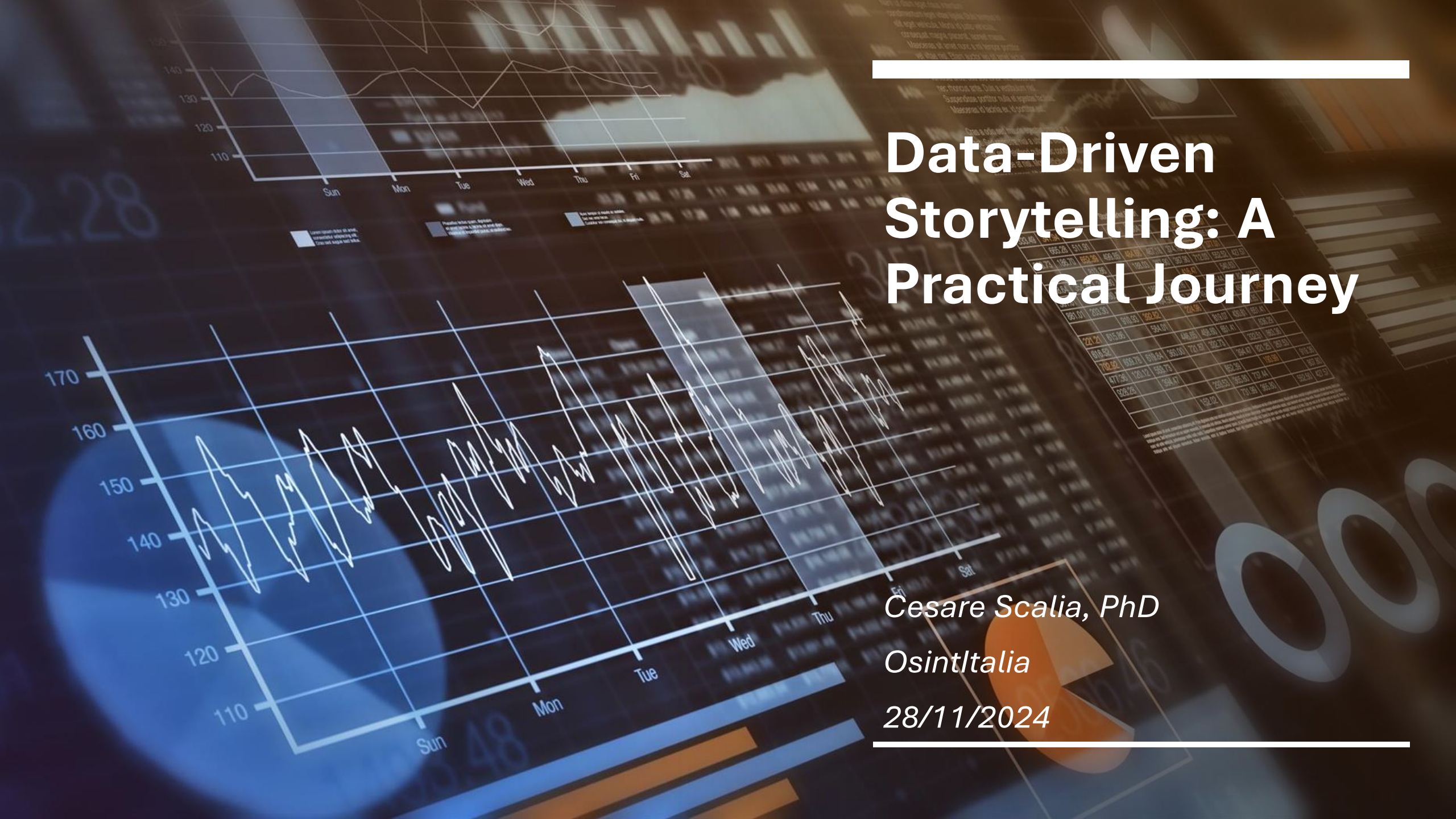


Data-Driven Storytelling: A Practical Journey

Cesare Scalia, PhD

OsintItalia

28/11/2024

Introduzione

Cos'è la data science

La Data Science si concentra sull'estrazione di conoscenza dai dati grezzi usando tecniche scientifiche, statistiche e informatiche, al fine di fornire decisioni migliori e prevedere comportamenti futuri.

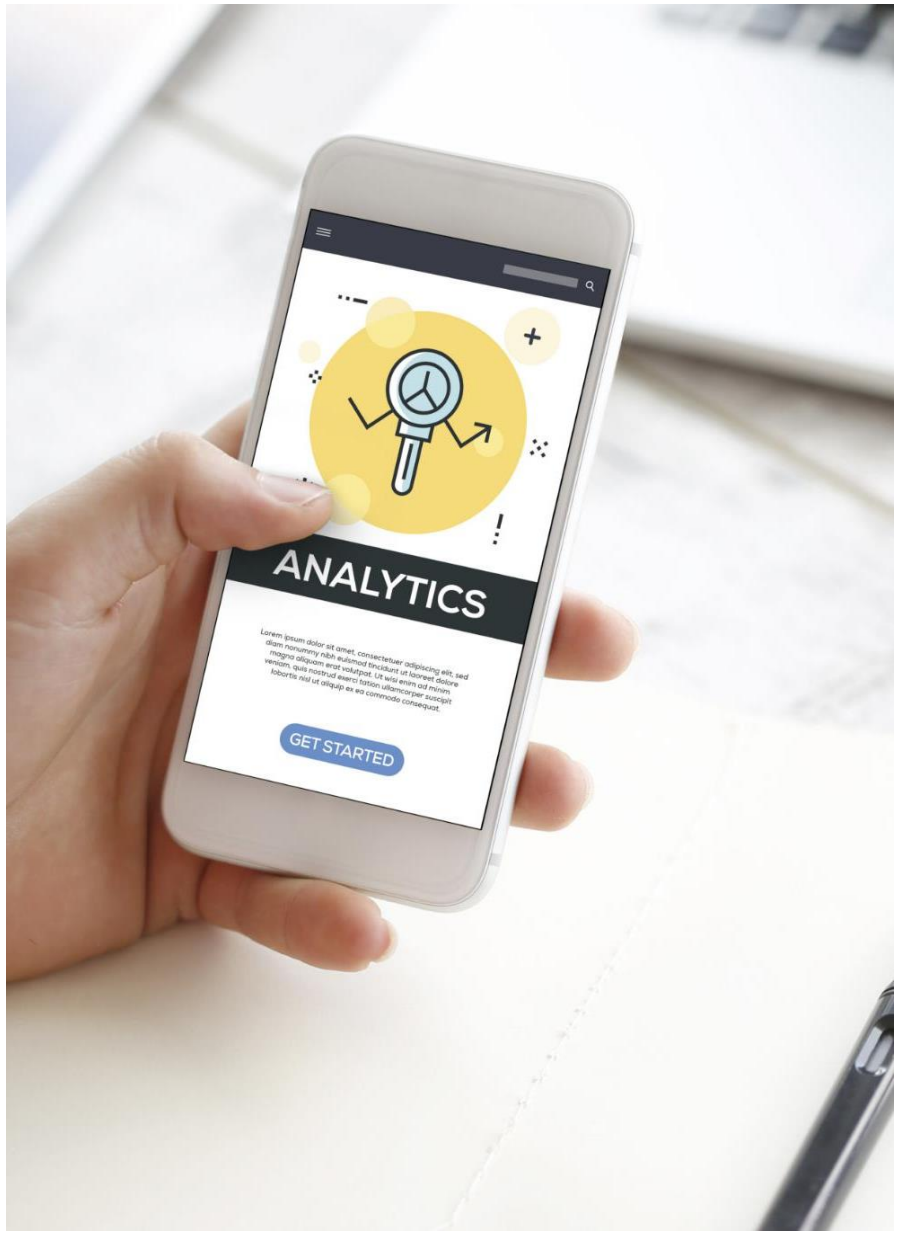
Python

Python è uno degli strumenti più importanti nella data science grazie alla sua flessibilità, semplicità e potenza. Con Python puoi analizzare dati, creare modelli di machine learning e visualizzare informazioni in modo chiaro e accattivante.

Le **librerie** sono insiemi di strumenti pronti all'uso che semplificano attività complesse, come la manipolazione di dati o la creazione di grafici.

- **Pandas:** libreria Python utilizzata per la manipolazione e l'analisi di dati strutturati. Sostituisce in pratica excel
- **NumPy:** utilizzata per la manipolazione di array numerici. È ampiamente utilizzata nella data science per la matematica computazionale, il machine learning e la visualizzazione dei dati.
- **Matplotlib:** per la visualizzazione dei dati. È ampiamente utilizzata nella data science per creare grafici e visualizzazioni di alta qualità per la presentazione dei dati.

Nel workshop però vedremo anche librerie avanzate per interazione con pagine web, machine learning e deep learning avanzato fino ad AI generativa



Applicazioni pratiche

Open Source Intelligence (OSINT)

La data science può essere utilizzata per analizzare fonti open source, come social media, immagini satellitari e documenti pubblici, per scoprire connessioni, tracciare movimenti e identificare schemi. Ad esempio, analisi di network e tecniche di text mining possono essere usate per monitorare la diffusione di disinformazione o mappare le catene di approvvigionamento globali.

