

MOTORI DI RICERCA

DAGLI ALGORITMI ALL'INTELLIGENZA ARTIFICIALE

Paolo Ferragina

PLS: 18 Novembre 2022

UNIVERSITÀ DI PISA

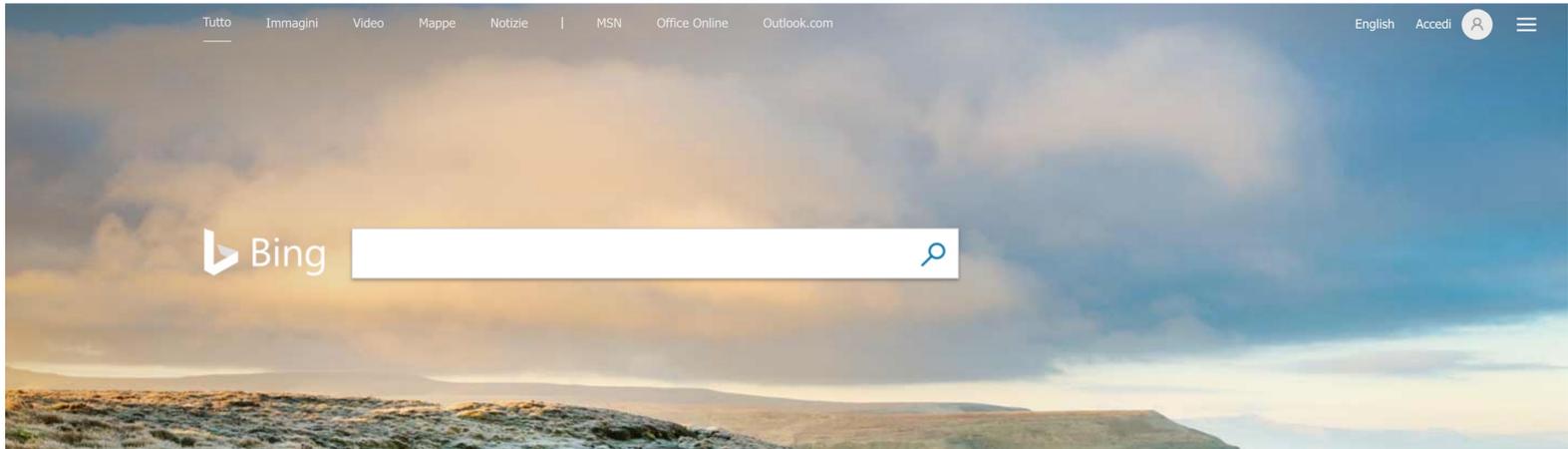


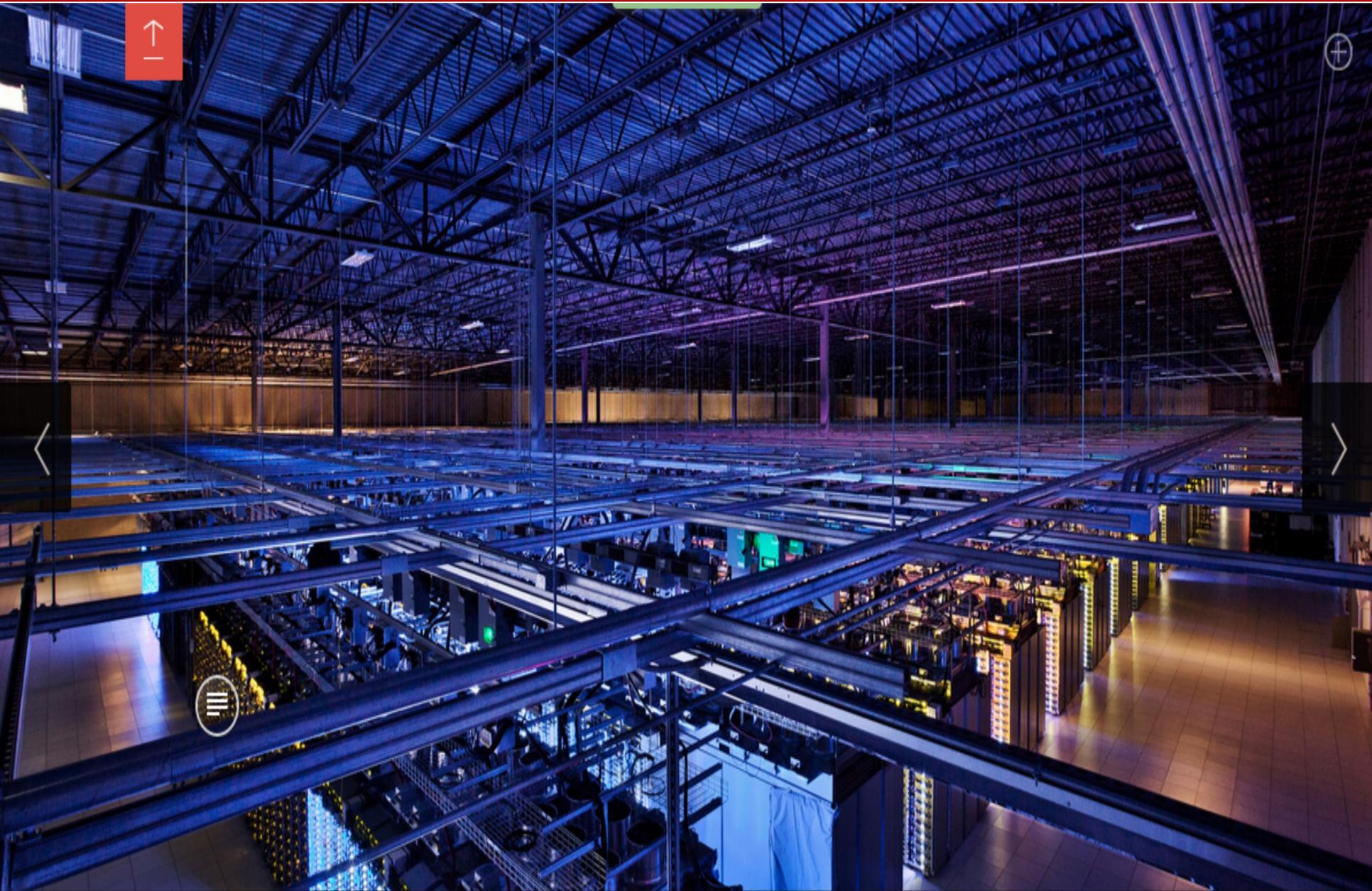
A large, empty search input field with a thin blue border. On the right side of the field, there is a small, colorful microphone icon, indicating voice search functionality.

