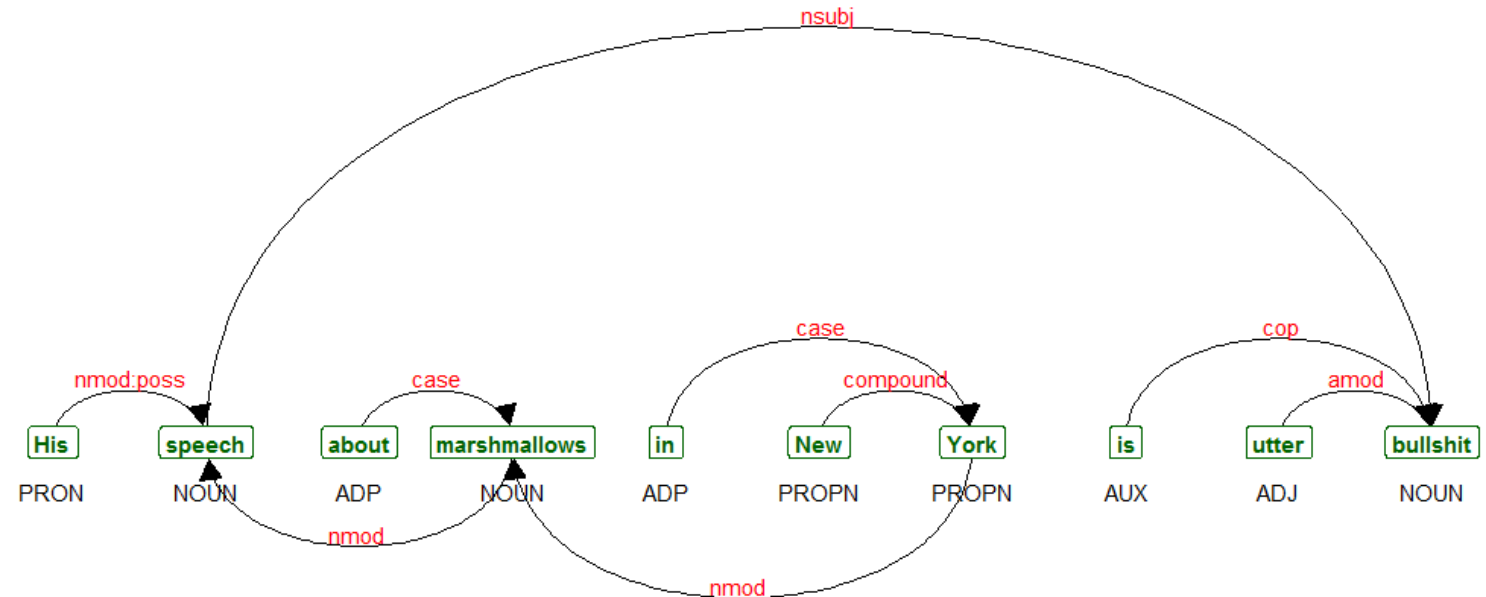
SOCIAL HISTORY: He is a non-cigarette smoker. He has occasional glass of wine. He is married. He has one biological child and three stepchildren. He works for