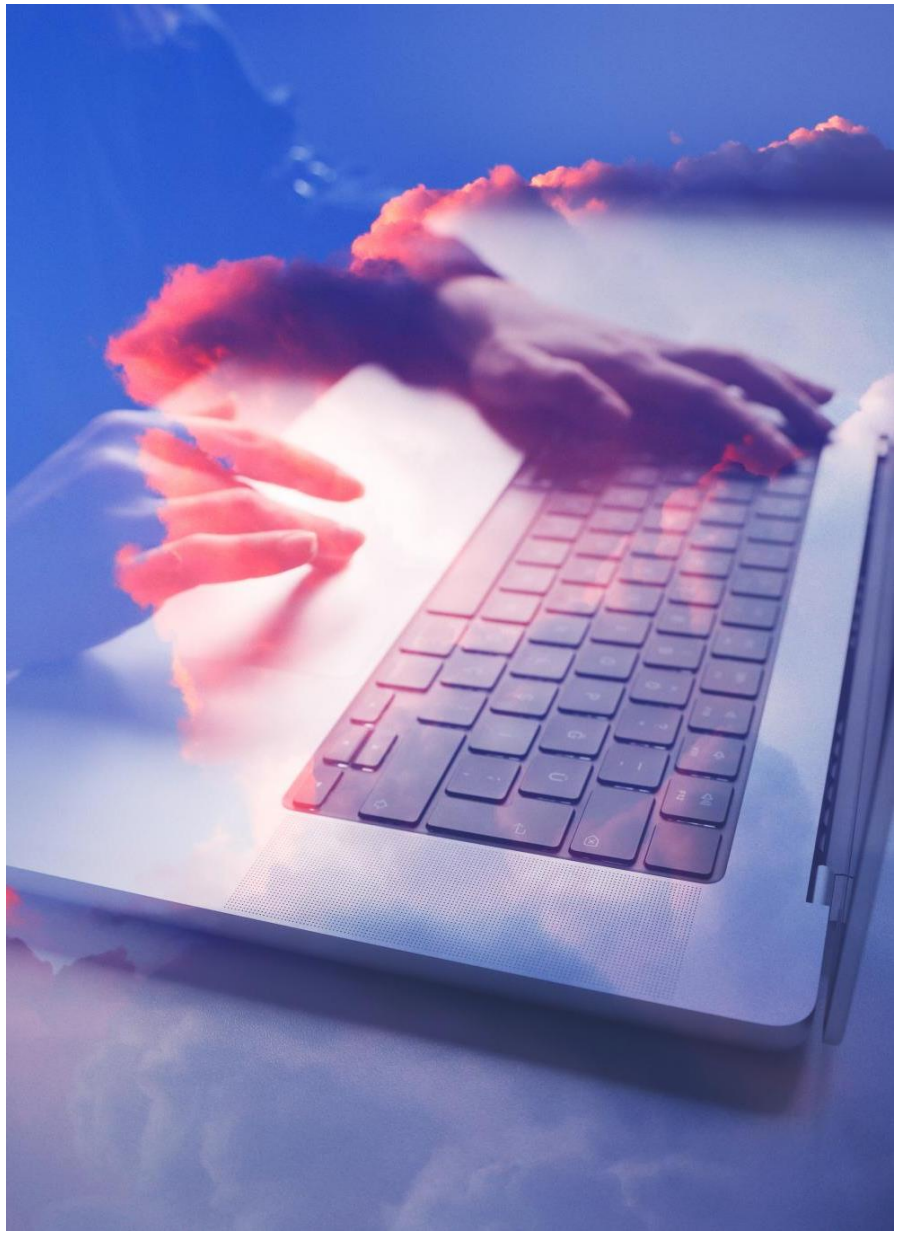
Giornalismo

La data science può aiutare i giornalisti a indagare su storie complesse analizzando grandi quantità di dati. Ad esempio, nei Panama Papers, tecniche di visualizzazione e analisi di grafi sono state usate per rivelare connessioni tra società offshore e personaggi influenti. Inoltre, strumenti come sentiment analysis e analisi di network possono aiutare a comprendere l'opinione pubblica su determinati temi.

Investigazioni

La data science può supportare investigazioni su crimini, frodi o violazioni dei diritti umani, analizzando grandi dataset per identificare anomalie o modelli nascosti. Organizzazioni come Bellingcat hanno usato immagini satellitari e dati pubblici per investigare su crimini di guerra, mentre tecniche di machine learning sono impiegate per riconoscere schemi o eventi specifici in video e foto.

Analisi da Open Data



Open Data: Trasparenza e Innovazione

Open Data

Iniziative governative e istituzionali per rendere i dati pubblici, liberamente accessibili e riutilizzabili da chiunque per ricerca, analisi e innovazione.

- Dati.gov.it: il portale nazionale raccoglie dataset su demografia, economia, trasporti, ambiente, salute ecc
- EU Open Data Portal: raccoglie dati da istituzioni e agenzie dell'Unione Europea su economia, ambiente, energia, innovazione ecc
- Data.gov: il portale statunitense aggrega dati pubblici su energia, salute, educazione, trasporti, e clima.

Banche dati aperte

Le banche dati aperte sono raccolte di dati gestiti da organizzazioni o istituzioni che sono disponibili al pubblico per l'uso e la condivisione

- World Bank Data: indicatori economici e sociali globali su povertà, sviluppo, salute, energia, istruzione e altro.
- UN Data: dati raccolti dalle agenzie delle Nazioni Unite su popolazione, sviluppo sostenibile, economia e più.
- OECD Data: statistiche e analisi economiche, sociali e ambientali dei paesi membri dell'OCSE.

Wikipedia

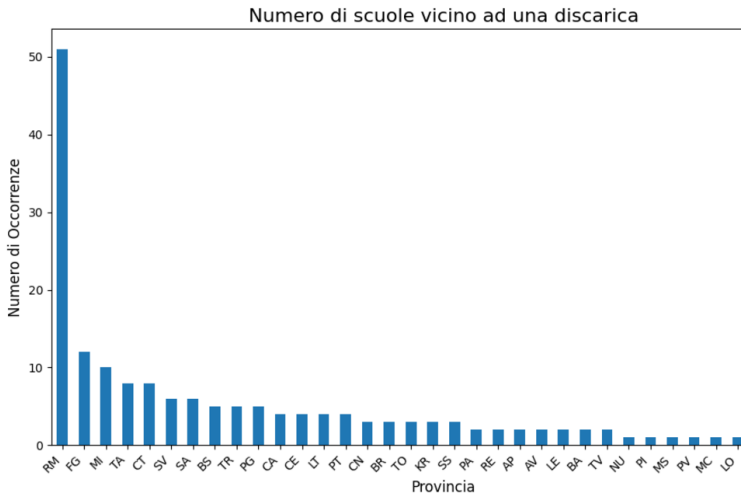
Anche se non strutturata in modo tabellare wikipedia è un dato a licenza aperta ed accesso pubblico molto utile per analizzare contenuti informativi o trend sociali (es analisi delle visite delle pagine

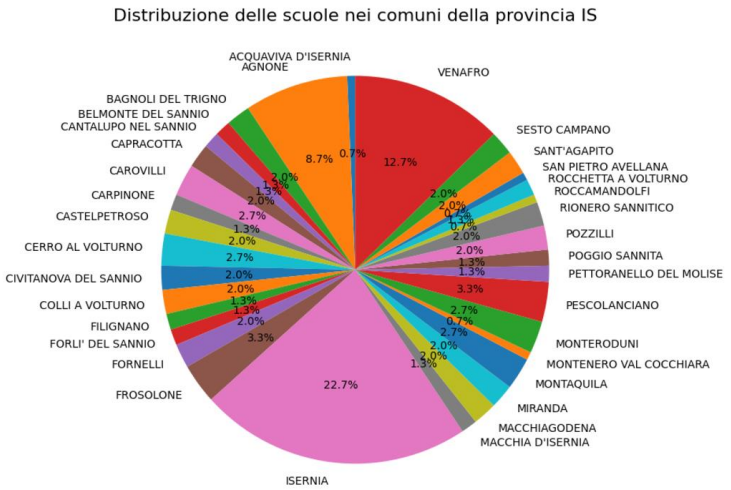
Dove Trovare i Dati in Italia

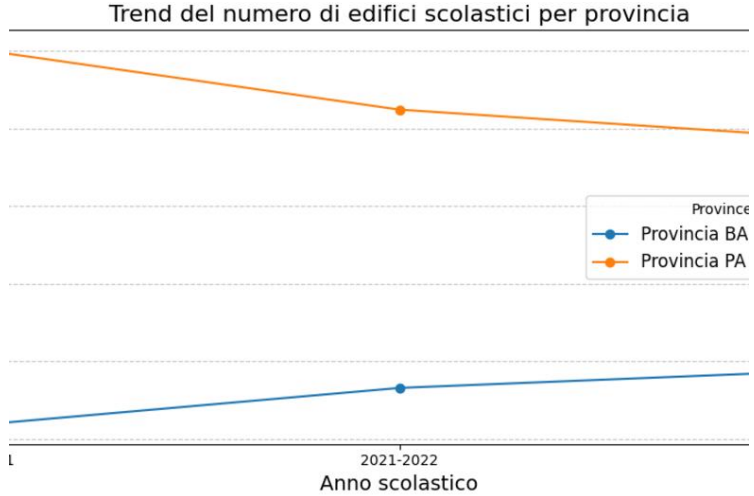
In Italia, l'apertura dei dati pubblici è regolamentata dal **Codice dell'Amministrazione Digitale** (CAD) e promossa dall'**Agenzia per l'Italia Digitale** (AgID). Le amministrazioni sono obbligate a pubblicare dati per favorire trasparenza, riuso dei dati e supportare ricerca e innovazione

Esempio fonti dati italiane:

- ISTAT: dati statistici su popolazione, economia, ambiente, ecc.
- Dati.gov.it: portale nazionale degli open data pubblici.
- OpenCoesione: dati sui progetti finanziati con fondi pubblici.
- Ministero della Salute: dataset su sanità, salute pubblica e prevenzione.
- Ministero dell'Economia e delle Finanze: dati su fisco, immobili e tributi.
- Geoportale Nazionale: dati territoriali e ambientali.
- Portale OpenData Piemonte: esempi regionali di open data.







Esempio: Open Data Scuola

Vediamo quali insight ed informazioni possiamo estrarre analizzando l'Open Data del Ministero dell'Istruzione relativo all'Edilizia Scolastica

Web Scraping per la Raccolta di Dati

Web scraping

Tecnica di estrazione dati utilizzata per recuperare informazioni da un sito web. È spesso utilizzato per raccogliere grandi quantità di dati in modo efficiente.

