Python Biella Group

# Python per il testo e i documenti

## Come usare Python per grandi collezioni di testo?

a cura di: **Andrea Guzzo**

# Di cosa parliamo oggi?

In un'azienda spesso si hanno tantissimi documenti testuali o sorgenti di testo

- Grandi collezioni di testo (o anche pdf e documenti)
- Molto sparse e differenti tra di loro (lingua, sintassi, argomenti, ...)
- Difficili da memorizzare e gestire per estrarre informazioni utili

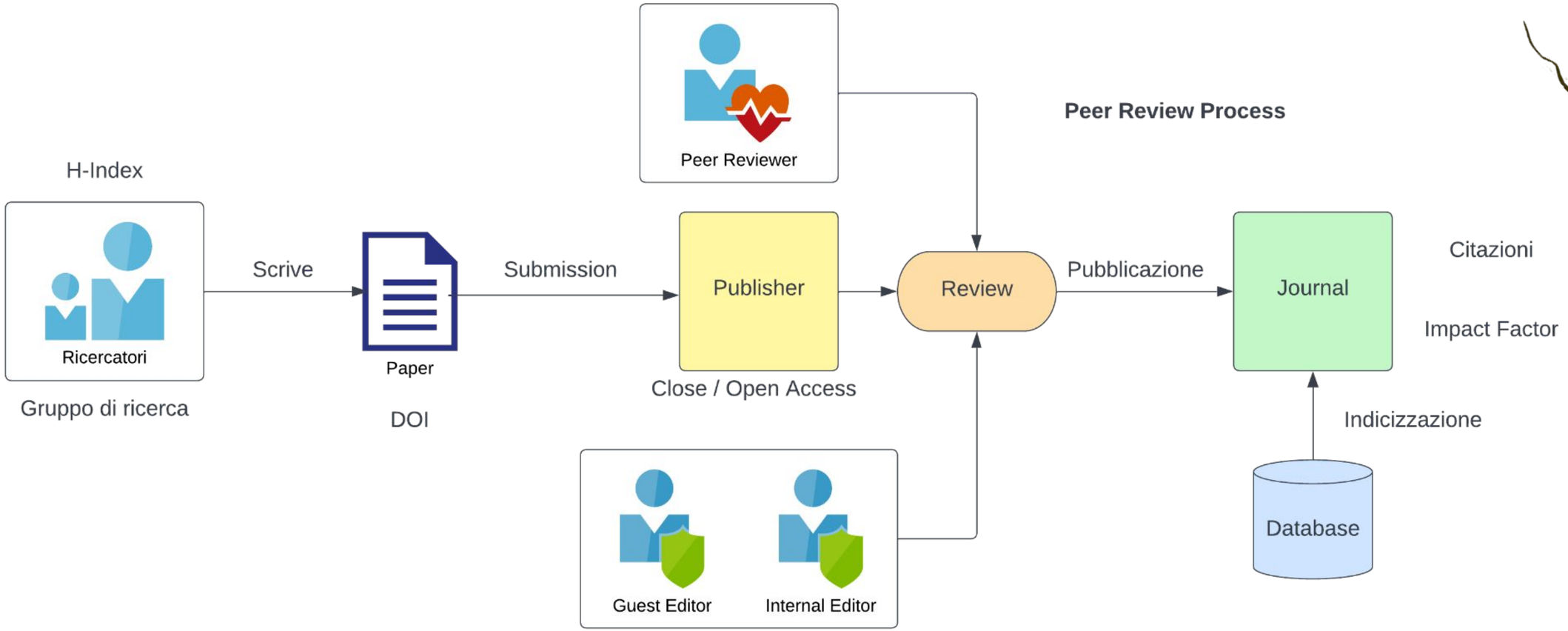
Come fare per analizzare tantissimi documenti assieme per estrarre informazioni utili?