# Cos'è «Natural Language Processing»

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

## udpipe output

tokenisation, parts of speech tagging & dependency relations



# Cos'è «Natural Language Processing»

- 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



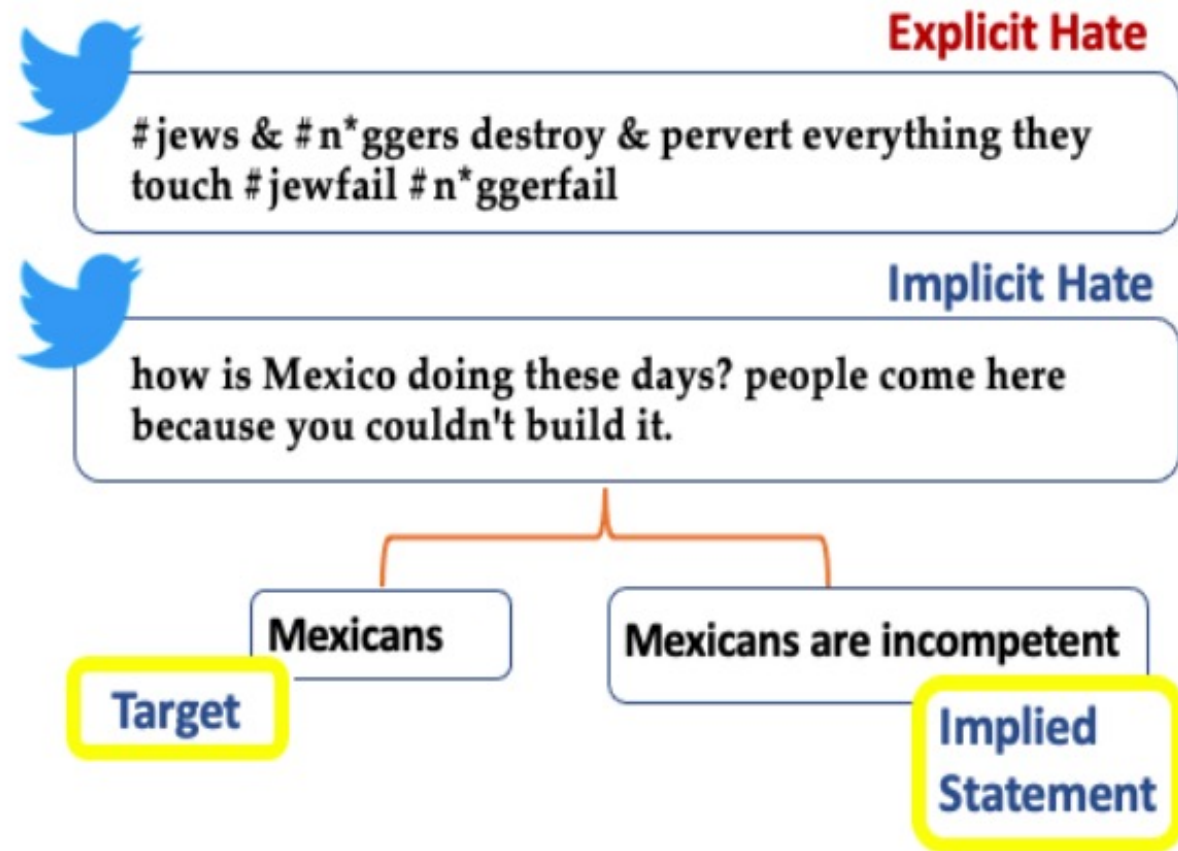
# Cos'è «Natural Language Processing»

- 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*



# Cos'è «Natural Language Processing»

- 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*



# Cos'è «Natural Language Processing»

- 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*



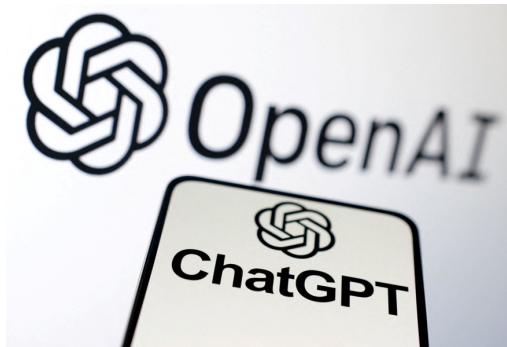


# Cos'è «Natural Language Processing»

- Large Language Models



<https://bit.ly/459pPSO>



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




Please, use the statistics Python package that in your opinion is the best in the public domain, and create a program to compute the conditional probability that when I have tossed three coins I obtain at least two heads



There are several popular Python packages for statistics, including ``scipy``, ``numpy``, and ``statsmodels``. Depending on your specific use case, one may be more appropriate than the others. For example, ``scipy`` is often used for scientific computing, while ``statsmodels`` is geared more towards statistical modeling.