Come funziona il Web Scraping

Il web scraping utilizza programmi software per analizzare il codice HTML di una pagina web e trovare le informazioni desiderate. Le informazioni vengono quindi estratte e salvate in un formato utilizzabile.

Sfide del Web Scraping

Il web scraping può essere difficile a causa delle politiche di sicurezza dei siti web, delle limitazioni di velocità e della complessità del codice HTML. Inoltre, è importante assicurarsi di avere il permesso di estrarre i dati da un sito web.

NB Utilizzare le informazioni raccolte responsabilmente è fondamentale per rispettare la privacy e la proprietà intellettuale.

```python
from usp.tree import sitemap_tree_for_homepage
tree = sitemap_tree_for_homepage("https://lefolliedinaruto.wordpress.com/category/manga-naruto/it-IT.sitemap.xml")
```

Librerie e metodologie

```python
import trafilatura
from tqdm import tqdm

# Funzione per estrarre il testo con Trafilatura
def extract_text(url):
    downloaded = trafilatura.fetch_url(url)
    if downloaded:
        return trafilatura.extract(downloaded)
    else:
        return None  # Nessun testo disponibile


# Configura tqdm per visualizzare il progresso
tqdm.pandas(desc="Processing URLs")

# Applica la funzione di estrazione testo a ogni URL con barra di avanzamento
df['text'] = df['url'].progress_apply(extract_text)

# Visualizza il risultato
display(df.head())
```

Trafilatura Trafilatura è una libreria semplice per l'estrazione del testo HTML dentro un url. Semplice, efficiente pero non permette di estrarre informazioni da pagine piu complesse

Ultimate_sitemap_parser

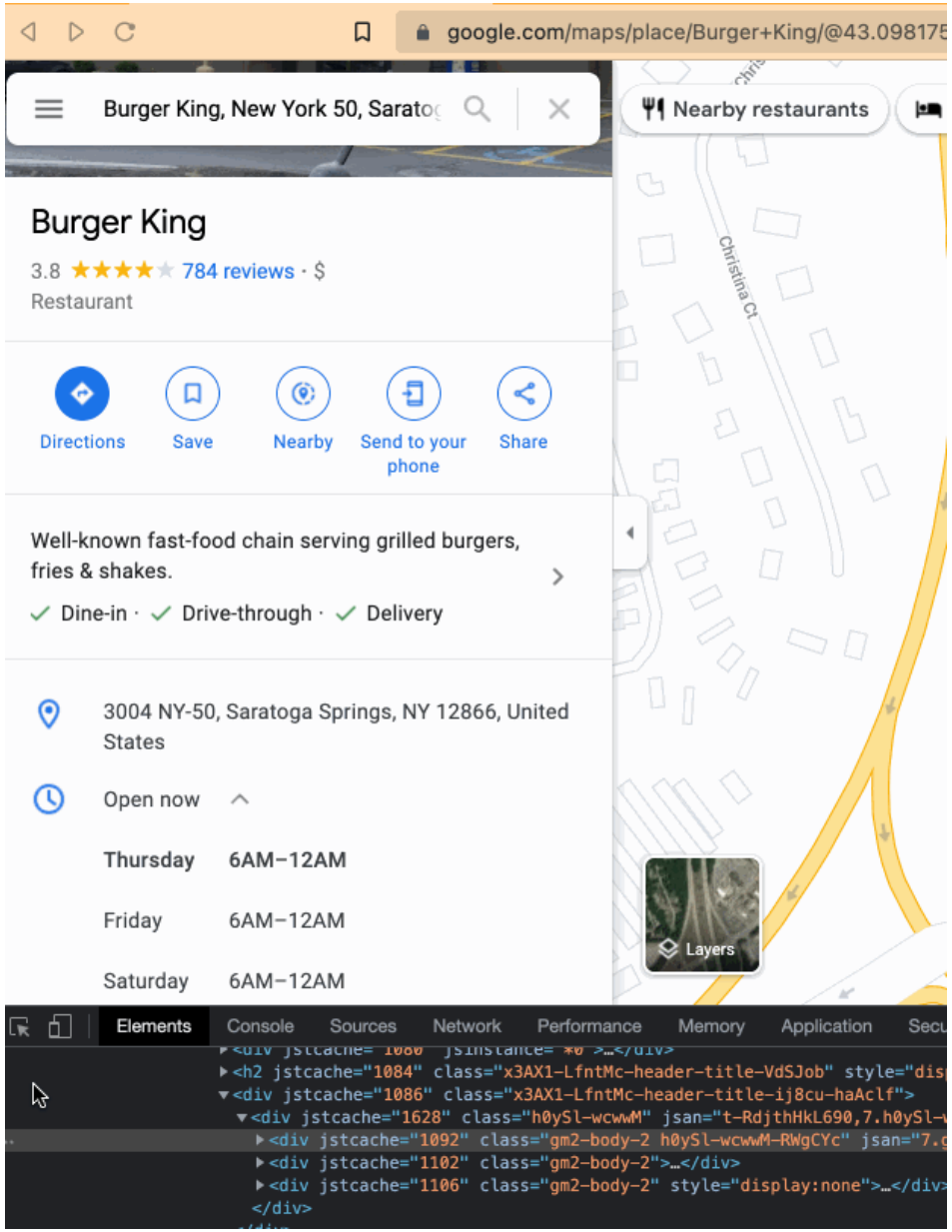
Permette di estrarre tutti i link presenti in un sito (semplice) in modo da usare trafilatura o beautifulsoup per estrarre l'intero contenuto di un sito

```
Processing URLs: 100%|██████████| 46/46 [00:12<00:00,  3.77it/s]
```

	url	text
0	https://lefolliedinaruto.wordpress.com/2022/05...	Introduzione:\nEsistono due emulatori validi: ...
1	https://lefolliedinaruto.wordpress.com/2021/01...	L'articolo sarà suddiviso nei seguenti punti:\...
2	https://lefolliedinaruto.wordpress.com/2019/05...	Più che il manga, possiamo dire che è Kishimot...
3	https://lefolliedinaruto.wordpress.com/2018/04...	(le parole sottolineate contengono link clicca...
4	https://lefolliedinaruto.wordpress.com/2018/03...	Avevo detto che non avrei più scritto articoli...

```python
df['text'].iloc[0]
```

```
'Introduzione:\nEsistono due emulatori validi: Duckstation e Beetle PSX HW (Retroarch). Questa guida è per Beetle, a mio parere il migliore, anche se l'emulatore perfetto non esiste…tutt'oggi potrest
e avere qualche problema con l'emulazione in HD, perciò usate entrambi questi ottimi emulatori…se un gioco non funziona bene su uno, usate l'altro.\nEsempio: Resident Evil 2 su Duckstation riproduce
sfondi prerenderizzati con minor qualità di Beetle, mentre per Metal Gear, Duckstation riproduce gli effetti framebuffer meglio di Beetle (difetti cerchiati in rosso).\nLa guida:\nScaricate Retroarch
qui, scrollate in basso e cliccate sotto Windows: "Download 64 bit".\nNon dovete installare niente, scaricherà un archivio che potete estrarre dove volete.\nProcuratevi il bios qui, consiglio sia il
giapponese, americano ed europeo. Quindi scaricate: scph5500.bin, scph5501.bin, scph5502.bin.\nRinominateli come li ho scritti, senza maiuscole, e posizionateli nella cartella di Retroarch\\system\\B
IOS (...'
```



Librerie e metodologie

BeautifulSoup

BeautifulSoup è la libreria di Python più usata per estrarre informazioni da pagine in html. Può essere utilizzato per estrarre informazioni diverse all'interno della pagina

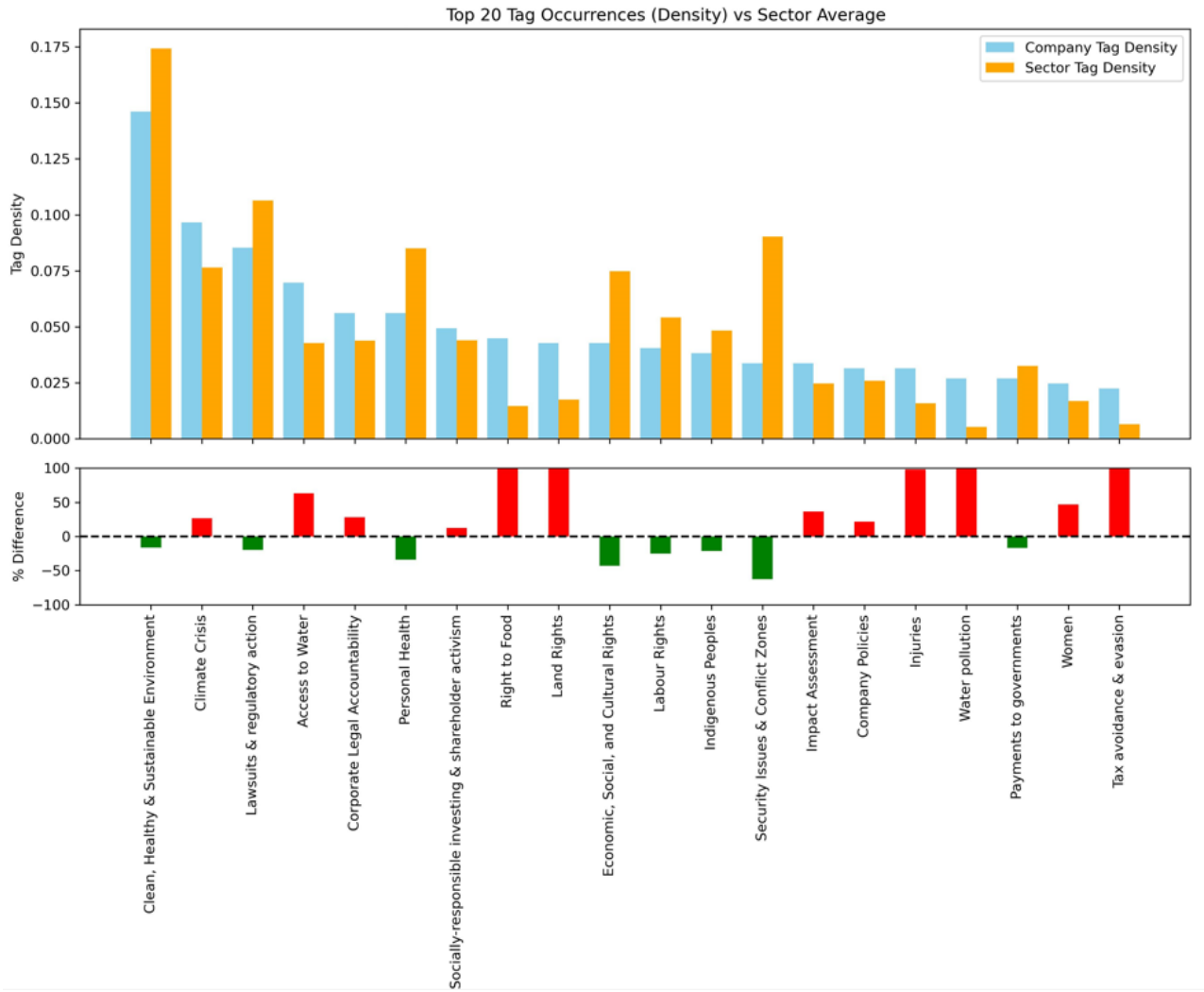
Selenium

Tool di scraping piu potente che simula la navigazione di un utente, cliccando bottoni, aprendo link e permettendo di navigare il sito ed estrarre informazioni durante il percorso