- Capire di cosa parla un testo (analisi semantica)
- Costruire applicazioni per ricercare agilmente informazioni (correlazioni, similarità, ...)
- Capire se ci sono delle analogie tra i documenti
- Scoprire di cosa parlano

Per fare questo dobbiamo usare: matematica, statistica, informatica e Python :)

# Come funziona il mondo della ricerca?

Il mondo della ricerca è molto grande e complesso: ma sorregge il mondo :)




Stiamo parlando di 400.000 articoli l'anno (almeno)

# Un esempio?

Website Google Scholar

1 December 2021

Conference Paper › Published by Institute of Electrical and Electronics Engineers (IEEE) in 2021 6th International Conference on Information Technology Research (ICITR)

Part of 2021 6th International Conference on Information Technology Research (ICITR), 2021-12-1 - 2021-12-3, Moratuwa, Sri Lanka

p. 1-6. https://doi.org/10.1109/icitr54349.2021.9657164

## How to pretrain an efficient cross-disciplinary language model: The ScilitBERT use case

Jean-Baptiste de la Broise Nolwenn Bernard Jean-Philippe Dubuc Andrea Perlato Bastien Latard

### Abstract

Transformer based models are widely used in various text processing tasks, such as classification, named entity recognition. The representation of scientific texts is a complicated task, and the utilization of general English BERT models for this task is suboptimal. We observe the lack of models for multidisciplinary academic texts representation, and on a broader scale, a lack of specialized models pretrained on specific domains, for which general English BERT models are suboptimal. This paper introduces ScilitBERT, a BERT model pretrained on an inclusive cross-disciplinary academic corpus. ScilitBERT is half as deep as RoBERTa, and has a much lower pretraining computation cost. ScilitBERT obtains at least 96% of RoBERTa's accuracy on two academic domain downstream tasks. The presented cross-disciplinary academic model has been publicly released [1] [1] https://github.com/JeanBaptiste-dlb/ScilitBERT. The results obtained show that for domains that use a technolect and have a sizeable amount of raw text data; the pretraining of dedicated models should be considered and favored.

### Keywords

VOCABULARY TEXT RECOGNITION COMPUTATIONAL MODELING BIT ERROR RATE BENCHMARK TESTING TRANSFORMERS DATA MODELS

### Related Publications

Improving Real-time Recognition of Morphologically Rich Speech with Transformer Language Model

OFDM Based on Low Complexity Transform to Increase Multipath Resilience and Reduce PAPR
IEEE Transactions on Signal Processing

Robust color image watermarking using the Spatio-Chromatic Fourier Transform and semi-random LDPC codes

Acoustic environment classification using discrete hartley transform features

The Out-of-Vocabulary Spelling Correction Model in a New Domain for End-to-End Automatic Speech Recognition

Adaptation of Precision Matrix Models on Large Vocabulary Continuous Speech Recognition

Emformer: Efficient Memory Transformer Based Acoustic Model for Low Latency Streaming Speech Recognition

A Novel Reversible Data Hiding Scheme in Encrypted Images using Arnold Transform

Multi-level language modeling and decoding for open vocabulary end-to-end speech recognition

Correction of Automatic Speech Recognition with Transformer Sequence-To-Sequence Model

https://app.scilit.net/publications/2f886c5a1a8858cba1e81d837276cf1c

# Ma quindi cosa si può fare?

Siccome abbiamo tanti paper e journal possiamo ad esempio:

- Trovare delle similarità semantiche: Article Similarity
- Trovare simili autori e peer reviewer: Peer Reviewer Finder
- Capire di cosa parlano certi articoli: Topic Finder
- Raggruppare articoli simili: Article Clustering
- Estrarre parole chiave: Topic Modelling
- Capire se ci sono delle copiature o dei problemi: Ethical Project
- Migliorare il processo di revisione manuale
- Suggerire riviste e articoli agli utenti: Finder
- Creare dashboard per l'analisi dei dati
- E tanto tanto altro ancora...