Cerca con Google

Mi sento fortunato

Tutto Immagini Video Mappe Notizie | MSN Office Online Outlook.com English Accedi  

The Bing homepage features a scenic background image of a landscape with rolling hills and a dramatic, cloudy sky. In the foreground, the Bing logo is positioned to the left of a white search bar. The search bar has a magnifying glass icon on its right end. The navigation menu at the top includes links for "Tutto", "Immagini", "Video", "Mappe", "Notizie", "MSN", "Office Online", and "Outlook.com". On the right side of the menu, there are links for "English" and "Accedi" (Login), followed by a user profile icon and a menu icon.



Council Bluffs, Iowa

Il nostro data center di Council Bluffs si estende per 10.700 metri quadrati. Sfruttiamo in modo ottimale ogni



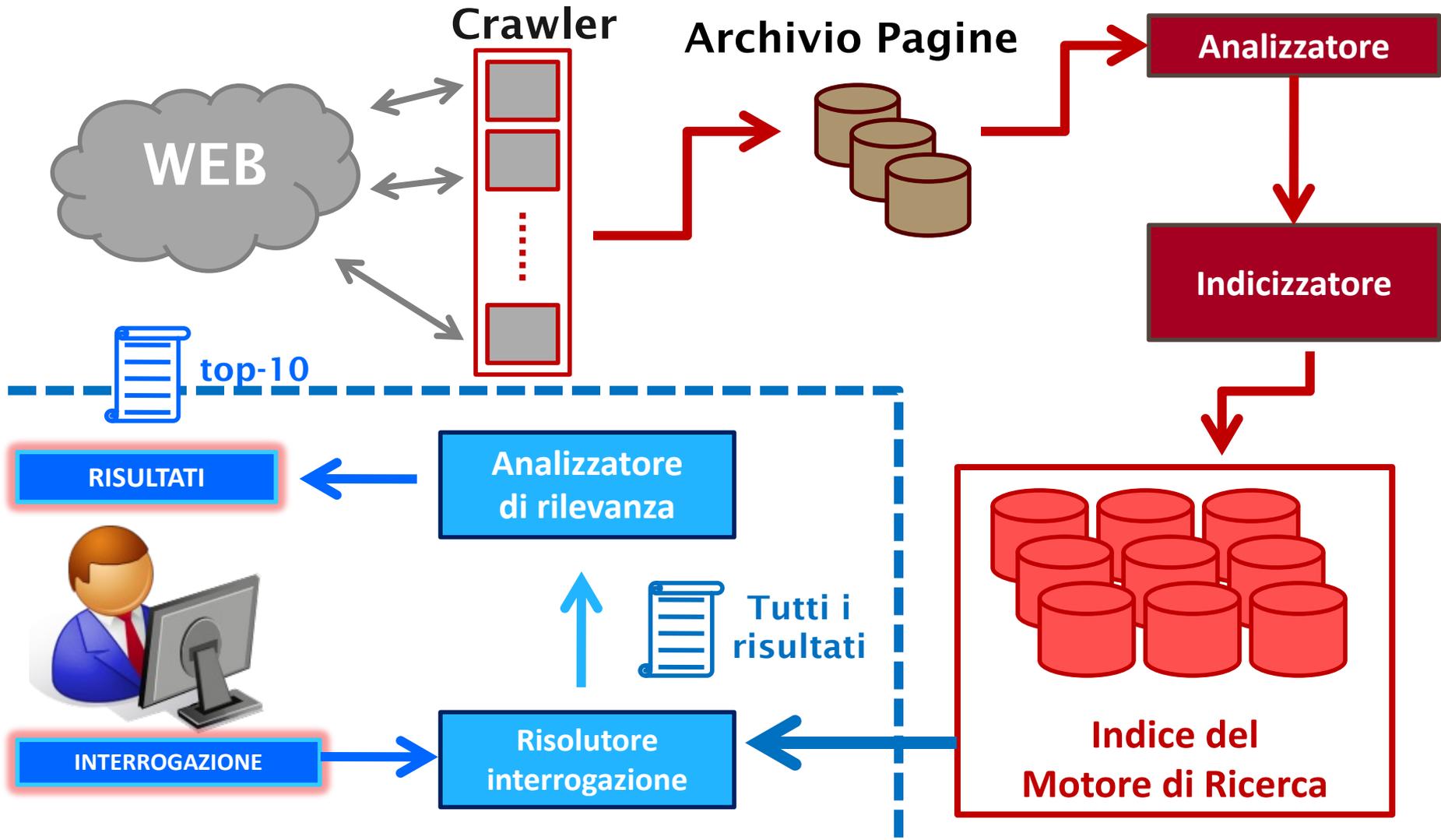
Condividi

+10m

Alcune domande per iniziare

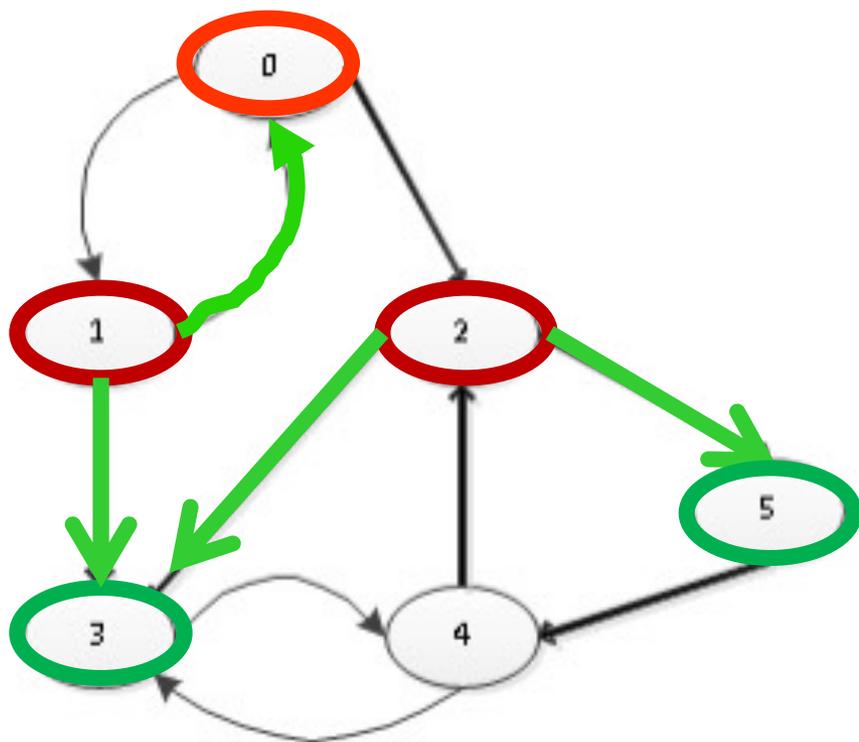
- Quando specificate una o più parole in una ricerca, Google/Bing «esplorano» il Web alla ricerca delle pagine che le contengono ?
- Dobbiamo specificare tutte le parole, anche articoli e preposizioni?
- Secondo quale «criterio di rilevanza» Google/Bing ordinano i risultati di una ricerca ?
- Cosa sono le «ricerche semantiche» eseguite oggi da Google/Bing ?

