Relazione tecnico scientifica

Paolo Valente 862489
Simone Tranquillo 888591

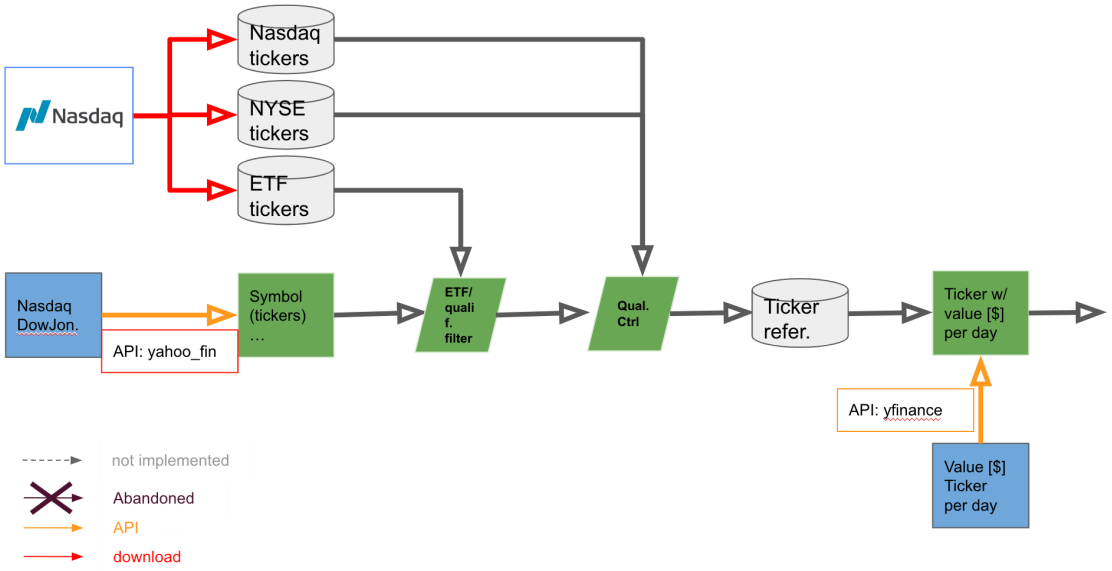
Problema da affrontare

L'obiettivo del progetto è associare ai dati storici giornalieri dei titoli azionari di XYZ società americane del NYSE, i post pubblicati sui social media.
Nel nostro caso si è scelto Reddit come social da cui estrarre i dati dei post.
L'idea alla base del progetto è quella di poter sfruttare tali dati per monitorare/osservare potenziali correlazioni tra le due entità.

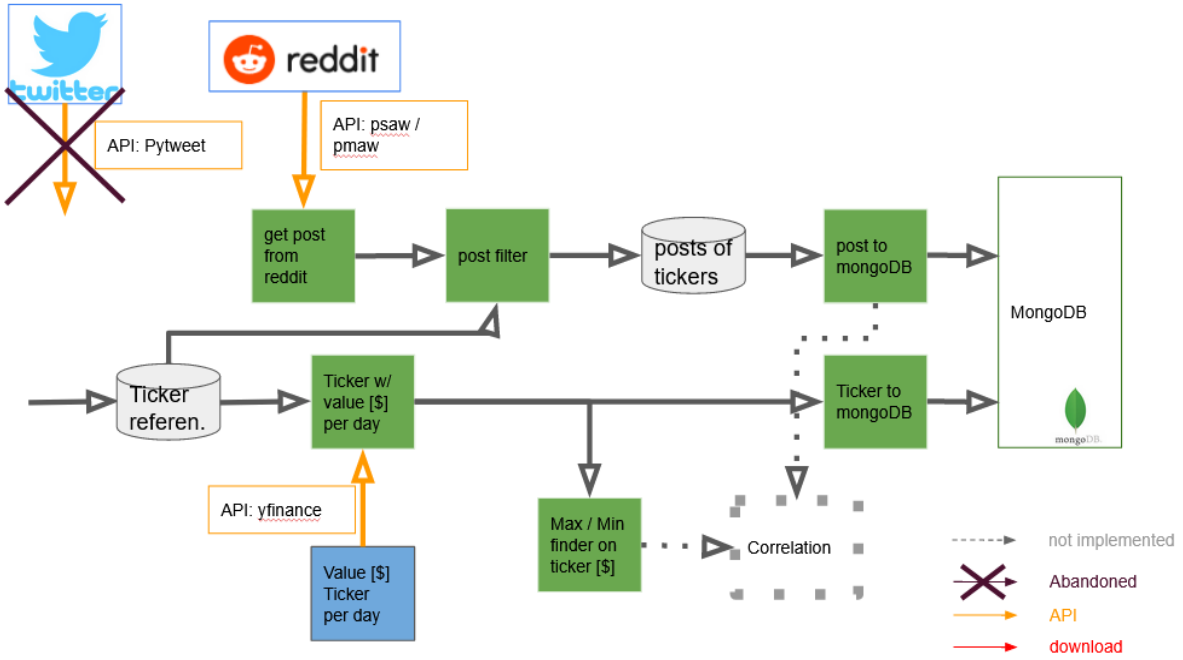
Tra le fonti di ispirazione di questo lavoro citiamo la vicenda "Gamestop", entrata sotto i riflettori di tutto il mondo a partire da Gennaio 2021, oltre che i vari tweet pubblicati da Elon Musk per i quali è stato accusato di market manipulation in quanto grande "influencer".