# Come si può fare?

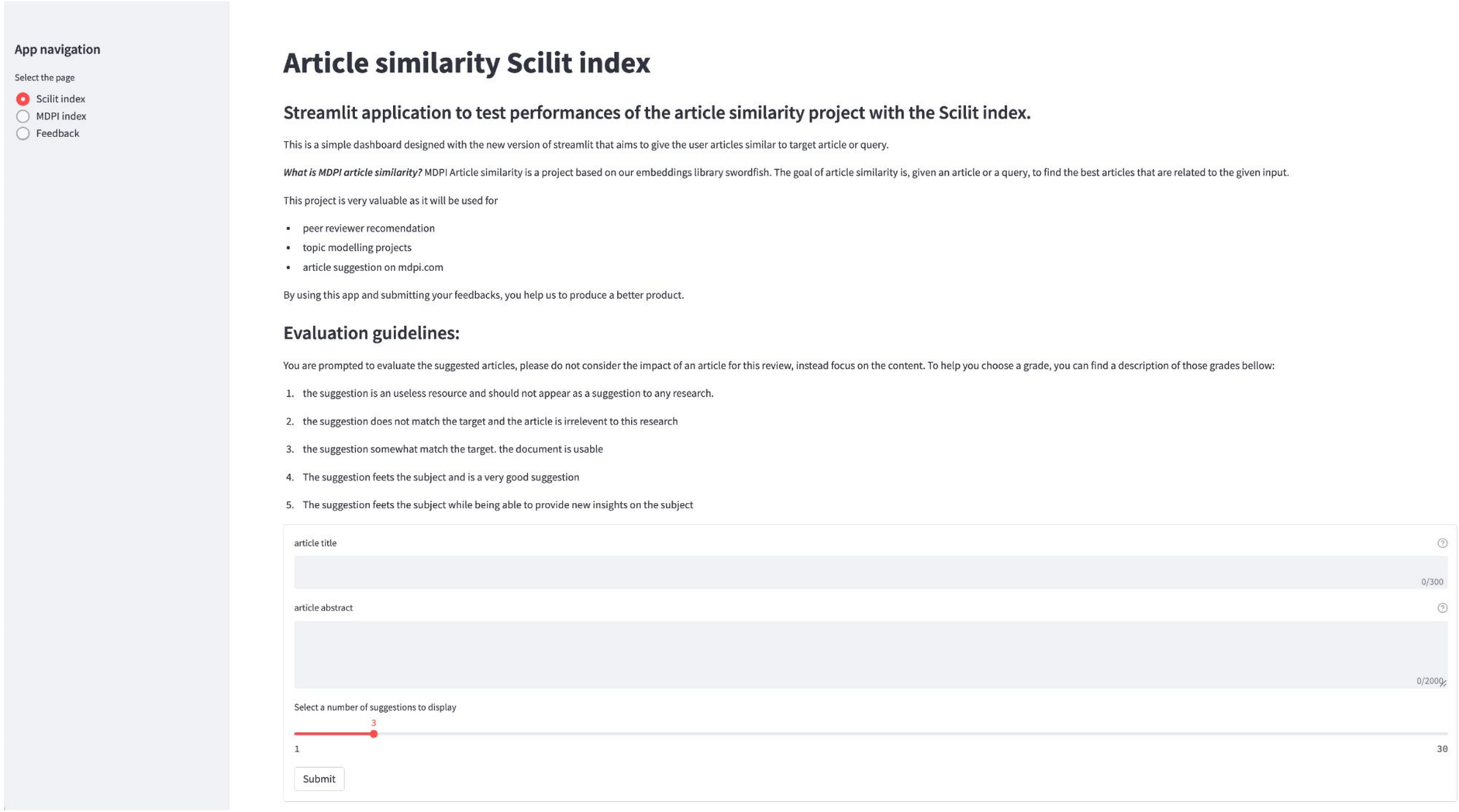
Prima di riuscire a realizzare quelle applicazioni servono tanti passi:

1. Organizzare i dati in modo che siano interpretabili
2. Avere una rappresentazione digitale dei documenti e del testo
3. Eseguire interrogazioni molto velocemente sulla base dati
4. Avere degli ambienti di sviluppo consoni (server, cloud, ...)
5. Iniziare a costruire algoritmi e prototipi
6. Architettura di produzione che consenta di utilizzare i servizi realizzati

In particolare ci servono degli algoritmi di NLP (Natural Language Processing) basati su:

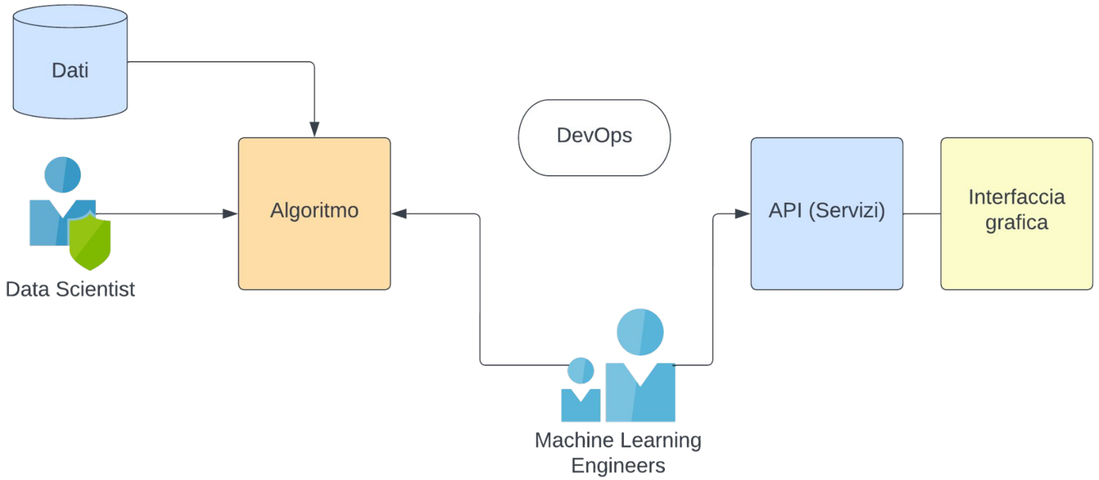
- Algoritmi classici: TF-IDF, Stemmer, Lemmatizers, ...
- Machine Learning: HDBScan, Riduzioni dello spazio, ...
- Deep Learning: Bert, SPECTER, Transformers, ...