La struttura di un motore di ricerca



Crawling

- Quanto esplorare nel Web ?
- Quanto esplorare di un singolo sito ?
- Quanto spesso visitare una pagina ?



E' un processo 24/7,
svolto off-line

Il Crawler

1. **S = «pagine seme»**
2. Finché S non è vuoto
 - a) Sia P una pagina di S
 - b) Scarica P dal web
 - c) Analizza il contenuto di P
 - d) Inserisce i link di P in S,
ma solo se ...??...



AUTO BY TEL USA CANADA
Buy and insure new cars & trucks online

Car Buying & Car Insurance Pain Relief

LOW-COST

[Click here for advertising information - reach millions every month!](#)

Search and Display the Results

Search with Digital's Alta Vista [[Advanced Search](#)] [[Add URL](#)]



[Download free demo versions of AltaVista Technology software](#)

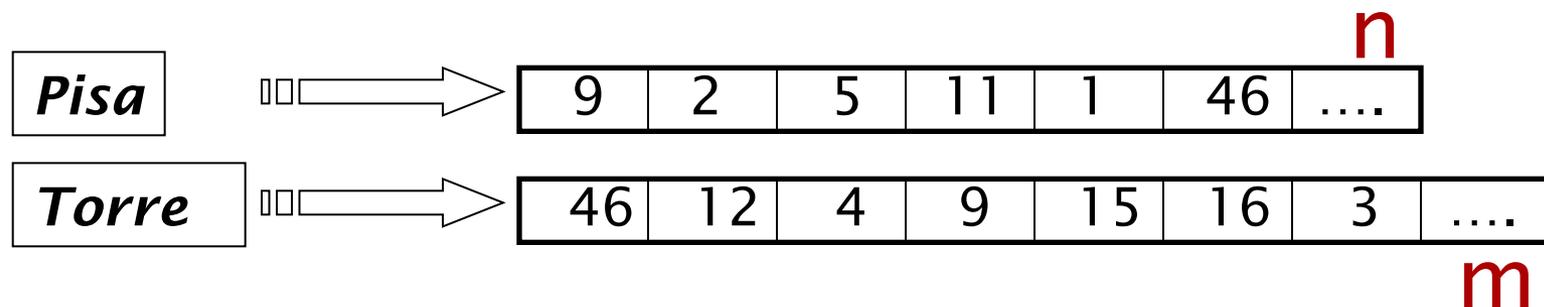


[\[Creative\]](#)[\[Search\]](#)[\[Humor\]](#)[\[Email\]](#)

Indicizzava solo il contenuto testuale delle pagine:

- Font, dimensione, posizione nella pagina [*proximity query*],....
- Frequenza delle parole nel documento ?

Il cuore dei motori di ricerca

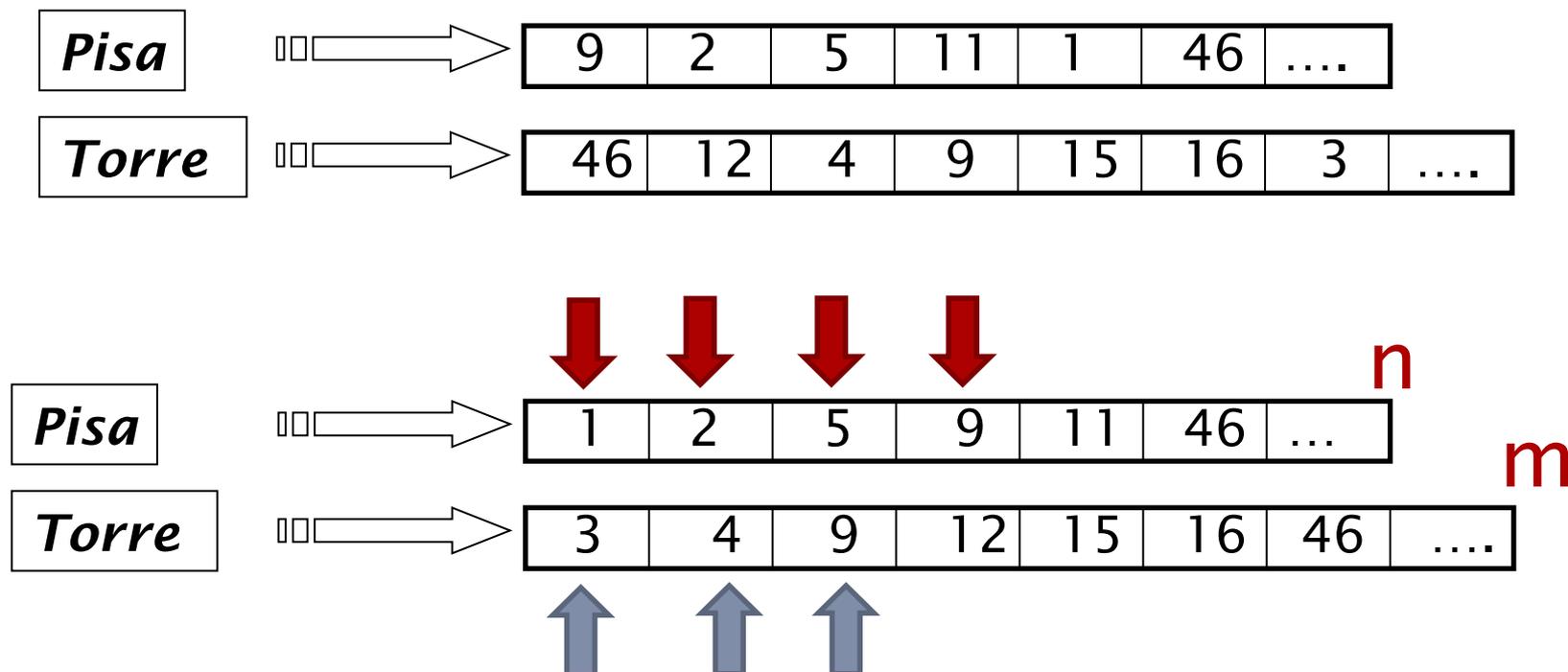


Sono eseguiti $n * m$ confronti (cfr)