Summary delle attività

1) Acquisizione dati di borsa
    a) Generazione della lista dei tickers
    b) Controllo qualità



2) Acquisizione post di Reddit
    a) Creazione dataframe dei post
    b) Filtro e assegnazione ticker
3) Memorizzazione nel database MongoDB
4) Integrazione dei dati



Generazione della lista dei ticker

I simboli azionari sono noti in lingua inglese come “Ticker” ed identificano univocamente i titoli di una determinata compagnia (e.g.: AAPL si riferisce alle azioni di Apple Inc., TSLA a Tesla Inc. ecc…).

Download dei ticker

La lista dei ticker è stata ottenuta seguendo una procedura complicata. Tra i molteplici titoli e mercati azionari esistenti ci si è focalizzati solo su quelli di Nasdaq e Dow Jones.
L’acquisizione stessa dei tickers, inizialmente ottenuta tramite un parser HTML da diversi siti, è stata poi realizzata attraverso l’API ***stock_info*** di **yahoo_fin.**
Il motivo è da attribuire sia alla relativa facilità nell’adoperare l’API, sia alla policy di sicurezza restrittiva dei siti consultati.

Le seguenti funzioni permettono per esempio di acquisire i ticker da NASDAQ, Dow Jones, SP500:
stock_info.tickers_nasdaq()
stock_info.tickers_dow()
stock_info.tickers_sp500()
Tramite la definizione di specifici parametri si sono dunque scaricati i codici di Nasdaq e Dow Jones.

Esplorazione dei dati con i ticker

La lista dei ticker scaricati dopo un' esplorazione si è rivelata molto più ricca rispetto alle aspettative. Quindi è stato necessario intervenire per eliminare i titoli non qualificati, i fondi e gli ETF, per questo motivo a valle è stato applicato un doppio filtraggio.

Filtro dei titoli non qualificati e dei titoli Fund ed ETF

Ottenuti i tickers è stato fatto un primo filtering dei titoli non qualificati, cioè quelli con la quinta lettera pari a W, R, P e Q.
In seguito si è passati ad un secondo filtering tramite cui i titoli di fund ed ETS sono stati eliminati.
Questo metodo è stato realizzato incrociando i dati dei ticker ottenuti dall'API con i tickers scaricati direttamente dal sito del NASDAQ[1]; da quest'ultimo, infatti, è stata scaricata la lista contenente tutti gli ETF e Fund.
Il file contenente tali informazioni è stato salvato come ***nasdaq_etf_screener.csv:***