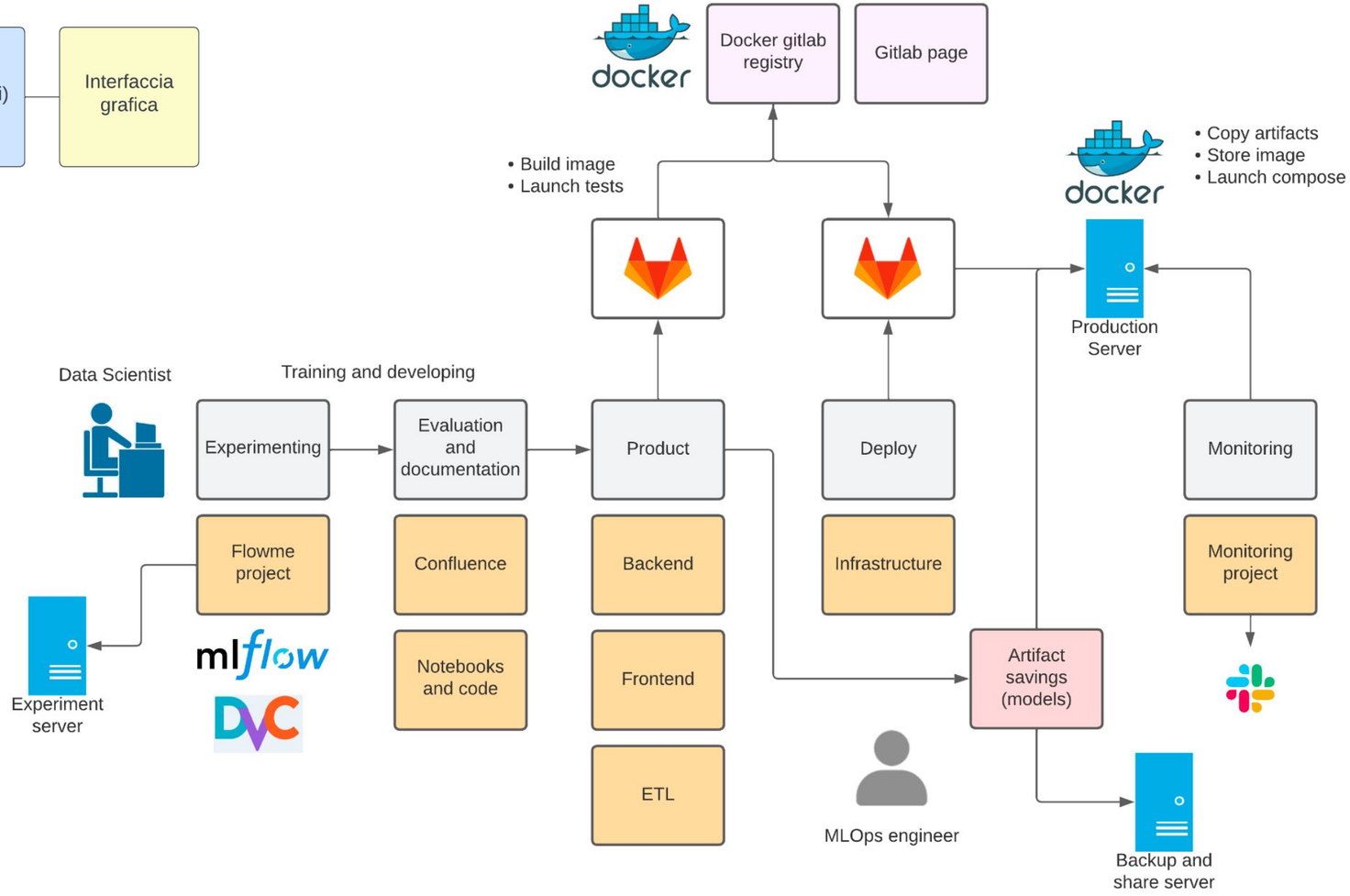
# Vediamo qualche esempio



# Scendiamo nel dettaglio

## Come si realizza un'applicazione di questo tipo?

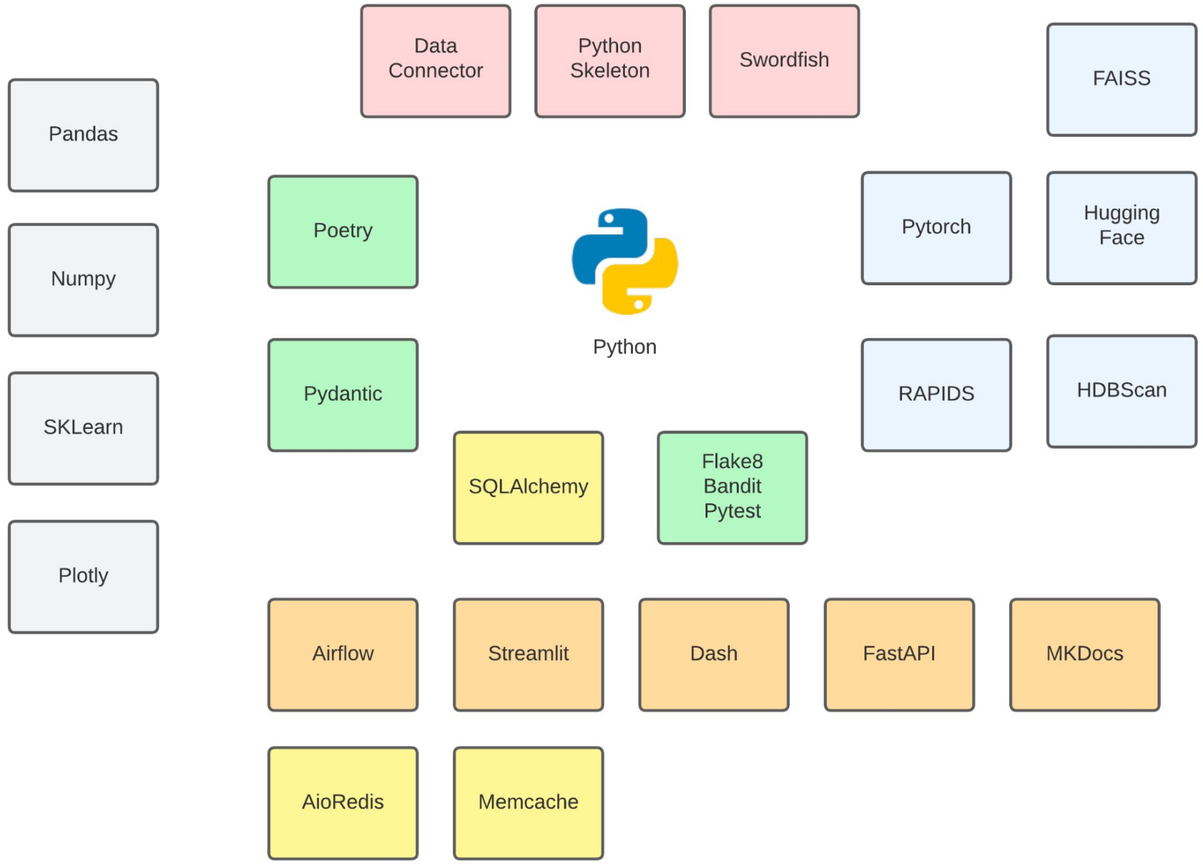