One package that is well suited for probability calculations is the ``scipy.stats`` module. You can use it to compute the conditional probability of getting at least 2 heads when tossing 3 coins. Here's an example of how you could use this package to calculate this probability:

python

 Copy code

```
from scipy.stats import binom

# The number of trials
n = 3

# The probability of success in each trial
p = 0.5

# The number of successful outcomes
k = 2

# P(X >= k)
result = 1 - binom.cdf(k-1, n, p)

print(result)
```

This will give you the probability of getting at least two heads in three coin tosses, which is

0.375

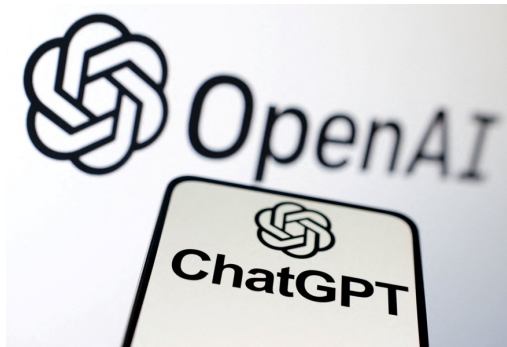
 Regenerate response

# Cos'è «Natural Language Processing»

- Large Language Models



<https://bit.ly/459pPSO>



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



Could you rewrite the first piece of code in a more Pythonic way please?



Sure! Here is an example of how the first piece of code could be written in a more "Pythonic" way:

python

Copy code

```
from math import comb

n = 3 # number of coin tosses
k = 2 # number of heads

# probability of getting at least 2 heads
result = sum(comb(n, i) * 0.5**i * 0.5**(n-i) for i in range(k, n+1))

print(result)
```

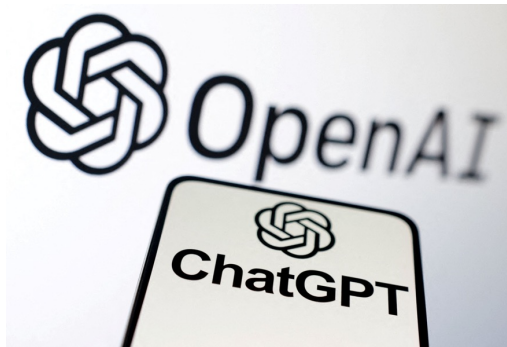
This way, you can avoid importing scipy package, and use python built-in combinatorics method.

# Cos'è «Natural Language Processing»

- Large Language Models



<https://bit.ly/459pPSO>



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



Scrivimi die strofe di una canzone nello stile di Fabrizio De André



Eccoti le prime quattro strofe di una canzone ispirata allo stile di Fabrizio De André:



Nella città che scivola via  
tra le case e le luci della città  
c'è chi cammina senza una meta  
e chi corre per la propria verità

C'è chi vive per il domani  
e chi rimpiange il passato già  
ma la vita è un susseguirsi di attimi  
che non si possono fermare mai

E ci sono i sognatori  
che cercano di cambiare il mondo  
con le parole e le idee  
che portano dentro al cuore profondo

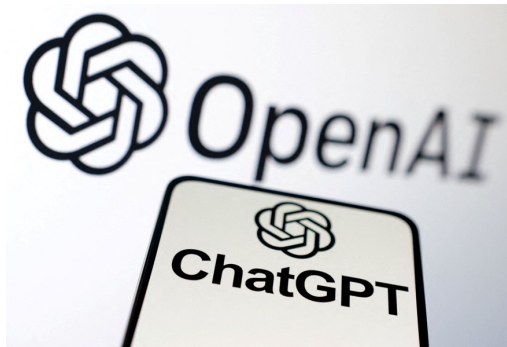
Ma le strade sono piene di ombre  
e le notti spesso sono vuote  
e il tempo passa inesorabile  
mentre il cuore batte forte

# Cos'è «Natural Language Processing»

- Large Language Models



<https://bit.ly/459pPSO>



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



Alice porta al mercato 25 chili di patate che vende a 3.5€ all'etto. Quanto guadagna se il costo del trasporto ammonta a 10€?



Per risolvere questo problema, dobbiamo prima calcolare la quantità di patate vendute in etti e poi calcolare il ricavo totale sottraendo il costo del trasporto.



25 chili corrispondono a 25000 grammi (1 chilo = 1000 grammi). Se vendiamo le patate a 3.5€ all'etto, dobbiamo calcolare il prezzo per 100 grammi (1 etto = 100 grammi).