	SYMBOL	NAME	LAST PRICE	NET CHANGE	% CHANGE	DELTA	1 yr % CHANGE
0	JHMM	John Hancock Multifactor Mid Cap ETF	$44.7200	0.5400	1.2075134168157424%	up	-14.36%
1	JHMS	John Hancock Multifactor Consumer Staples ETF	$36.6274	0.5467	1.492598437235512%	up	6.80%
2	JHMT	John Hancock Multifactor Technology ETF	$69.7511	-0.5289	-0.7582676115502122%	down	-22.01%
3	JHMU	John Hancock Multifactor Utilities ETF	$36.2858	0.9027	2.487750028936939%	up	13.29%
4	JHSC	John Hancock Multifactor Small Cap ETF	$29.5346	0.3146	1.0651913349088866%	up	-15.93%
...	...	...	...	...	...	...	...
2653	ZHDG	ZEGA Buy and Hedge ETF	$16.9200	0.1685	0.9958628841607565%	up	-15.57%
2654	ZIG	The Acquirers Fund	$24.8201	0.2789	1.123686044778224%	up	-8.41%
2655	ZROZ	PIMCO 25 Year Zero Coupon U.S. Treasury Index...	$108.4400	0.0900	0.08299520472150498%	up	-25.04%
2656	ZSL	ProShares UltraShort Silver	$33.3800	1.1900	3.565008987417615%	up	42.35%
2657	Data as of : 7/1/2022 8:00:00 PM	NaN	NaN	NaN	NaN	NaN	NaN

2658 rows × 7 columns

A questo punto la lista iniziale di NASDAQ e Dow Jones si è ridotta a **4459 tickers** da usare per le query successive e per indicizzare il database a valle.

Controllo della qualità dei tickers

Sono stati utilizzati due metodi per il controllo qualità dei ticker in ingresso.
Il primo riguarda la presenza di missing value, il secondo, più complesso, ha riguardato il controllo di esistenza e veridicità dei ticker scaricati dall'API.

Per effettuarlo sono state scaricate dalla sorgente fidata del Nasdaq due liste di tickers: la prima relativa al Nasdaq e la seconda al NYSE che comprende i titoli del Dow Jones.
Anche in questo caso, quindi, è stato utilizzato lo screener messo a disposizione dal sito del Nasdaq e sono state prodotte le seguenti liste:

nasdaq_screener_NASDAQ.csv:

4869	ZWRK	Z-Work Acquisition Corp. Class A Common Stock	$9.83	-0.0100	-0.102%	282612500.0	United States	2021.0	284585	Industrials	Consumer Electronics/Appliances
4870	ZWRKW	Z-Work Acquisition Corp. Warrant	$0.1625	0.0425	35.417%	0.0	United States	2021.0	31880	Industrials	Consumer Electronics/Appliances
4871	ZY	Zymergen Inc. Common Stock	$1.29	0.0900	7.50%	133051574.0	United States	2021.0	760129	Industrials	Industrial Specialties
4872	ZYNE	Zynerba Pharmaceuticals Inc. Common Stock	$1.17	0.0000	0.00%	51007272.0	United States	2015.0	215779	Health Care	Biotechnology: Pharmaceutical Preparations
4873	ZYXI	Zynex Inc. Common Stock	$8.45	0.3800	4.709%	325325000.0	United States	NaN	125773	Health Care	Other Pharmaceuticals

4874 rows × 11 columns

nasdaq_screener_NYSE.csv:

3146	ZTS	Zoetis Inc. Class A Common Stock	$173.11	-1.1000	-0.631%	8.147053e+10	United States	2013.0	1195069	Health Care	Biotechnology: Pharmaceutical Preparations
3147	ZUO	Zuora Inc. Class A Common Stock	$9.50	0.4000	4.396%	1.227400e+09	NaN	2018.0	648730	Technology	EDP Services
3148	ZVIA	Zevia PBC Class A Common Stock	$2.88	0.2400	9.091%	1.125133e+08	NaN	2021.0	81171	Consumer Staples	Specialty Foods
3149	ZWS	Zurn Elkay Water Solutions Corporation Common ...	$27.73	0.2500	0.91%	3.493080e+09	United States	2012.0	415074	Utilities	Environmental Services
3150	ZYME	Zymeworks Inc. Common Shares	$5.75	0.2100	3.791%	3.321844e+08	Canada	2017.0	632603	NaN	NaN

3151 rows × 11 columns

In definitiva il file con i tickers di riferimento è stato salvato come
downloaded_symbol.csv:

	Symbol
0	DBTX
1	IVCP
2	BHSE
3	NAII
4	GNACU
...	...