Esempio

Estraendo da un sito di news tag come aziende, categorie, settori e location possiamo per esempio fare una analisi di benchmark tra il numero di eventi (articoli) accaduti ad una singola azienda rispetto quelli medi suo settore di appartenenza

title	date	tags	location	article_body	industry	companies	type
PVH's response - Business & Human Rights Resou...	10 May 2019	['Freedom of Association', 'Labour Rights', 'L...	['Ethiopia']	... PVH takes the allegations raised in the WR...	['Clothing & textile']	['PVH (Phillips-Van Heusen)']	NaN
UK: Workers' watchdog could ban sweat-shop clo...	9 Jun 2021	['Labour Rights', 'Forced Labour & Modern Slav...	['United Kingdom']	'Workers' watchdog could ban sweat-shop clothe...	['Clothing & textile']	['Boohoo']	Tipo non disponibile
Uzbekistan: ILO welcomes lifting of Cotton Cam...	16 Mar 2022	['Forced Labour & Modern Slavery', 'Child labo...	Location non trovata	Agricultural and economic reforms have led to ...	['Clothing & textile']	[]	Tipo non disponibile
Myanmar: Garment workers forced to sleep overn...	28 Aug 2024	['Gender Discrimination', 'Labour Rights', 'Ch...	['Myanmar']	In March 2024, it was reported that workers at...	['Clothing & textile']	['Fast Retailing', 'Solamoda Garments Co., Ltd...	NaN
Sri Lanka: Trade unions yet to officially join...	14 Apr 2022	['Protests', 'Freedom of Association', 'Freedo...	['Asia & Pacific', 'Sri Lanka']	"What are Sri Lanka's trade unions doing amid ...	['Diversified/Conglomerates', 'Clothing & text...	[]	NaN



Social listening

Permette di raccogliere e analizzare dati pubblicamente disponibili da social media e altre fonti per ottenere insight utili in ambiti investigativi, strategici o di monitoraggio

- API Social Media: API Twitter, Reddit
- Provider di dati: Sprinklr, Brandwatch, Meltwater, ecc
- Strumenti di scraping
  - Apify: Versatile per il web scraping su diverse piattaforme.
  - Exportcomments: Specifico per l'estrazione di commenti pubblici.
  - PhantomBuster: Automazione di attività di scraping e ricerca su social media.

Applicazioni

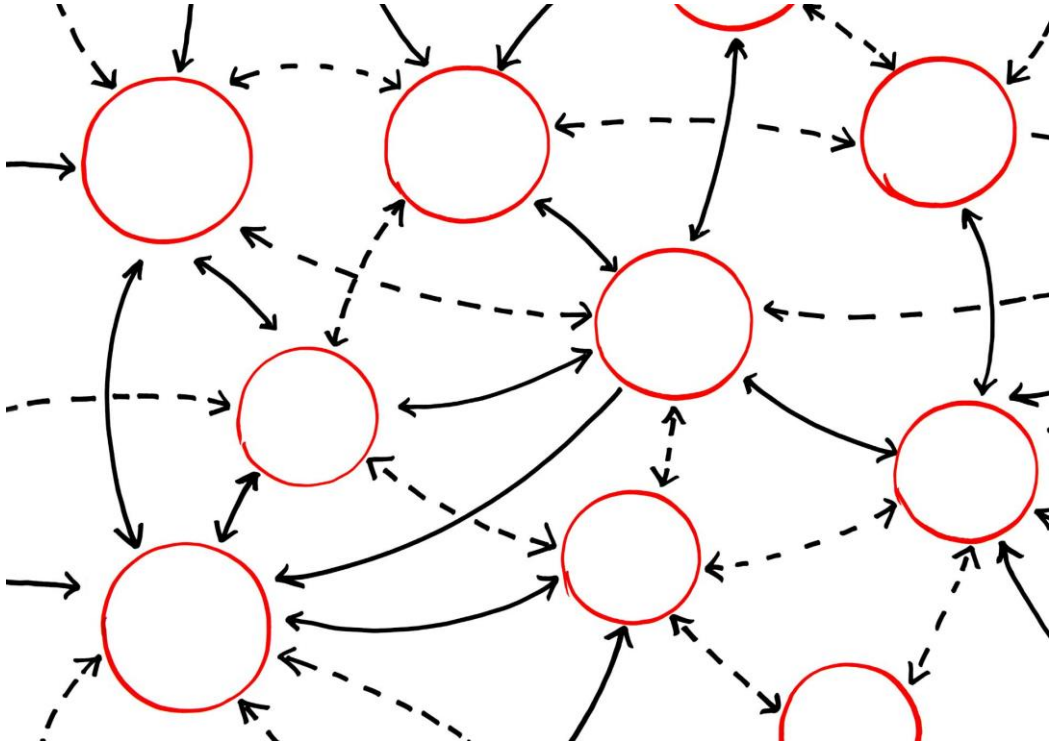
- ✓ Monitoraggio di Eventi in Tempo Reale
- ✓ Analisi Geopolitica e di Sicurezza
- ✓ Indagini su Minacce o Fake News
- ✓ Profilazione e Network Analysis

Dataset

Per esercizio prendiamo il dataset pubblico «How ISIS Uses Twitter» proveniente da kaggle e contenente 17000 tweets di circa 100 autori pro ISIS

Impareremo a:

- Individuare e clusterizzare gli utenti più influenti tramite tecniche di grafi
- Analizzare le conversazioni e produrre un mini report automaticamente
- Misurare il sentiment
- Individuare conversazioni pro e contro ISIS

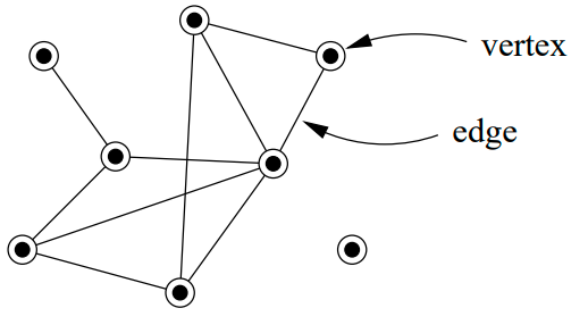


RETI

Un grafo o «rete» è un insieme di elementi, chiamati vertici o talvolta **nodi**, con connessioni tra loro chiamate **spigoli** (link).

L’interazione tra utenti in un social network può essere modellata come rete in cui gli utenti sono nodi e i link sono le interazioni (es utente menziona altro utente)

Reti in python



username	tweets	Mentions
theflamesofhaqq	RT @Free_lance_jour: Reminder:\nThis happened ...	[Free_lance_jour]
mobi_ayubi	RT @Malcolmite: Confirmed: Syrian Rebels captu...	[Malcolmite]
Uncle_SamCoco	@FranceTn @_DavidThomson le savoir français e...	[FranceTn, _DavidThomson]
warreporter2	@Abulzzadeen2 https://t.co/zCelsz2dd8	[Abulzzadeen2]
Abu_Azzzam25	RT @AbuNaseeha_03: #Israel Defence Forces are ...	[AbuNaseeha_03]

Per creare reti utilizziamo la libreria “networkx” , creando l’elenco degli utenti “source” che menzionano gli utenti “target”

```
# Creare una lista di sorgenti, destinazioni e pesi degli archi
sources = []
targets = []
edges = {}  # Dizionario per salvare i pesi degli archi

for row in df.iterrows():
    source = str(row[1]["username"])  # Utente che scrive il tweet
    targets = str(row[1]["Mentions"])  # Utenti menzionati
    if targets != '[]':  # Consideriamo solo tweet con menzioni
        for target in targets.split(","):
            target = re.sub(r'[^\w\s]', ' ', target).strip()  # Rimuovere punteggiatura e spazi bianchi
            weight = (source, str(target))  # Crea una tupla (sorgente, destinazione)
            if weight in edges:  # Incrementa il peso se l'arco esiste
                edges[weight] += 1
            else:  # Altrimenti, inizializza il peso a 1
                edges[weight] = 1

# Stampare un esempio di arco con il suo peso
for knot, weight in edges.items():
    print(knot, '|', weight)
    break
```

```
('GunsandCoffee70', 'KhalidMaghrebi') | 2
```