Il prezzo per 100 grammi è:

$3.5€ / 10 \text{ etti} = 0.35€ \text{ per etto}$

Il ricavo totale è quindi:

$25000 \text{ grammi} / 100 \text{ grammi per etto} = 250 \text{ etti di patate venduti}$

$250 \text{ etti} \times 0.35€ \text{ per etto} = 87.5€ \text{ di ricavo totale}$

Sottraendo il costo del trasporto di 10€, otteniamo il guadagno totale di Alice:

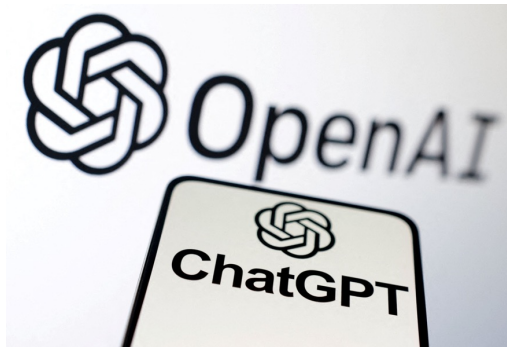
$87.5€ - 10€ = 77.5€ \text{ di guadagno totale.}$

# Cos'è «Natural Language Processing»

- Large Language Models



<https://bit.ly/459pPSO>



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



Ma il prezzo di vendita da me indicato era già un prezzo per etto di patate. Puoi riformulare la soluzione?



Certamente, grazie per la precisazione.



Se Alice vende 25 chili di patate al prezzo di 3.5€ per etto, dobbiamo prima calcolare quanti etti di patate ha in totale:

$25 \text{ chili} \times 10 \text{ etti per chilo} = 250 \text{ etti di patate}$

Il ricavo totale della vendita sarà:

$250 \text{ etti} \times 3.5\text{€ per etto} = 875\text{€}$

Tuttavia, Alice deve anche sostenere un costo di 10€ per il trasporto. Quindi, il suo guadagno totale sarà:

$875\text{€} - 10\text{€} = 865\text{€}$

Alice guadagna quindi 865€ dalla vendita delle patate, dopo aver sostenuto un costo di trasporto di 10€.



# Cos'è «Natural Language Processing»

- Tutto questo richiederà un po' di *appoggio esterno*
  - Le conoscenze acquisite nel modulo di Analisi dei Big Data
    - Algoritmi di Machine Learning/Deep Learning
    - Analisi statistica dei nostri dati

# Il Syllabus

- Le informazioni complete sugli obiettivi didattici del corso, il programma delle lezioni e i libri di testo si trovano nella *Scheda di Trasparenza* aggiornata sul repository del corso
  - [Elaborazione del Linguaggio Naturale](#)

# II Syllabus

- Testi consigliati
  - Daniel *Jurafsky*, James H. Martin “Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”, 3d Edition draft, 2023, disponibile on line all’indirizzo: <https://web.stanford.edu/~jurafsky/slp3/>
  - Steven *Bird*, Ewan Klein, and Edward Loper, “Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit” versione on line all’indirizzo: <https://www.nltk.org/book/>



# Il Syllabus

ORE	Lezioni Frontali	Testo rif.
1	Introduzione al Corso.	Slide docente
2	Processing del testo: tokenizzazione, segmentazione, normalizzazione, lemmatizzazione, stemming.	Estratti dal Jurafsky cap. 2
2	Modelli del linguaggio con N-grammi.	Estratti dal Jurafsky cap. 3
4	Part-of-Speech Tagging e Named Entity Recognition.	Estratti dal Jurafsky cap. 8
2	Semantica vettoriale ed embedding non contestuali: word2vec, GloVE, fasttext.	Estratti dal Jurafsky cap. 6

# Il Syllabus

ORE	Lezioni Frontali	Testo rif.
4	Modelli neurali del linguaggio: reti feed-forward, reti ricorrenti e meccanismo di attenzione.	Estratti dal Jurafsky capp. 7 e 9
3	Traduzione automatica: il meccanismo di self-attention e i Transformer.	Estratti dal Jurafsky cap. 10
3	Masked Language Models pre-addestrati: BERT, RoBERTa e derivati.	Estratti dal Jurafsky cap. 11
3	Large Language Models: ChatGPT e LLAMA2, prompting, RAG e fine-tuning.	Slide docente