Se $n, m \approx 10^6$, si eseguono $n * m \approx 10^{12}$ cfr

Se CPU esegue 10^9 cfr/sec, intersezione $\approx 10^3$ sec

Una soluzione più efficiente...



Se $n, m \approx 10^6$, si eseguono $n + m \approx 10^6$ cfr

Se CPU esegue 10^9 cfr/sec, intersezione $\approx 10^{-3}$ sec

Interrogazione per frase

Q = “pensiero computazionale”

- *pensiero:*

<2: 1, 17, 74, 222, 551>; <4: 8, 16, 190, 429, 433>; <7: 13, 23, 191>; ...

- *computazionale:*

<1: 17, 19>; <4: 11, 14, 17, 197, 291, 430, 435 >; <5: 14, 19, 101>; ...

Interrogazione per frase

Q = “pensiero computazionale”

- *pensiero:*

<2: 1, 17, 74, 222, 551>; <4: 8, 16, 190, 429, 431>; <7: 13,23,191>; ...

- *computazionale:*

<1: 17, 19>; <4: 11, 14, 17, 197, 291, 430, 435>; <5: 14,19,101>; ...

Quali sono i top-10 risultati ?

Possiamo usare solo la frequenza delle parole?

rango	forma	frequenza	rango	forma	frequenza
1	e	1752	16	ma	290
2	di	1338	17	i	283
3	che	1019	18	come	234
4	a	932	19	da	233
5	il	925	20	io	225
6	la	711	21	mi	219
7	un	708	22	le	211
8	non	507	23	più	210
9	per	481	24	l'	206
10	in	453	25	disse	202
11	Pinocchio	415	26	lo	199
12	si	393	27	burattino	195
13	gli	364	28	se	189
14	una	360	29	con	188
15	è	296	30	era	185

Le 30 parole più frequenti in *Pinocchio*

TF-IDF score: funzione della frequenza di una parola **in** e **tra** docs

Una famosa misura di rilevanza

$$\text{rilevanza}(t, d) = \text{freq}(t, d) \times \text{rarietàColl}(t)$$

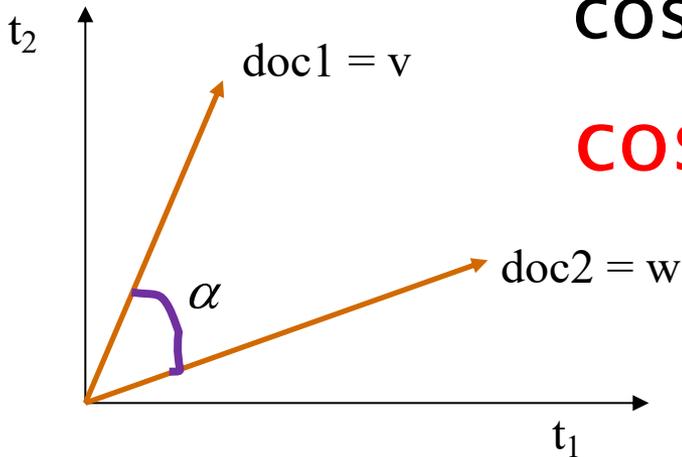
- Rilevanza('e' , «Pinocchio») \rightarrow (freq = 1752) * (rarietà \cong 0)
- Rilevanza('burattino' , «Pinocchio») \rightarrow (freq = 195) * (rarietà \cong 1)
- Ogni documento **d** (p.e. «Pinocchio») è rappresentato quindi con un vettore di pesi di rilevanza, uno per ogni possibile termine **t**

abaco, casa, gatto, computer, ...

Pinocchio = [0.0, 0.8, 0.9, 0.0,]

Siamo in uno **Spazio Vettoriale**
con milioni di dimensioni, tante quanti sono i termini del dizionario

Similarità del Coseno



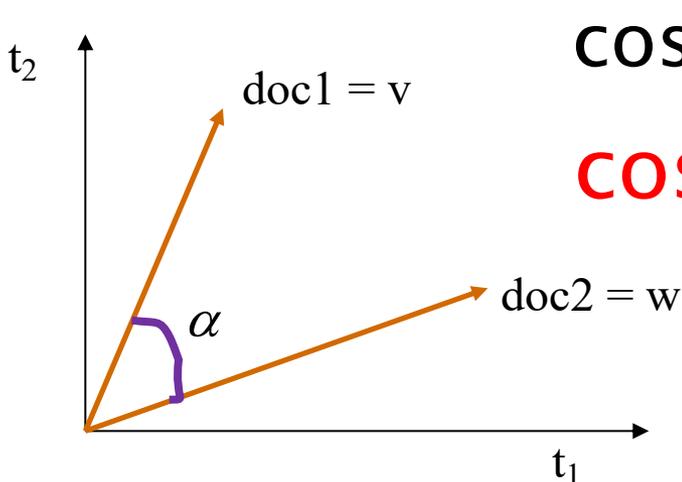
$\cos(\alpha) \in [0 \text{ (dissimile)}, 1 \text{ (simile)}]$

$$\cos(\alpha) = v \cdot w / \|v\| * \|w\|$$

	Doc1 = v	Doc2 = w
term 1	2	4
term 2	3	1

$$\cos(\alpha) = 2*4 + 0*0 + 3*1 / \sqrt{2^2 + 3^2} * \sqrt{4^2 + 1^2} \cong 0,75 \rightarrow 40^\circ$$

Similarità del Coseno



$\cos(\alpha) \in [0 \text{ (dissimile)}, 1 \text{ (simile)}]$

$$\cos(\alpha) = v \cdot w / \|v\| * \|w\|$$

Problema computazionale

Non si può confrontare il vettore della query con i vettori di tutti i documenti, richiederebbe troppo tempo.