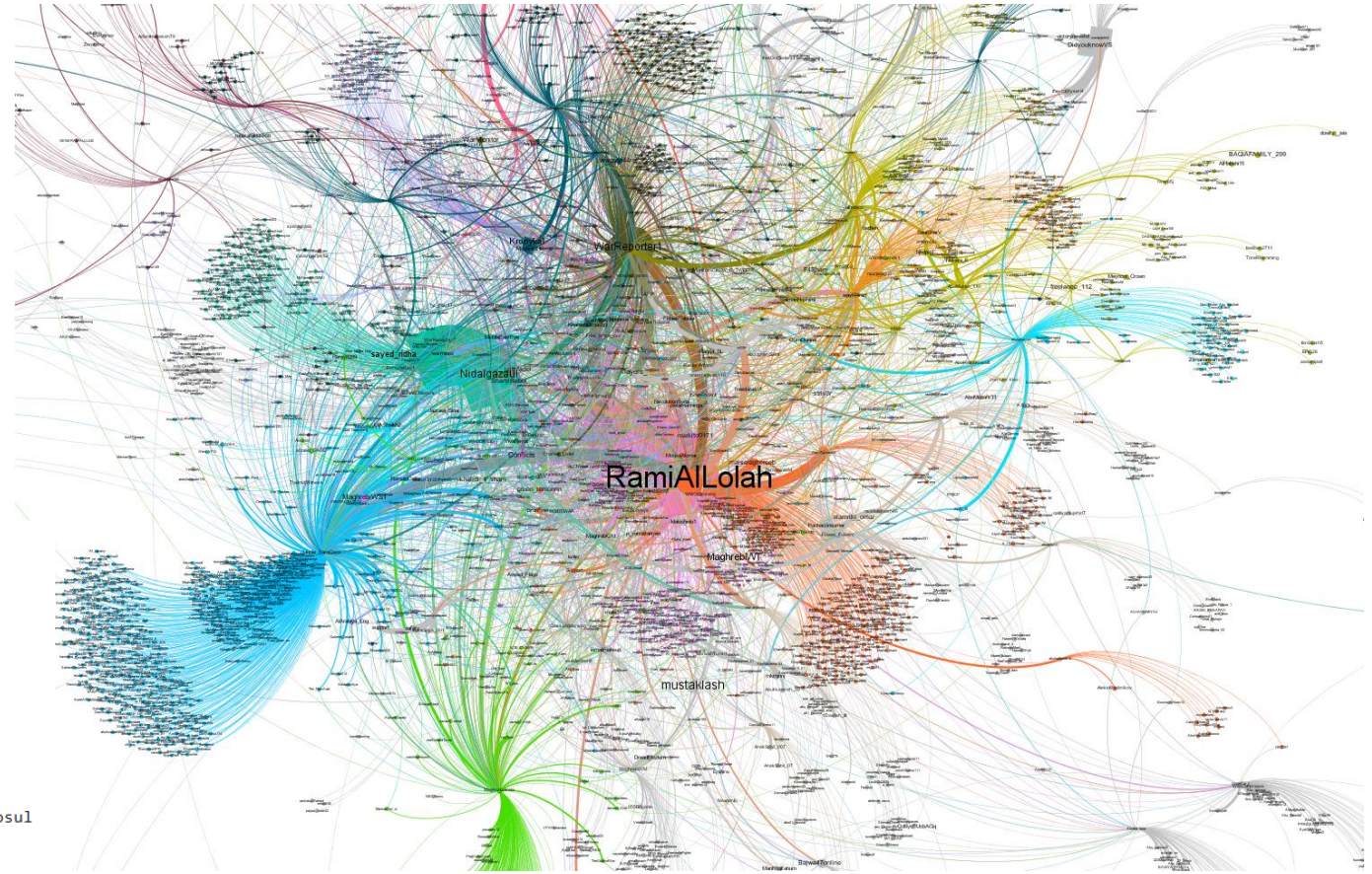
```
# Creare un grafo diretto
DG = nx.DiGraph()
for k, v in edges.items():
    source = k[0]
    target = k[1]
    weight = v
    DG.add_edge(source, target, weight=weight)  # Aggiungere l'arco al grafo con il peso
```

Utenti piu influenti

Tramite la struttura a grafo possiamo usare metriche per individuare gli utenti più influenti della rete e trovare le community in cui si distribuiscono

```
# Calcolare la centralità di PageRank
page_rank_dict = nx.pagerank(DG)  # Esegue l'algoritmo di PageRank
```

```
# Applicare l'algoritmo di Louvain
partition = community_louvain.best_partition(G, resolution=2)
```

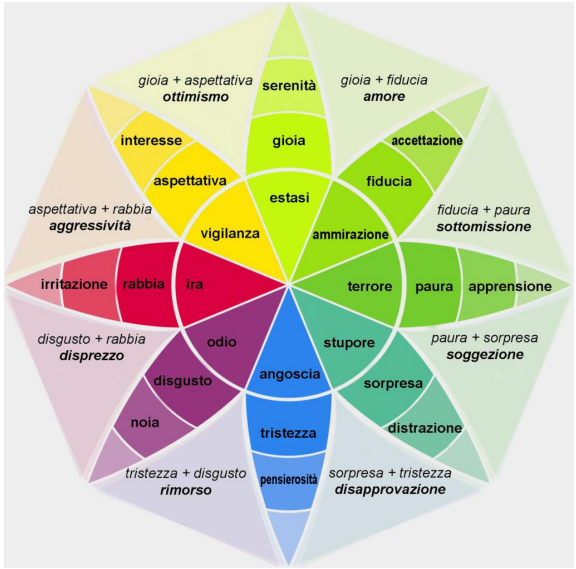


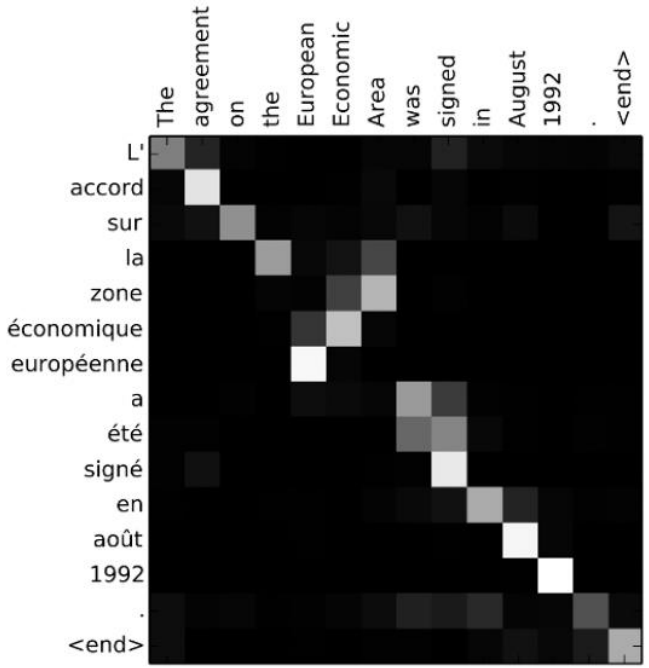

```
Username: RamiAlLolah
PageRank: 0.0015
Tweet: .@Mohammed_Aloush اللي جايكم على جنيف عم ينجرللكم الخازوق بقا حضروا حالكم..
----------------------------------------------------
Username: btt_ar
PageRank: 0.0009
Tweet: 🏴#WilayatDijlah
■||" Their #Assembly will be #Defeated, and They will Show Their Back "||
All links ⬇️
https://t.co/MNJYY0r7Wq
----------------------------------------------------
Username: Al_Battar_Engl
PageRank: 0.0009
Tweet: Aspects of the work of center services - the construction of a park for children in the area of mansions #Mosul
https://t.co/XmEwfjfB63
----------------------------------------------------
```

Mi piace la pizza!

Il film era veramente brutto

Mario è andato al parco





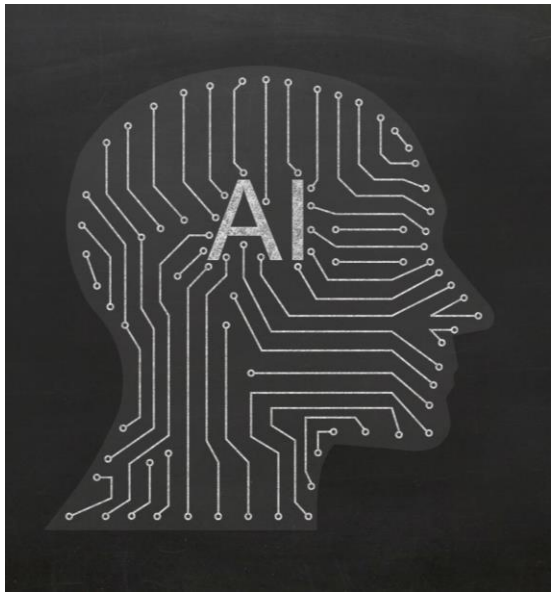
Natural Language Processing

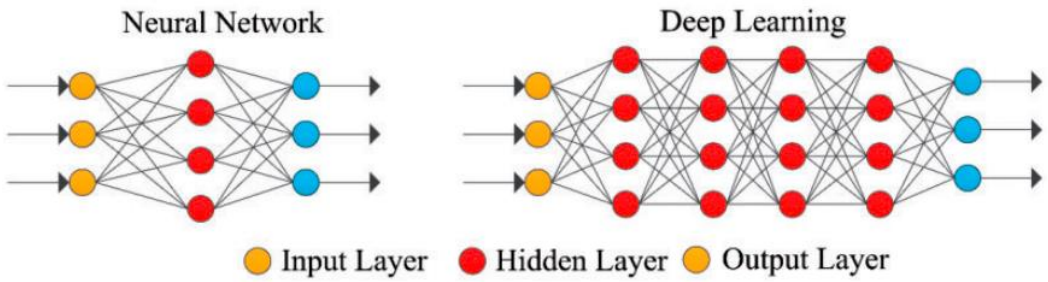
Insieme delle tecniche di linguistica e computer science che permettono a modelli matematici di processare ed elaborare testi

Esempi:

- Analisi di sentiment ed emozioni
- Traduzioni
- Similarità tra testi
- Estrazione di topic e concetti
- Generazione di testi