# Il Syllabus

ORE	Esercitazioni
3	Introduzione a PyTorch.
3	Text processing con NLTK e modelli a N-grammi.
3	POS e NER con NLTK.
3	Embedding e reti neurali per la text classification.
3	Trasformer per la traduzione automatica con la piattaforma Hugging Face.
3	Uso di BERT da Hugging Face per i task NLP già illustrati in precedenza.
3	Uso di RoBERTa e modelli multilingua da Hugging Face per lo svolgimento di task in italiano.
3	Uso di LLAMA2-7b da Hugging Face per lo svolgimento di task NLP.

- Il riferimento per le esercitazioni saranno dei Notebook Colab predisposti dal docente
- Il libro di riferimento per NLTK è il Bird

# Il materiale didattico

- Le slide da sole *non sono* materiale didattico: esse sono a compendio dei libri di testo, della spiegazione orale del docente e degli *appunti* presi dallo studente

# Il materiale didattico

- Repository GitHub del corso
  - [https://github.com/fredffsixty/Natural Language Processing](https://github.com/fredffsixty/Natural_Language_Processing)
  - Contiene:
    - I file pdf di tutte le slide (incluse queste)
    - I codici delle esercitazioni (rilasciati *dopo* lo svolgimento delle stesse)
    - I dati utilizzati nelle esercitazioni

# Gli esami

- Progetto:
  - Parteciperete, formando dei gruppi, ad una challenge di NLP in italiano o inglese
  - Sarete liberi di implementare la soluzione, tra quelle studiate, che riterrete più opportuna per affrontare il task
  - Cominceremo lo sviluppo del progetto nelle ultime lezioni del corso
  - Vi sarà fornito un training set e un development test (test set per affinare la vostra soluzione)
  - Vi sarà indicata una procedura per l'upload della vostra predizione su un test set, rilasciato dopo la fine della challenge, che servirà da benchmark per costituire la graduatoria dei gruppi

# Gli esami

- *Il posizionamento del vostro gruppo in graduatoria non pregiudicherà la possibilità di ottenere un voto alto*
- Il voto del progetto costituisce proposta di voto finale
- I singoli potranno accedere ad un eventuale orale per migliorare il proprio voto di progetto

# Le tesi di laurea

- Vi verranno proposti dei possibili argomenti di tesi di laurea da condurre presso il nostro Laboratorio (CHILab – Laboratorio di Interazione Uomo-Macchina) su temi inerenti il Deep Learning e l'IA:
  - NLP per l'analisi di corpora testuali di interesse medico
  - Deep Learning a supporto del Drug Discovery
  - Elaborazione di immagini mediche (TAC/RM)
  - Sistemi di Integrazione delle predizioni di una DNN negli standard di gestione delle immagini mediche
- Altra alternativa possono essere le tesi aziendali che abbiano attinenza con Il Natural Language Processing e *siano di interesse per il nostro laboratorio*



# Le tesi di laurea

- Vincoli sull'assegnazione della tesi
  - Che ci sia uno slot libero (max 5 tesisti in contemporanea, altrimenti deve prima laurearsi qualcuno per poter avere la tesi)
  - Che siano garantiti almeno sei mesi effettivi di lavoro
    - Consecutivamente al netto delle altre materie e del tirocinio
    - Non ha senso chiedere la tesi un anno prima, quando ancora si devono sostenere altri esami e si «scompare» per mesi
    - Manifestate comunque il vostro interesse!!

# Le tesi di laurea

- Vincoli sull'assegnazione della tesi
  - Che si concordi su una delle tematiche proposte e su un livello minimo di obiettivi concordato al momento dell'assegnazione
  - Che l'argomento di tesi industriale risulti di interesse per il nostro gruppo di ricerca
- Si verrà affidati ad uno o più dottorandi del laboratorio con i quali si dovranno avere incontri (anche on line) *al più bi-settimanali* sullo stato di avanzamento lavori