The heart of the Elastic Stack

Elasticsearch is a distributed, RESTful search and analytics engine capable of addressing a growing number of use cases. As the heart of the Elastic Stack, it centrally stores your data so you can discover the expected and uncover the unexpected.

[Start free trial](#)

SPAM results

Pagina personale di Paolo Ferragina

Mi chiamo Paolo Ferragina, e vivo a Pisa.

Pagina personale di Paolo Ferragina

calcio, calcio, calcio, calcio, calcio, calcio,
calcio, calcio, calcio, calcio, calcio, calcio,...

Mi chiamo Paolo Ferragina, e vivo a Pisa.

Query: calcio



Stavano svolgendo un PhD in Computer Science a Stanford

Google!

Search the web using Google!

10 results



Google Search

I'm feeling lucky

Index contains ~25 million pages (soon to be much bigger)

[About Google!](#)

[Stanford Search](#) [Linux Search](#)

Sfruttare i testi àncora e la struttura del grafo del Web



Web Images Groups News Froogle Local more »

miserable failure

Search

Advanced Search Preferences



Web Results 1 - 10 of about 969,000 for miserable failure. (0.06 seconds)

Biography of President George W. Bush

Biography of the president from the official White House web site.
www.whitehouse.gov/president/gwbbio.html - 29k - [Cached](#) - [Similar pages](#)
[Past Presidents](#) - [Kids Only](#) - [Current News](#) - [President](#)
[More results from www.whitehouse.gov »](#)

GOOGLE bombing

Welcome to MichaelMoore.com!

Official site of the gadfly of corporations and the television show The Awful Truth
www.michaelmoore.com/ - 35k - Sep 1

BBC NEWS | Americas | 'Miserable failure'

Web users manipulate a popular search engine to the president's page.
news.bbc.co.uk/2/hi/americas/3298443

Google's (and Inktomi's) 'Miserable failure'

A search for **miserable failure** on Google returns a Bush biography from the US White House.
searchenginewatch.com/sereport/article.php?id=682 x 472

Google quanto è alto un nano?

Web Immagini Video Shopping Notizie Altro Strumenti di ricerca

Circa 1.870.000 risultati (0,44 secondi)

I cookie ci aiutano a fornire i nostri servizi. Utilizzando tali servizi, accetti l'utilizzo dei cookie da parte nostra. [Informazioni](#) [OK](#)

1,65 m
Silvio Berlusconi, Altezza



Feedback

Notizie relative a quanto è alto un nano?

"Quanto è alto un nano?", Google risponde Berlusconi (FOTO)
Blitz quotidiano - di Filippo Limoncelli - 1 ora fa
ROMA – Sui social network impazza il risultato grafico della domanda "quanto è alto un nano" scritta su Google. Il motore di ricerca, infatti, ...

Altre notizie su **quanto è alto un nano?**

Quanto è alto un nano? E su Google spunta Berlusconi ...
www.giornalettismo.com/.../chiedi-a-google-quanto-e-alto-un-n...

Silvio Berlusconi

Ex Ministri degli affari esteri della Repubblica Italiana
Silvio Berlusconi è un politico e imprenditore italiano, conosciuto anche come "il Cavaliere" in ragione dell'onorificenza a cavaliere del lavoro cui ha rinunciato nel 2014, e conferitagli nel 1977 dal presidente della Repubblica Giovanni Leone. [Wikipedia](#)
Data di nascita: 29 settembre 1936 (età 77), Milano
Altezza: 1,65 m
Partner: Francesca Pascale (2012-)
Figli: Barbara Berlusconi, Marina Berlusconi, Pier Silvio Berlusconi, Eleonora Berlusconi, Luigi Berlusconi
Coniuge: Veronica Lario (s. 1990-2013), Carla Elvira Lucia Dall'Oglio (s. 1965-1985)
Genitori: Luigi Berlusconi, Rosa Bossi

Post recenti



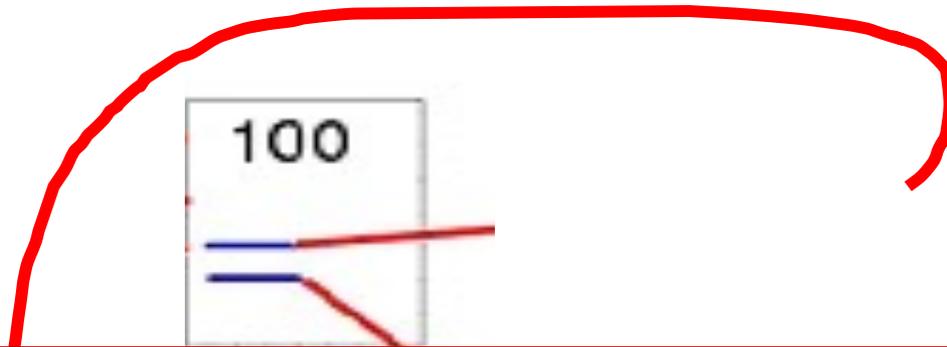
La sinistra, perseguendo la teoria gramsciana delle cosiddette case matte del potere, ha conquistato il mondo della cultura. Ha messo i suoi uomini nella scuola, ... 7 mag 2014

Ricerche correlate



Il (classico) PageRank

SPAM → SEO ?



Sorta di «misura di centralità» di un nodo nel grafo

*... oggi la «rilevanza» di una pagina è calcolata con tecniche di *AI* basate su centinaia di parametri (feature)... tra queste sicuramente varianti di TF-IDF e PageRank*

Definizione *«ricorsiva»*: la «rilevanza» di una pagina dipende dalla «rilevanza» delle pagine che puntano a essa nel grafo del Web

Rilevanza di un documento

È un concetto matematicamente **non ben definibile**,
essendo innanzitutto soggettivo e mutevole nel tempo

Per ogni pagina si calcolano
una serie di **caratteristiche**:

- TF-IDF delle parole
- PageRank
- Loro vicinanza
- Se nel titolo o URL
- Posizione nella pagina
- ...