Machine & Deep Learning





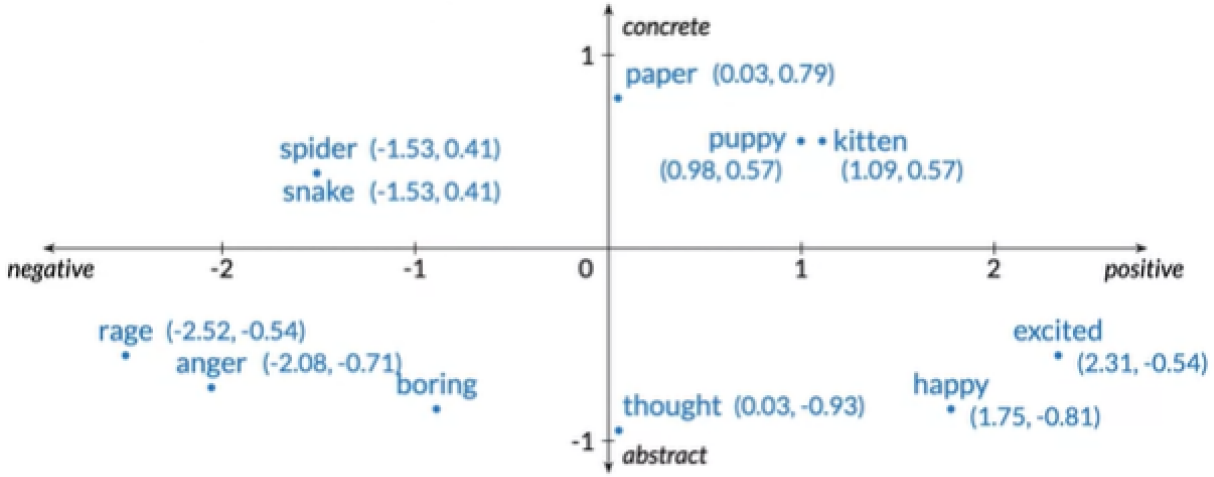
Machine Learning

studia l'abilita di modelli matematici di imparare a svolgere una certa attività senza essere esplicitamente programmati a farlo

Deep Learning

Evoluzione del Machine Learning che consiste nell’utilizzare reti neurali «profonde» capaci di modellare strutture e pattern complessi nei dati

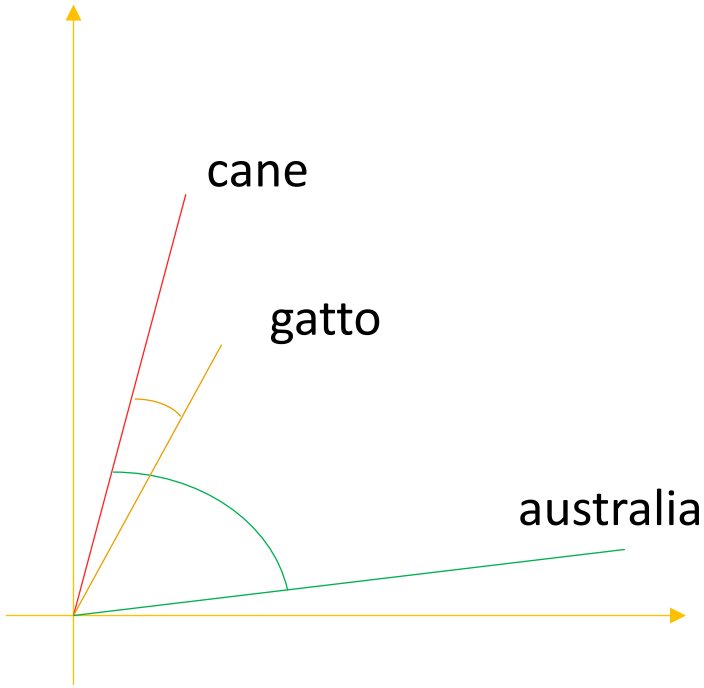
Come fa l AI a capire testi



Gli algoritmi sono oggetti matematici ed hanno bisogno di numeri!

EMBEDDING

Rappresentazioni che collegano parole a vettori matematici legati al significato delle parole



Clustering

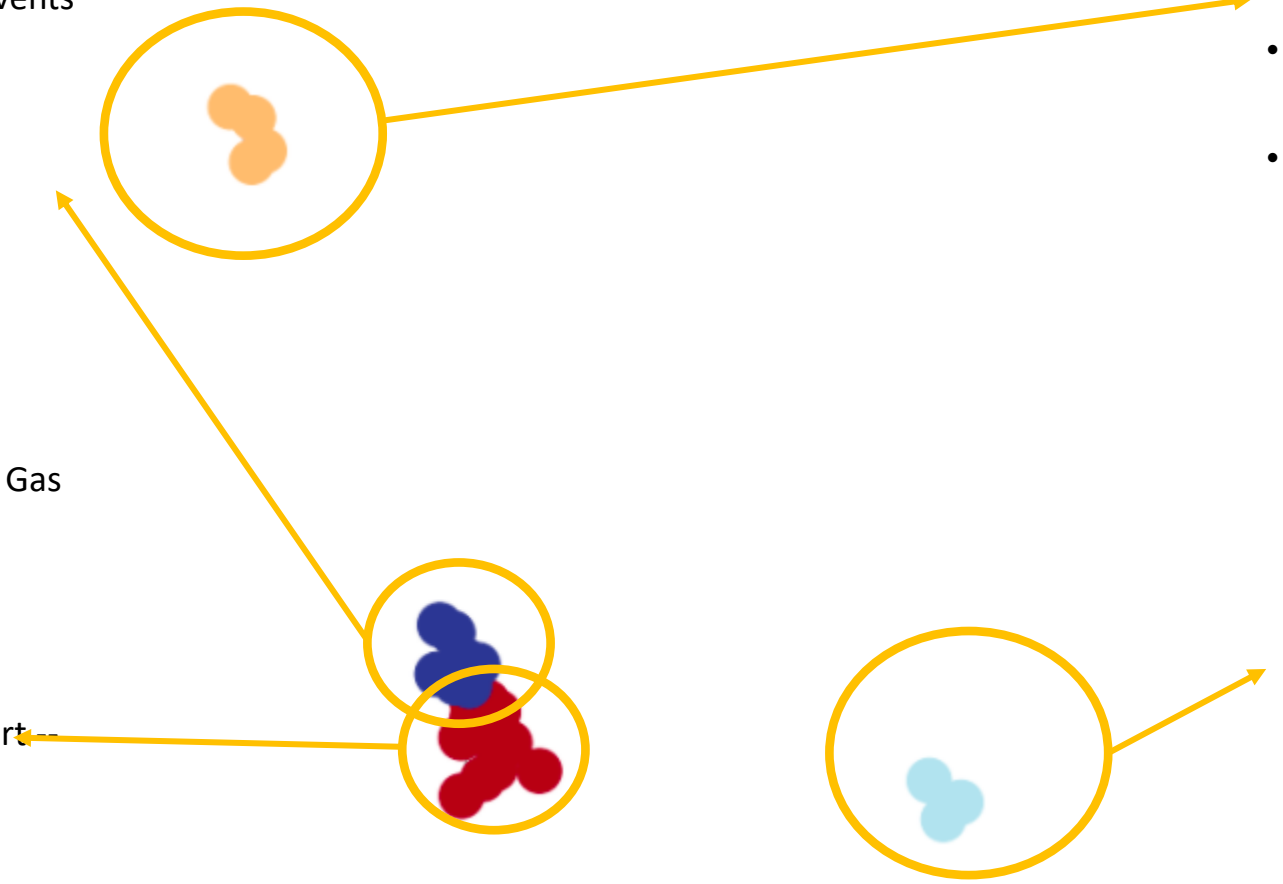
Dividere un dataset in classi non note

Notizie di cronaca

- 'Ndrangheta: 3 arresti per tentate estorsioni a Reggio Calabria
- 'Ndrangheta, tentata estorsione: 3 arresti nel reggino contigui cosche Libri-Morabito
- 'Ndrangheta: tentata estorsione a ditta nel Reggino,arresti
- 'Ndrangheta:arresti tentata estorsione, scoperta anche frode

Notizie legate al mondo corporate

- 'DJ Sweden Calendar of Corporate Events - Month Ahead
- 'DJ Norway Calendar of Corporate Events - Month Ahead
- 'DJ Portugal Calendar of Corporate Events - Month Ahead
- …

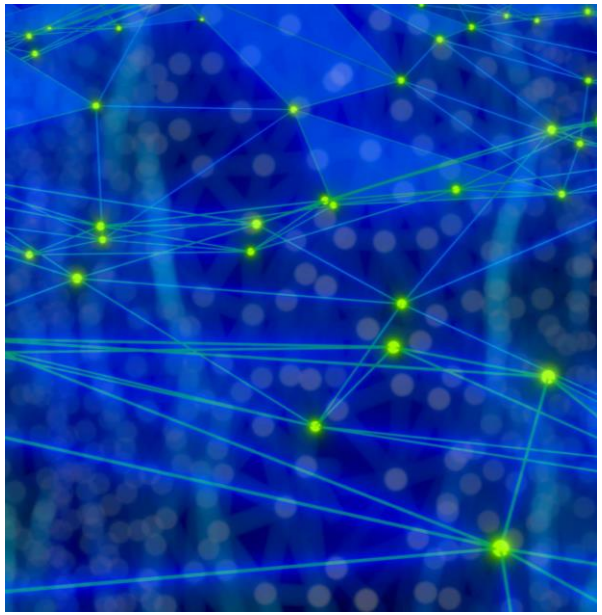


Notizie legate al mondo finanziario

- 'DJ Falcon Oil & Gas Ltd. Falcon Oil & Gas Ltd. - Full Year Results\\n',
- 'DJ BP 1Q Profit Rose More Than Expected; Launches $500M Share Buyback\\n',
- 'DJ Credit Suisse Board Loses Support -- WSJ\\n',
- 'DJ Nasdaq, S&P Jump to Records -- WSJ\\n',
- …

Notizie legate al mondo energetico

- DJ Falcon Oil & Gas Ltd. Falcon Oil & Gas Ltd.
- Venezuela's socialist-held congress to ratify 'anti-blockade' law, official says
- US weighs policy on Venezuela as Maduro signals flexibility



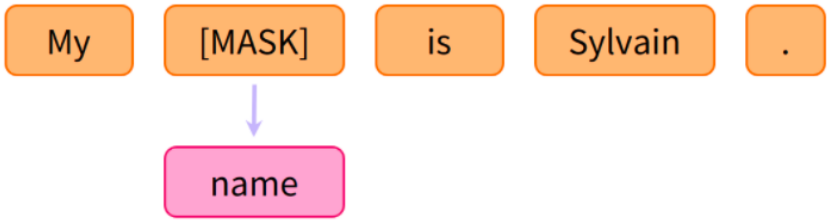
Analisi degli argomenti

BERT

BERT è un modello transformers di deep learning utilizzato per l'elaborazione del linguaggio naturale. Permette di converire grandi quantità di testo in embedding

BERTopic

BERTopic è un algoritmo di clustering che utilizza BERT per raggruppare gli elementi di testo in gruppi di argomenti correlati. Può essere utilizzato per analizzare i dati di Twitter e identificare i topic che emergono sui social media.