4454	HYRE
4455	NXPI
4456	INCY
4457	CMAX
4458	NLITU

4459 rows × 1 columns

Download dei valori di borsa

I valori di borsa sono stati ottenuti attraverso l' API ***Ticker*** presente in ***yfinance***.
La funzione ha come ingresso il ticker più una serie di parametri, tra i quali i più importanti sono quelli che definiscono la finestra temporale di estrazione dei dati e granularità. Nella presente attività è stato indicato l'intervallo temporale del primo trimestre del 2022 ed una granularità di un giorno.
Essa restituisce una grande varietà di informazioni che spaziano dal settore di borsa ai dividendi ai valori di borsa [2]. Nello specifico sono stati considerati i valori di ogni titolo all'apertura, alla chiusura, massimo, minimo ed il volume delle azioni scambiate.

```
price_history = yf.Ticker("ticker").history(period=, # valid
period:1d,5d,1mo,..,5y,10y,ytd,max
                        interval=_interval, # valid intervals: 1m,2m,5m,15m,..,,1d,...
                        actions=False)
```

Da Twitter a Reddit

Il primo social media utilizzato è stato Twitter; in seguito, tuttavia, si è optato per Reddit.
Il motivo di questo cambiamento risiede principalmente nell'esiguo numero di tweet associati ad un codice di borsa (o al nome esteso di una compagnia quotata).

Sull'altra piattaforma, invece, vi sono diverse subreddit specifiche per discussioni di trading; inoltre è "convenzione" intitolare i post secondo il codice di borsa relativo alla compagnia di cui si vuole discutere.

Download post da Reddit

Per ottenere i post da Reddit si è fatto affidamento ad un wrapper di Pushshift [3] API chiamato "pmaw".
Pushshift.io è un progetto di big-data storage sviluppato da Jason Baumgartner che funziona come repository di elementi postati su Reddit; essi vengono copiati nel momento stesso in cui sono pubblicati sulla piattaforma.
Il motivo per cui è stato preferito pmaw rispetto all'API ufficiale di Reddit è dato dall'uso del multithreading da parte del wrapper.
Esso infatti consente di scaricare, più rapidamente rispetto all'API ufficiale, enormi quantità di post.

pmaw ha un approccio minimalista e per tale motivo risulta relativamente semplice da usare.
Tramite la funzione "***search_submissions"*** si è in grado di scaricare post da reddit che possono essere direttamente salvati in un dataframe object di Pandas.

è inoltre possibile definire parametri in grado di personalizzare al meglio la tipologia di post da scaricare. Nel nostro caso abbiamo curato la definizione dei seguenti parametri:

- *after* → indicazione temporale che stabilisce il periodo successivamente al quale scaricare i post.
- *before* → come "*after"* ma indica il periodo precedentemente al quale scaricare i post; insieme essi definiscono il nostro orizzonte temporale.
- *subreddit* → indica in quale subreddit eseguire la ricerca e scaricamento dei post. Nel nostro caso è stata limitata a 3 subreddit specifiche "r/Wallstreetbets", "r/Stocks" e "r/Investing".
- *filter* → parametro per filtrare le variabili ottenute dalla ricerca ovvero le variabili che saranno a tutti gli effetti le colonne del dataframe. Abbiamo quindi tenuto i valori più rilevanti quali l'id del post, il nome dell'autore, la data di creazione, il numero di commenti, la subreddit in cui è stato pubblicato, il titolo ed infine il testo del post.
- *limit* → parametro usato in fase di "test" del wrapper in quanto permette di fissare un numero massimo di post da scaricare.
- *title* → essenzialmente una query dei titoli dei post da scaricare; se il contenuto della query è presente nel titolo del post allora esso viene scaricato.