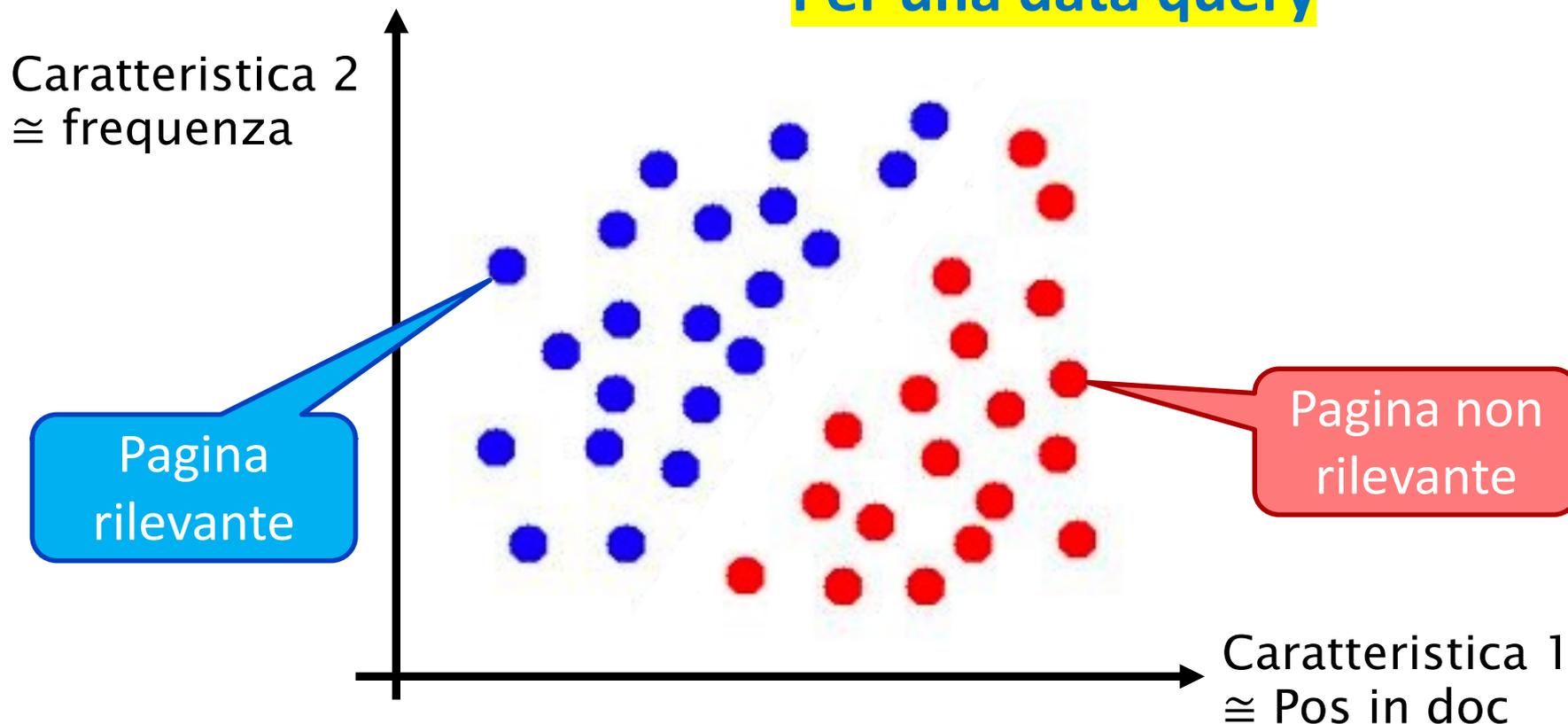
sono oltre 200 «misure»...

The image shows a Google search interface for the query "torre pisa". The search bar at the top contains the text "torre pisa" and the Google logo. Below the search bar, there are tabs for "Tutti", "Immagini", "Maps", "Notizie", "Video", "Altro", "Impostazioni", and "Strumenti". The search results section shows "Circa 3.860.000 risultati (0,83 secondi)". The first result is "Torre di Pisa - Wikipedia" with a snippet: "La torre di Pisa (popolarmente torre pendente e, a Pisa, la Torre) è il campanile della cattedrale di Santa Maria Assunta, nella celeberrima piazza del Duomo di ...". Other results include "TORRE DI PISA" from torrepisa.com and "La Torre di Pisa pende e ruota - Il Sole 24 ORE". On the right side, there is a knowledge panel for "Torre di Pisa" with a 4.5-star rating, 5,317 reviews, and a map showing the location in Piazza del Duomo, Pisa. The panel also includes details like "Edificio a Pisa, Italia", "Inirizzo: Piazza del Duomo, 56126 Pisa PI", "Altezza: 57 m", and "Costruzione iniziata: agosto 1173".

Rilevanza di una Pagina

Legata oggi alle tecniche di **AI e Machine Learning**

Per una data query



L'ultimo passo recente

il paparazzo ha fotografato una stella del cinema

L' astronomo ha fotografato una stella

Giorgio sta usando il browser di Microsoft

Luca è un fan di Internet Explorer



Leonardo da Vinci



Leonardo di ser Piero da Vinci was an Italian Renaissance polymath: painter, sculptor, architect, musician, scientist, mathematician, engineer, inventor, anatomist, geologist, cartographer, botanist, and writer.

Read more on en.wikipedia.org

Born: April 15, 1453, Anchiano
Died: May 2, 1519, Clos Lucé
Buried: St. Florentin's Church
Inventions: Viola organista, Double hull
Parents: Caterina da Vinci, Piero da Vinci

Maggio 2012



The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.



See it in action

Discover answers to questions you never thought to ask, powered by the Knowledge Graph.

Circa 56.200.000 risultati (0,49 secondi)

[Home | Galileo - Giornale di Scienza](#)

www.galileonet.it/ ▾

News, magazine, recensioni e dossier monografici sui temi della scienza. Il primo magazine Online italiano su scienza e problemi globali. On line dal 1996.

[Galileo Galilei - Wikipedia](#)

it.wikipedia.org/wiki/Galileo_Galilei ▾

Galileo Galilei (Pisa, 15 febbraio 1564 – Arcetri, 8 gennaio 1642) è stato un fisico, filosofo, astronomo e matematico italiano, considerato il padre della scienza ...
[Processo a Galileo Galilei](#) - [Categoria:Galileo Galilei](#) - [Casa di Galileo Galilei](#)

[Galileo Galilei - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Galileo_Galilei ▾ [Traduci questa pagina](#)

Galileo Galilei often known mononymously as **Galileo**, was an Italian physicist, mathematician, engineer, astronomer, and philosopher who played a major role ...