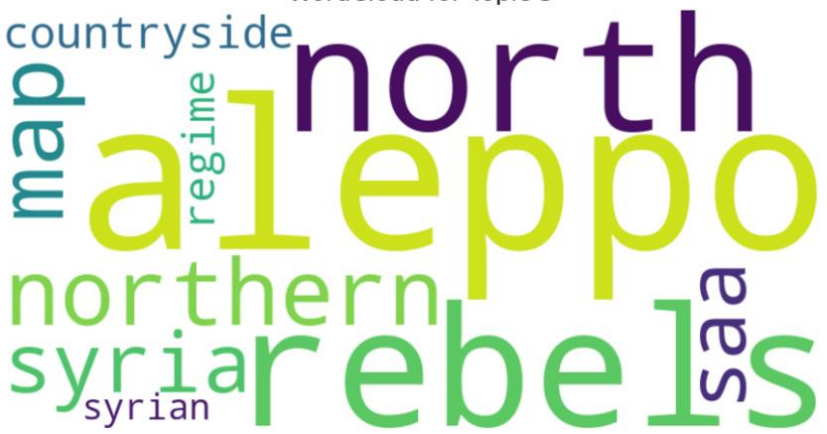
TopicModel

```
# Creare il modello BERTopic
# La variante "paraphrase-MiniLM-L3-v2" è più leggera rispetto a modelli più grandi.
topic_model = BERTopic(embedding_model="paraphrase-MiniLM-L3-v2")

# Applico il modello
topics, probs = topic_model.fit_transform(docs)
df['Topic']=topics
```

	Topic	Count	Name	Representation	Representative_Docs
0	-1	8829	-1_the_is_in_to	[the, is, in, to, isis, of, and, rt, https, co]	[No matter what others may put on you, forgive...
1	0	544	0_de_le_la_est	[de, le, la, est, les, je, des, pas, et, un]	[@Annabbii mais croire que Janviériste sont de...
2	1	518	1_co_https_دعم_ha	[co, https, دعم, ha, ameen, lol, this, read, l...	[Ha ha https://t.co/m7kisZ4500, Ha ha ha http...
3	2	222	2_في_من_على_الله	[في, من, على, الله, بسم, الدولة, الرحمن, الرحي...	[RT @ABUHIBBAN9: بسم الله الرحمن الرحيم \n\n\ال...
4	3	207	3_aleppo_rebels_north_northern	[aleppo, rebels, north, northern, map, syria, ...	[RT @Buzzriet: #Syria #Aleppo\nAnd another att...
...	...	...	...	...	...
192	191	10	191_haneefah_rulers_rebellion_supported	[haneefah, rulers, rebellion, supported, imam,...	[Imam Abu Haneefah's views about fighting rule...
193	192	10	192_muslims_mosque_unite_muslimfromchina	[muslims, mosque, unite, muslimfromchina, chin...	[RT @MuslimPrisoners: Muslims inviting society...
194	193	10	193_units_infiltrate_kurdish_shuyukh	[units, infiltrate, kurdish, shuyukh, kobani, ...	[#Breaking 3 Kurdish units killed yesterday as...
195	194	10	194_manjanik_news_siyono_hanya_yang	[manjanik_news, siyono, hanya, yang, bisa, kep...	[RT @manjanik_news: Tak Hanya Ustadz Ba'asyir,...
196	195	10	195_fake_shawjary900_bintislamiya19_her	[fake, shawjary900, bintislamiya19, her, accou...	[@bintislamiya19 @WarReporter1 @shawjARY900 th...

WordCloud for Topic 3



WordCloud for Topic 4

Cosa sono i Large Language Models

I Large Language Models sono algoritmi di deep learning addestrati su enormi volumi di dati e costituiti da **miliardi** di parametri che permettono di analizzare e generare testo in modo automatico, comprendendo il contesto e potendo essere adattati alle piu diverse applicazioni

Utilizzeremo LLM sfruttando le API di together.ai

GPT-4

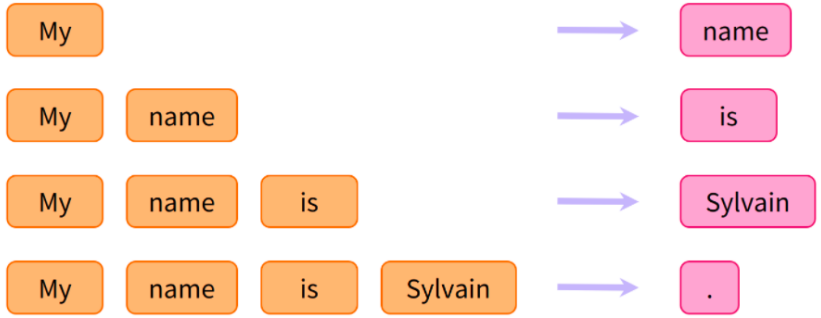
Prodotto da OpenAI, è uno dei più grandi modelli di linguaggio attualmente disponibili

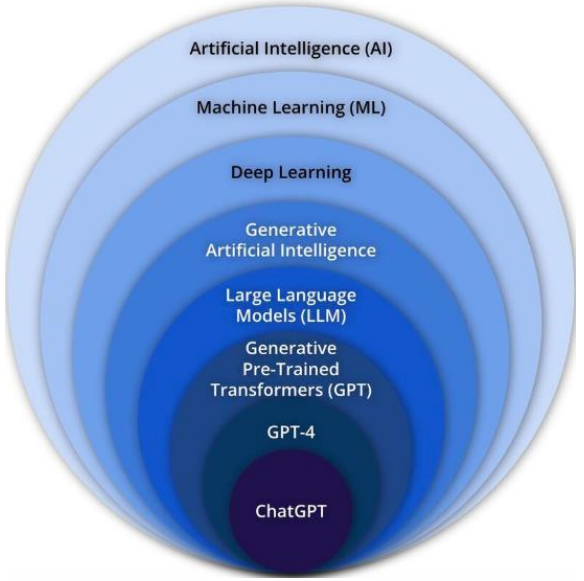
LLAMA

Prodotto da Facebook è opensource e disponibile in diverse versioni e grandezze

Per ulteriori approfondimenti «Prompt and predict»

CAMBIA TOTALMENTE IL MODO IN CUI L'UTENTE SI INTERFACCIA ALL'AI: QUI DEVI SCRIVERE IN LINGUAGGIO NATURALE!





Riassunti

```python
# Invio del testo direttamente al modello
response = client.completions.create(
    model="meta-llama/Meta-Llama-3.1-70B-Instruct-Turbo",
    prompt=f"""Sei un analista OSINT che analizza diversi tweet per estrarre informazioni
              Riassumi ed estrai il concetto chiave in una frase dai seguenti tweet:\n\n{combined_text}
              """

            ,
    max_tokens=300,
    temperature=0.5
)
```

```
Riassunto per Topic ID 0:
Riepilogo dei concetti chiave dei tweet:
                  I tweet parlano di vari argomenti, tra cui:
- L'importanza della perseveranza e della costanza nella lotta contro l'ingiustizia.
- La critica alle politiche occidentali e la loro influenza negativa sui paesi musulmani.
- La condanna delle azioni di gruppi estremisti come l'ISIS e la loro ideologia.
- La promozione della conoscenza e dell'educazione come strumenti per combattere l'ignoranza e la disinformazione.
- La riflessione sulla natura della verità e della realtà, e l'importanza di cercare la conoscenza e la saggezza.
- La critica alla società occidentale e alla sua cultura, accusandola di essere materialista e corrotta.
- La promozione della fratellanza e della solidarietà tra i musulmani e la necessità di lavorare insieme per raggiungere gli obiettivi comuni.
- La condanna della corruzione e dell'ingiustizia nei paesi musulmani e la necessità di riforme politiche e sociali.
- La riflessione sulla natura della fede e della spiritualità, e l'importanza di cercare
---

Riassunto per Topic ID 1:
Il tweet parla di vari argomenti tra cui politica, terrorismo, religione e cultura. Alcuni utenti discutono di eventi attuali come la guerra in Siria e l'attacco terroristico a Bruxelles, mentre altri
---

Riassunto per Topic ID 2:
Gli analisti OSINT analizzano i tweet per estrarre informazioni su Aleppo, in Siria. I tweet riportano eventi come la distruzione di un convoglio dell'esercito siriano, l'uccisione di soldati dell'ese
---

Riassunto per Topic ID 3:
Il concetto chiave in una frase che riassume i tweet è: "La lotta contro il terrorismo e la violenza in Medio Oriente, con un focus sui recenti attacchi in Siria, Iraq e Turchia, e le reazioni delle f
---

Riassunto per Topic ID 4:
I tweet sembrano essere una raccolta di messaggi di vari utenti su piattaforme di social media, principalmente Twitter. I messaggi coprono vari argomenti, tra cui politica, religione, tecnologia e att
---
```

NB: ha risposto in italiano a tweet in arabo!