Sfortunatamente il parametro “title”, che sembrava essere potenzialmente il più utile, sia in termine di tempo, sia di organizzazione del dataset, si è rivelato problematico.

Il parametro infatti non accetta variabili multidimensionali quali array o liste e per questo motivo è necessario l’uso di cicli for per cercare singolarmente ogni ticker contenuto nel csv descritto nei paragrafi precedenti.

Questa attività risulta essere particolarmente dispendiosa in termini di tempo di esecuzione.
L’alternativa, che abbiamo verificato essere più veloce, è l’utilizzo di un “filtro” dei dati post-download.

Filtro dei post - assegnazione ticker

In primo luogo abbiamo scaricato TUTTI i post provenienti dalle 3 subreddit citate nel primo trimestre del 2022 ovvero da Gennaio a Marzo.
Questo ha portato al download di più di 73.000 elementi.

author	created_utc	id	num_comments	permalink	selftext	subreddit	title	date	time
upbstock	2022-02-01 12:59:41	shudnp	0	/r/Optionmillionaires/comments/shudnp/stock_op...	NaN	Optionmillionaires	stock option open interest changes $TLRY $INTC...	2022-02-01	12:59:41
TheWizzr	2022-02-01 09:14:51	shqqi3	0	/r/StockInvest/comments/shqqi3/aapl_price_pred...	NaN	StockInvest	AAPL Price Predictions - Apple Inc. Stock Anal...	2022-02-01	09:14:51
ShortAlgo	2022-02-01 11:00:11	shsbqr	0	/r/UltraAlgo/comments/shsbqr/aapl_look_at_this...	NaN	UltraAlgo	$AAPL Look at this! 8 Trades executed, trade P...	2022-02-01	11:00:11
ShortAlgo	2022-02-01 12:04:43	shtf5o	0	/r/UltraAlgo/comments/shtf5o/aapl_waiting_for_...	NaN	UltraAlgo	$AAPL Waiting for short signal https://t.co/cH...	2022-02-01	12:04:43
ShortAlgo	2022-02-01 12:18:59	shtnq2	0	/r/UltraAlgo/comments/shtnq2/aapl_waiting_for_...	NaN	UltraAlgo	$AAPL Waiting for short signal https://t.co/cH...	2022-02-01	12:18:59

Si può notare come alcuni post presentino 0 commenti e non abbiamo testo; questo può ricondursi a molteplici fattori:

1) Molti post vengono cancellati poco dopo essere stati pubblicati nel caso in cui non vengano approvati dai moderatori della subreddit
2) Pushshift a volte perde la connessione con i server di Reddit e non riesce a scaricare alcuni elementi

3) Il post è stato copiato su Pushshift prima che potesse ricevere commenti

In seguito questi dati sono stati filtrati in modo da conservare solo i post che comprendessero almeno uno dei ticker presenti nel nostro .csv di riferimento.

Per fare ciò è stata utilizzata una funzione che, preso in ingresso l'array di simboli e il dataset "grezzo", producesse un subset di quest'ultimo tramite una "maschera". Essa, all'interno di un ciclo, passa in rassegna ogni simbolo, verificandone la presenza nel titolo del post.
Oltre a questo, se la presenza del simbolo è verificata, la funzione assegna ad una nuova variabile "symbol" il suo ticker.

Il risultato è stato salvato in un file chiamato ***Jan_Mar_2022_tickers.csv:***

author	created_utc	id	num_comments	permalink	selftext	subreddit	title	date	time	symbol
Vencero_JG	2022-03-23 19:18:26	tl8w5y	1	/r/wallstreetbets/comments/tl8w5y/i_hear_tendi...	[removed]	wallstreetbets	I HEAR Tendies Comin!	2022-03-23	19:18:26	HEAR
Vencero_JG	2022-03-23 19:39:09	tl9ufh	0	/r/wallstreetbets/comments/tl9ufh/i_hear_tendi...	So Turtle Beach Corp. (HEAR) makes some really...	wallstreetbets	I HEAR Tendies Comin!!	2022-03-23	19:39:09	HEAR
streetgainer	2022-03-25 18:50:40	tnw928	1	/r/wallstreetbets/comments/tnw928/where_did_th...	[removed]	wallstreetbets	Where did that $HEAR volume come from all of a...	2022-03-25	18:50:40	HEAR
SergeantGrimm	2022-02-02 00:09:39	si9w8m	1	/r/wallstreetbets/comments/si9w8m/legit_questi...	[removed]	wallstreetbets	Legit question, how can after hours volume dis...	2022-02-02	00:09:39	NXPI