[Aeroporto Galileo Galilei - Sito ufficiale - Aeroporto di Pisa ...](#)

www.pisa-airport.com/ ▾

Aeroporto Internazionale **Galileo Galilei**. Include le informazioni, gli orari dei voli, le infrastrutture.

Hai visitato questa pagina molte volte. Ultima visita: 05/09/14

[Liceo Classico Galilei Pisa](#)

www.lcgalilei.pisa.it/ ▾

Liceo Classico **Galileo Galilei** - Pisa. Il futuro ha un cuore antico. Home; Registro

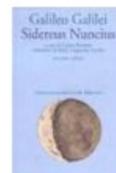


Altre immagini

Galileo Galilei

Fisico

Libri



Sidereus Nuncius
1610



Dialogo sopra i due massimi s...
1632



Discorsi e dimostraz... matematici...
1638



Il Saggiatore
1623



Lettera a Madama Cristina d...
1636

Ricerche correlate

Visualizza altri 15 elementi



Niccolò Copernico



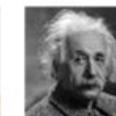
Isaac Newton



Giovanni Keplero



Aristotele



Albert Einstein

Il Grafo di Wikipedia

WIKIPEDIA L'enciclopedia libera

Voce **Diego Armando Maradona**

Da Wikipedia, l'enciclopedia libera.

Disambiguazione – "Maradona" rimanda qui. Se stai cercando altri significati, vedi *Maradona (disambigua)*.

Diego Armando Maradona (Lands, 30 ottobre 1960) è un allenatore di calcio, dirigente sportivo ed ex calciatore argentino, di ruolo centrocampista offensivo, capitano della Nazionale argentina di calcio vincitrice del Mondiale del 1986.

Noto anche come *El Pibe de Oro* (*il Ragazzo d'Oro*), è considerato uno dei più grandi calciatori di tutti i tempi^[R] e, da molti, il migliore in assoluto.^[R] In una carriera da professionista più che ventennale, ha militato nell'Argentinos Juniors, nel Boca Juniors, nel Barcellona, nel Napoli, nel Siviglia e nei Newell's Old Boys. Con la Nazionale argentina ha partecipato a quattro edizioni dei Mondiali (1982, 1986, 1990 e 1994): 01 incontri disputati e la 34^a nel realizzate in Nazionale costituiscono due record, successivamente battuti.^[R] Il suo gol realizzato contro la Nazionale inglese nei quarti di finale del Mondiale 1986 è considerato il gol del secolo, e segue di cinque minuti l'altro famoso e controverso episodio per cui si spesso ricordati, quello della *mano di Dio*.

Non è mai potuto entrare nella graduatoria del Pallone d'oro, perché fino al 1995 il premio era riservato solo ai giocatori europei. Proprio per questo nel 1995 vinse il Pallone d'oro alla carriera.

L'11 dicembre 2000, Maradona ha inoltre ricevuto il premio ufficiale FIFA come "Miglior



Maradona al Napoli nel 1986

Nazionalità Argentina

(FR) Fédération Internationale de Football Association

FIFA

For the Game. For the World.

Discipline	Calcio
	Calcio a 5
	Beach soccer
Fondazione	1904
Giurisdizione	Mondiale
Federazioni affiliate	211
Confederazione	CIO ASOIF
Sede	 Zurigo
Presidente	 Gianni Infantino
Motto	<i>For the Game. For the World.</i> ^[1]
Sito ufficiale	www.fifa.com [ⓘ]

Modifica dati su Wikidata - Manuale

Boca Juniors
Calcio



Xeneizes (Genovesi),
La mitad más uno (la metà più uno)

Segni distintivi
Uniformi di gara



Casa



Trasferta

Colori sociali **Giallo e blu**

Inno *La marcha de Boca Juniors*
Victoriano "Toto" Caffarena

A.C. Milan
Calcio



Rossoneri, Diavolo^[1]

Segni distintivi
Uniformi di gara



Casa



Trasferta



Terza divisa

Colori sociali **Rosso e nero**

Simboli **Diavolo**

Inno *Milan, Milan*
Tony Renis e Massimo Guantini^[2]