# Come ci può aiutare Python?

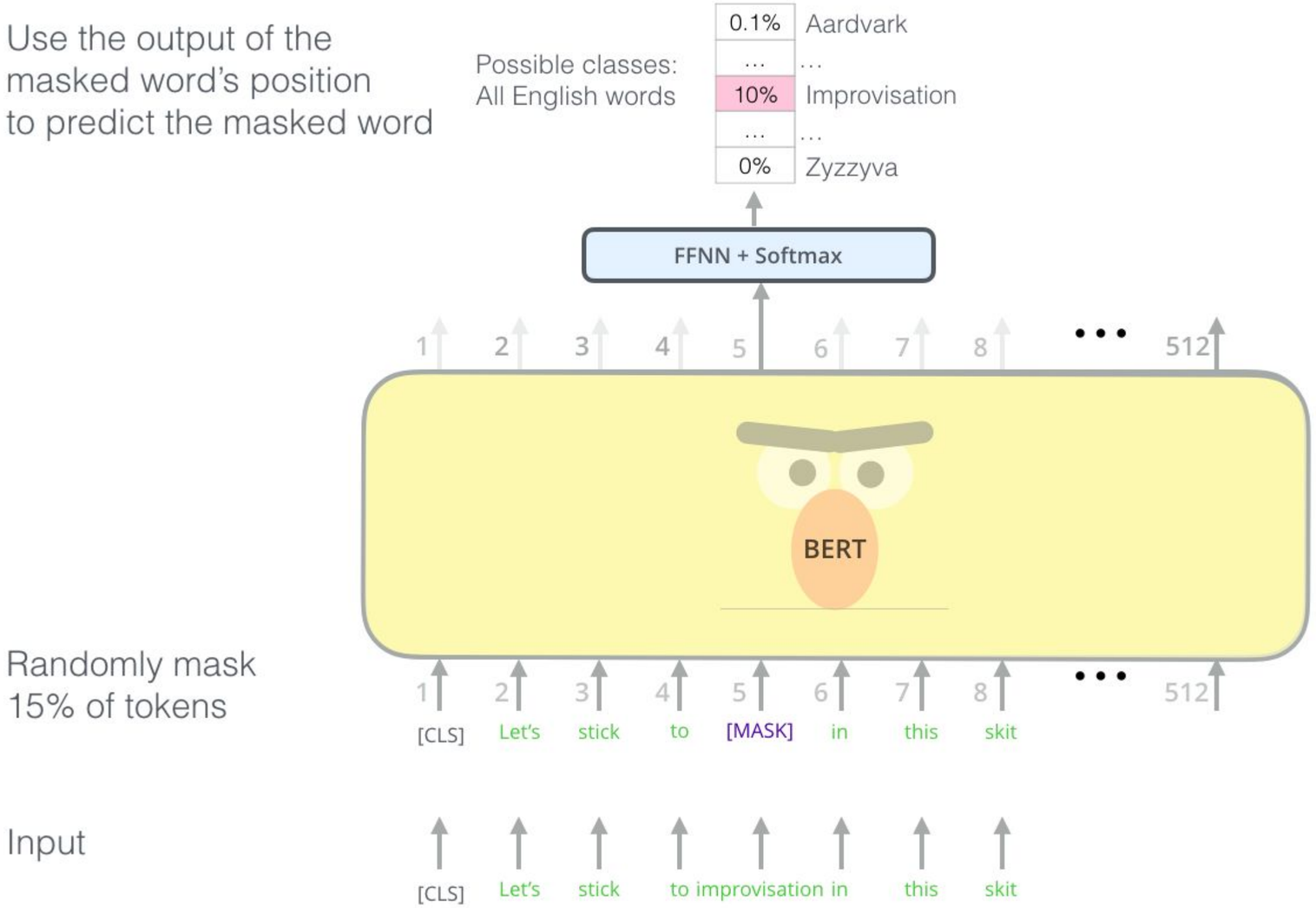
Python è il linguaggio generale che ci aiuta a costruire tutto questo

- Utilizzando framework e librerie
- Un solo linguaggio per tutto il flusso di lavoro