Intero report!

```
# Invio del testo direttamente al modello
response = client.completions.create(
    model="meta-llama/Meta-Llama-3.1-70B-Instruct-Turbo",
    prompt=f"""Sei un analista OSINT che analizza diversi tweet per estrarre informazioni.
                Partendo dai riassunti dei topic analizzati:
                {topic_string}

               Scrivi in italiano
               Attieniti a quanto riportato nei topic
               Riassumi in un report di intelligence per capire come viene utilizzato twitter
               per propaganda ISIS in particolare se è un canale efficace o antagonista e
               se emergono tematiche particolari
               """
                ,
    max_tokens=1000,
    temperature=0.5
)

# Estrai il riassunto dalla risposta
summary = response.choices[0].text.strip()
print(summary)
```

La propaganda dell'ISIS su Twitter è un fenomeno complesso e multiforme. Dai tweet analizzati emerge un quadro articolato di argomenti e strategie utilizzate dall'organizzazione terroristica per diffondere la propria ideologia e influenzare l'opinione pubblica.

In primo luogo, è evidente che l'ISIS utilizza Twitter per **diffondere messaggi di odio** e violenza contro gli avversari, compresi i governi occidentali e le forze armate che combattono contro di loro. I tweet contengono spesso minacce e incitamenti alla violenza, nonché immagini e video di attacchi e uccisioni.

Tuttavia, l'analisi dei tweet suggerisce che l'ISIS non utilizza solo Twitter per diffondere messaggi di violenza e odio. L'organizzazione cerca anche **di promuovere la propria ideologia e di attrarre nuovi adepti**, utilizzando argomenti come la critica alla società occidentale e la promozione della conoscenza e dell'educazione come strumenti per combattere l'ignoranza e la disinformazione.

Inoltre, i tweet analizzati suggeriscono che l'ISIS cerca di sfruttare la piattaforma per diffondere notizie e informazioni false, al fine di influenzare l'opinione pubblica e creare confusione. Ad esempio, alcuni tweet riportano notizie false su attacchi e uccisioni, mentre altri condividono articoli e video che promuovono la propaganda dell'ISIS.

Tuttavia, è importante notare che l'uso di Twitter da parte **dell'ISIS non è esente da critiche e opposizioni**. Molti utenti utilizzano la piattaforma per condannare le azioni dell'organizzazione e per esprimere solidarietà con le vittime degli attacchi terroristici. Inoltre, alcuni utenti condividono informazioni e notizie che **smentiscono la propaganda** dell'ISIS, fornendo una visione più equilibrata e obiettiva degli eventi.

Hugging Face

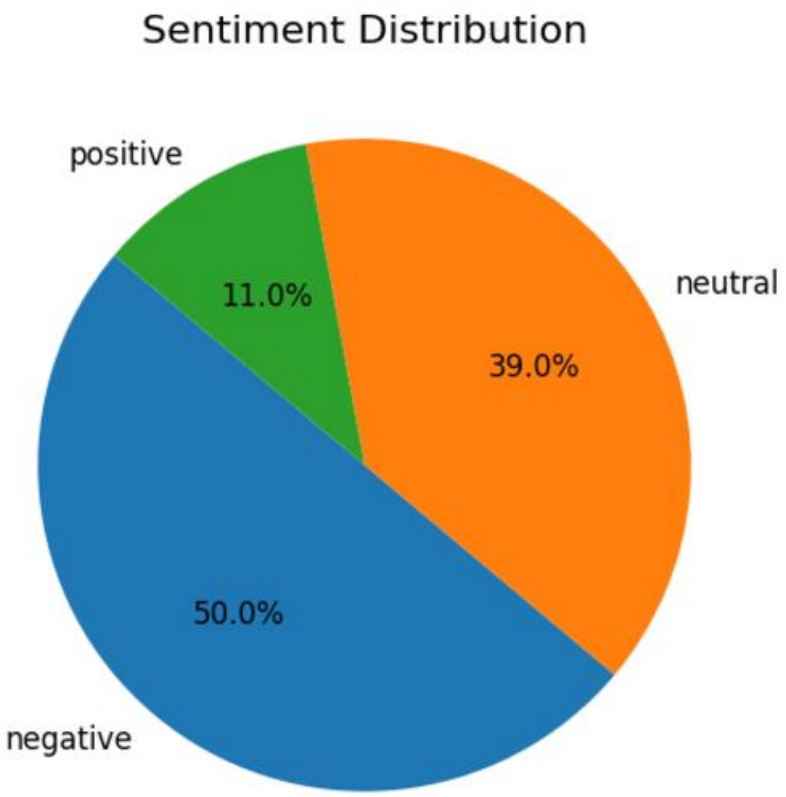
Hugging Face è una azienda statunitense fondata nel 2016. Contiene numerosi modelli open-source di intelligenza artificiale per l'elaborazione del linguaggio naturale come traduzioni, analisi delle emozioni, estrazioni di informazioni ecc

Sentiment

Utilizzeremo l’algoritmo multilingua RoBERTa XLM per analizzare il sentiment da dati di Twitter

(NB cercate i modelli con pipeline)

Sentiment




```python
from transformers import pipeline

model_path = "cardiffnlp/twitter-xlm-roberta-base-sentiment"
sentiment_task = pipeline("sentiment-analysis", model=model_path, tokenizer=model_path)
```

```python
# Funzione per analizzare il sentiment di un tweet
def analyze_sentiment(tweet):
    result = sentiment_task(tweet)
    # Restituisce il label (es. POSITIVE, NEUTRAL, NEGATIVE) o un altro valore utile
    return result[0]["label"]

df_small = df.sample(100)
df_small["sentiment"] = df_small["tweets"].astype(str).progress_apply(analyze_sentiment)
```

	username	tweets	sentiment
13233	al_zaishan10	#AmaqAgency \n#IslamicState \n#WilayatHalab \n...	neutral
10359	wayyf44rer	Beirut buses provide rare bridge to IS turf ht...	neutral
3515	mobi_ayubi	RT @lion_faisal: USA has rewarded #Iran with 1...	negative
7029	_IshfaqAhmad	Ramiz Raja you unbeauty. #AsiaCupT20Final	negative
16718	Uncle_SamCoco	RT @SimNasr: A good read https://t.co/4BEqUWtGdI	positive

E’ un sentiment commerciale, è quello che ci serve?

Zero shot learning

I Large Language Models sono talmente potenti da poter svolgere compiti specifici **senza essere specificamente addestrati** per svolgerli, potendo generalizzare!

Questo è proprio il «zero shot learning» in cui gli chiediamo di fare qualcosa senza dargli esempi (ma spiegando bene cosa ci serve)

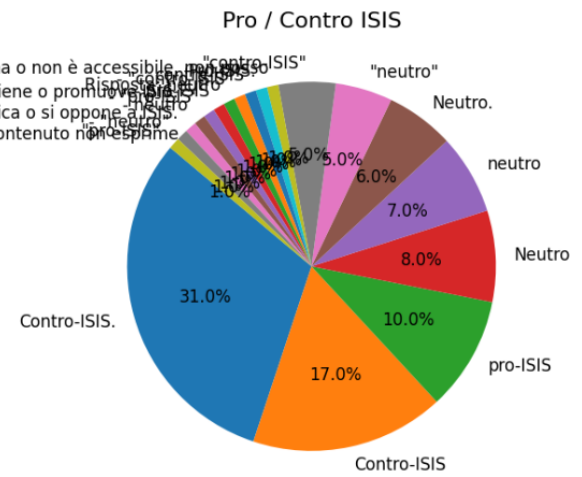
Possiamo quindi analizzare i tweet per capire se sono pro o contro l ISIS senza dover addestrare un classificatore apposito (come ROBERTAXLM per il sentiment)!

Pro / contro ISIS

	username	tweets	classification
13233	al_zaishan10	#AmaqAgency \n#IslamicState \n#WilayatHalab \n...	- pro-ISIS
10359	wayyf44rer	Beirut buses provide rare bridge to IS turf ht...	Contro-ISIS.
3515	mobi_ayubi	RT @lion_faisal: USA has rewarded #Iran with 1...	Contro-ISIS.
7029	_IshfaqAhmad	Ramiz Raja you unbeauty. #AsiaCupT20Final	Neutro.
16718	Uncle_SamCoco	RT @SimNasr: A good read https://t.co/4BEqUWtGdl	Contro-ISIS.

```python
# Funzione per classificare un tweet
def classify_tweet(tweet):
    prompt = f"""
Sei un analista OSINT. Analizza il seguente tweet:
"{tweet}"
Classificalo come:
- "pro-ISIS" se il contenuto sostiene o promuove ISIS.
- "contro-ISIS" se il contenuto critica o si oppone a ISIS.
- "neutro" se il contenuto non esprime una posizione chiara.
Rispondi solo con una delle tre categorie.
    """
    # Invia la richiesta al modello
    response = client.completions.create(
        model="meta-llama/Meta-Llama-3.1-70B-Instruct-Turbo",
        prompt=prompt,
        max_tokens=50,
        temperature=0.3
    )
    # Restituisce la classificazione come testo
    return response.choices[0].text.strip()
```

Con il link fornito non è possibile classificare il tweet senza ulteriori informazioni. Tuttavia, posso fornire una risposta generica. Se il link non funziona o non è accessibile, - "pro-ISIS" se il contenuto sostiene o promuove ISIS. - "contro-ISIS" se il contenuto critica o si oppone a ISIS. - "neutro" se il contenuto non esprime



Conclusione

Il mondo è pieno di dati: la sfida non è solo raccoglierli, ma trasformarli in informazioni utili e saperci porre le giuste domande.

Open Data e web scraping

Ottenere insight e comprendere fenomeni attraverso analisi accessibili a tutti

Tecniche di data science

Abilitano l'analisi di grandi volumi di dati in modo veloce ed efficiente

AI non è un sostituto

Ma un acceleratore che amplifica la capacita di analisi e catalogazione

Materiale

LinkedIn @cesare-scalia-phd