Dati societari

Argentina




Uniformi di gara



Casa

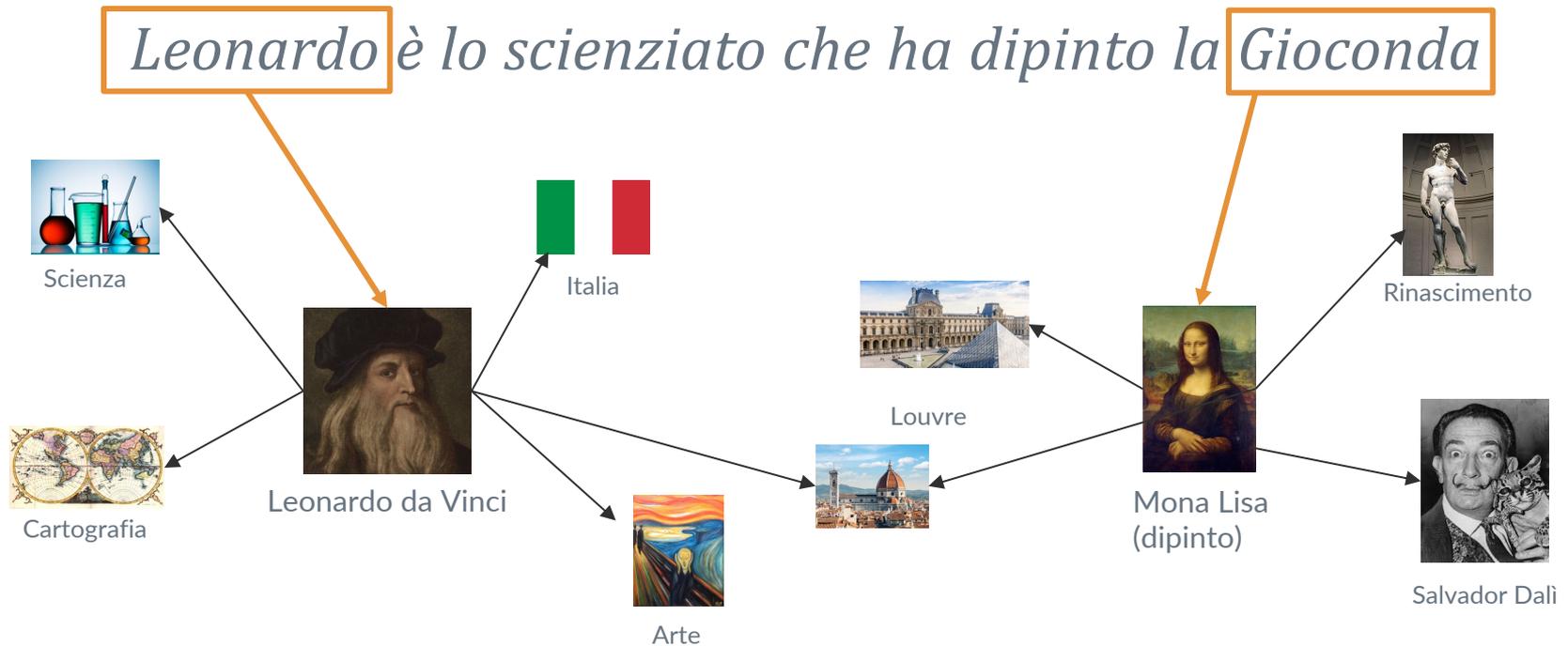


Trasferta



Dalle parole ai concetti

Leonardo è lo scienziato che ha dipinto la Gioconda



Query: surrealismo

Surrealismo

Da Wikipedia, l'enciclopedia libera.

 **Questa voce o sezione sull'argomento movimenti artistici non cita le fonti**
Puoi migliorare questa voce aggiungendo citazioni da fonti attendibili secondo le linee gu

Il **surrealismo** è un movimento artistico e letterario d'avanguardia del Novecento, nato negli anni 20 a Parigi con coinvolse tutte le arti, toccando anche letteratura e cinema; nel 1924 ne fu scritto il primo manifesto.^[1] Esso vuole fatta di irrazionale e di sogno, per rivelare gli aspetti più profondi della psiche.

Il surrealismo ebbe come principale teorico il poeta André Breton, che impersonò la vitalità distruttiva del dadaismo lettura de *L'Interpretazione dei sogni* di Freud del 1900; dopo averlo letto arrivò alla conclusione che fosse inace l'inconscio avessero avuto così poco spazio nella civiltà moderna, e pensò quindi di fondare un nuovo movimnto avessero un ruolo fondamentale. Nacque così il surrealismo, che aveva avuto tra i suoi precursori recenti il poeta morto nel 1918.



Service Overview

TagMe API

Demo

WAT Api

SWAT Api

The Entity Linking tools by Acube lab

Welcome to the Tagme Virtual Research Environment. From here, you can access all Entity Linking tools provided by the [Acube lab](#) at the University of Pisa.



Entity linker, ideal for annotating noisy text.

Available languages: **en, de, it**



Entity linker, ideal for annotating well-formed text.

More accurate than TagMe, but still experimental.

Available languages: **en**



Entity linker for web search queries.

Available languages: **en**



Entity Salience service: assigns a relevance score to the entities mentioned by a document.

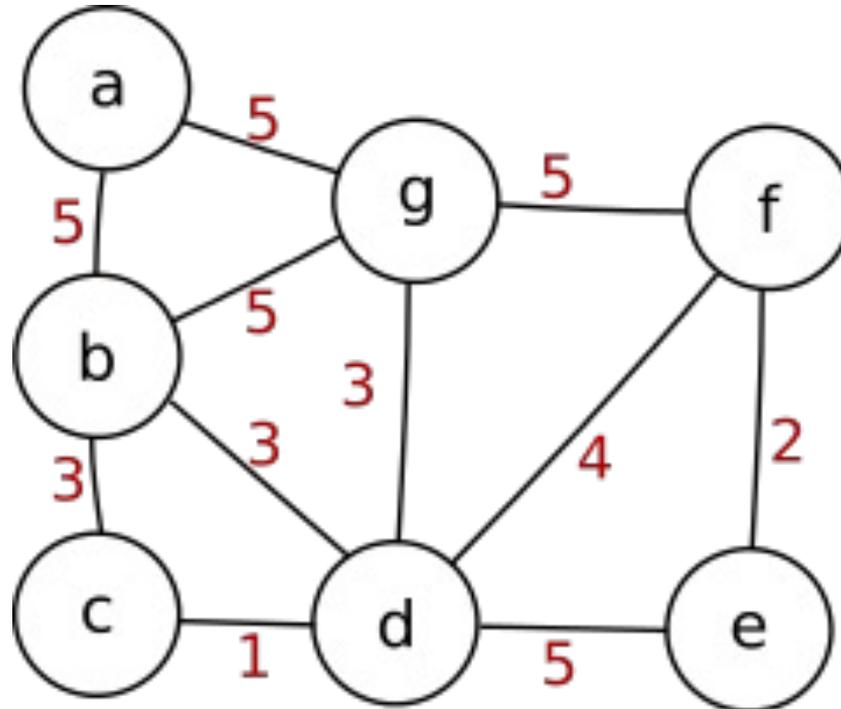
Available languages: **en**

Credits

Current and former members of this lab, who contributed to the development and deployment of these services, include **Paolo Ferragina, Marco Cornolti, Francesco Piccinno, Marco Ponza, Ugo Scaiella, Daniele Vitale**.

<https://sobigdata.d4science.org/web/tagme/>

Grafi: una struttura dati potente



Consigliati in base ai tuoi interessi



Simili a prodotti che hai già visto [Visualizza altro](#)



Altri prodotti da tenere presente [Visualizza altro](#)



Take-away msg #1

- Esistono motori di ricerca per ogni tipo di dato
- Riescono a scalare a milioni di documenti anche nella loro versione open-source



elasticsearch

- Sono diventati molto sofisticati e basati non solo sulla «keyword search»
- Estraggono loro stessi «keywords» che quindi non devono necessariamente essere assegnate a mano



Take-away msg #2

- La rappresentazione tabellare è insufficiente per espressività, efficacia ed efficienza dell'elaborazione
- I dati devono essere strutturati in modo interconnesso, con proprietà sui nodi e sugli archi
- Tanto più il grafo è di «qualità» quanto più efficace sarà l'estrazione dei dati
- Varianti del PageRank sono comunemente usate per definire la «rilevanza» di nodi e quindi utenti, prodotti, documenti,
- Abbiamo già forme di «reasoning» automatico