# Fantascienza?

## No dai! Solamente tanta matematica, statistica e informatica




HUGGING FACE

NVIDIA CUDA

PyTorch

RAPIDS

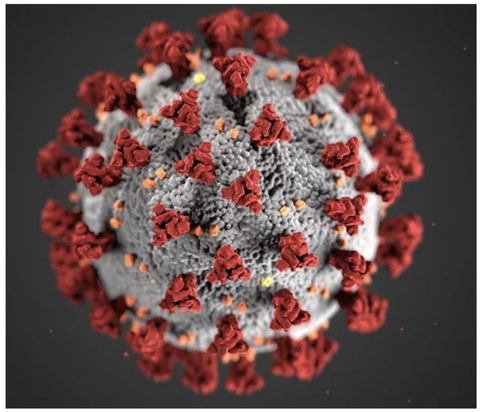
https://jalammar.github.io/illustrated-bert/

# E il futuro? Cosa ci riserva?

Stiamo lavorando a tantissimi nuovi strumenti:

- Global finder models
- Progetti legati all'etica e plagiarismo (molto molto importante)
- Graph machine learning e database basati su grafo (knowledge graph)

### Retracted coronavirus (COVID-19) papers



via CDC

We've been tracking retractions of papers about COVID-19 as part of our database. Here's a running list, which will be updated as needed. (For some context on these figures, see this post, our letter in *Accountability in Research* and the last section of this *Nature* news article. Also see a note about the terminology regarding preprint servers at the end.)

### The Retraction Watch Leaderboard

Who has the most retractions? Here's our unofficial list (see notes on methodology), which we'll update as more information comes to light:

1. Yoshitaka Fujii (total retractions: 183) See also: Final report of investigating committee, our reporting, additional coverage
2. Joachim Boldt (163) See also: Editors-in-chief statement, our coverage
3. Hironobu Ueshima (119) See also: our coverage
4. Yoshihiro Sato (106) See also: our coverage
5. Ali Nazari (87) See also: our coverage
6. Jun Iwamoto (82) See also: our coverage
7. Diederik Stapel (58) See also: our coverage
8. Yuhji Saitoh (55) See also: our coverage
9. Adrian Maxim (48) See also: our coverage
10. Chen-Yuan (Peter) Chen (43) See also: SAGE, our coverage
11. Fazlul Sarkar (41) See also: our coverage
12. Shahaboddin Shamshirband (41) See also: our coverage
13. Hua Zhong (41) See also: journal notice
14. Shigeaki Kato (40) See also: our coverage
15. James Hunton (37) See also: our coverage
16. Hyung-In Moon (35) See also: our coverage
17. Antonio Orlandi (34) See also: our coverage
18. Dimitris Liakopoulos (33) (NB: We're counting a book he co-authored as a single retraction. The book has 13 retracted chapters with DOIs that are not included in this figure.) See also: our coverage

https://retractionwatch.com/

### Springer Nature slaps more than 400 papers with expressions of concern all at once


Cartoon by Hilda Bastian (license)

A total of 436 papers in two Springer Nature journals are being subjected to expressions of concern, in the latest case of special issues — in this case, "topical collections" — likely being exploited by rogue editors or impersonators.

# Forte vero? :)

# JOIN US!

## C’è sempre qualcosa da imparare per migliorarci e crescere... insieme!

- **Nostri canale e materiale:  https://linktr.ee/PythonBiellaGroup**

Tutto questo è stato reso possibile grazie a

- **Tutta la community di P.B.G.**
- **Maria Teresa Panunzio**: https://www.linkedin.com/in/maria-teresa-panunzio-27ba3815/
- **Mario Nardi**: https://www.linkedin.com/in/mario-nardi-017705100/
- **Andrea Guzzo**: https://www.linkedin.com/in/andreaguzzo/
- **Davide Airaghi:** https://www.linkedin.com/in/airaghidavide/

# INIZIAMO!

https://github.com/PythonBiellaGroup/MaterialeSerate/

Python Biella Group

# Agenda

## Incontri e serate

**01** Lorem ipsum

**02** Lorem ipsum

**03** Lorem ipsum

**04** Lorem ipsum

## Materiale e codice sorgente

https://github.com/PythonBiellaGroup/

Python Biella Group

# Obiettivo della serata

## Descrizione obiettivo

- Obiettivo 1
- Obiettivo 2
- Obiettivo 3

# MACRO ARGOMENTO

## Argomento 1

- Sotto argomento 1
- Sotto argomento 2
- Sotto argomento 3

## Argomento 2

- Lista 1
- Lista 2
- Lista 3

## Argomento 3

- Sotto argomento 1
- Sotto argomento 2