Problematiche relative al nuovo dataset

L'attività di filtraggio e assegnazione dei ticker ha generato un dataset di circa 32.700 elementi.
Tale risultato, considerando il filtro applicato, non dovrebbe sorprendere in quanto vi sono diversi post nei quali vengono discussi più di 1 simbolo alla volta e per questo motivo essi vengono "duplicati", uno per ogni simbolo discusso.

Questo ha però portato anche alla creazione di dati "errati".
Il filtro, nel best case scenario, duplica quei post che contengono più simboli e gli assegna correttamente il ticker per cui è stato selezionato.
In fase poi di integrazione, questo consentirà di matchare correttamente i simboli dei post con i relativi dati di borsa.
Nel worst case scenario, tuttavia, riconosce simboli che appartengono a parole/verbi/parti di frasi non inerenti a titoli azionari.

$OPEN versus $Z	2022-01-13	10:27:12	OPEN
MARKET OPEN HYPE!	2022-01-14	11:29:05	OPEN
Markets are OPEN!	2022-01-17	15:29:21	OPEN

L'immagine riporta uno dei molteplici casi di riconoscimento errato. La prima riga riconosce correttamente $OPEN come ticker mentre le altre due non hanno niente a che fare con il simbolo azionario.

Una possibile limitazione di tale problema potrebbe essere cercare i titoli con i simboli preceduti dal dollaro americano.
Tuttavia questa soluzione limiterebbe di molto il dataset finale in quanto sono relativamente poche le volte in cui viene adottata tale "nomenclatura" dagli utenti di Reddit.

Non trovando una soluzione migliore in merito abbiamo optato per utilizzare comunque questo dataset.

EDA - Reddit posts

```
RangeIndex: 32779 entries, 0 to 32778
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   author        32779 non-null  object
 1   created_utc   32779 non-null  object
 2   id            32779 non-null  object
 3   num_comments  32779 non-null  int64
 4   permalink     32779 non-null  object
 5   selftext      21658 non-null  object
 6   subreddit     32779 non-null  object
 7   title         32779 non-null  object
 8   date          32779 non-null  object
 9   time          32779 non-null  object
 10  symbol        32779 non-null  object
dtypes: int64(1), object(10)
```

Non vi sono valori nulli e l'unico valore numerico riguarda il numero di commenti.

Il motivo dei valori mancanti della variabile selftext sono stati già evidenziati precedentemente.

```
subreddit
investing          2124
stocks             4276
wallstreetbets    26379
```

Notiamo anche come la maggior parte dei post provengano dalla subreddit "wallstreetbets" che è anche comprensibile data la recente popolarizzazione che l'ha portata ad avere più di 10 milioni di utenti.

	symbol	count
1214	V	3961
737	ME	2613
869	ON	1222
1088	SP	1124
1311	Z	1031
1145	TH	910
349	EA	577
1125	TA	523
665	LE	504
1175	TSLA	469

Questo risultato dovrebbe indicare i simboli più discussi ma è sicuramente falsato dai problemi evidenziati in precedenza.
Ad esempio ME molto probabilmente si riferisce anche ai post che trattano di GME e così via.

Per questo motivo, data la popolarità dello stock ancora adesso e dato l'errore nel riconoscimento di quest'ultimo, abbiamo cambiato il simbolo dei post etichettati come ME in post etichettati come GME.

Dataset importato in MongoDB

I dati provenienti dai due flussi di principali (post di Reddit e valori di borsa associati ai tickers) sono stati caricati nel database documentale MongoDB.
La motivazione principale della scelta di questo tipo di database deriva soprattutto dalla natura dei dati che si è deciso di caricare. In particolare i post di Reddit associati ai valori di borsa ne hanno orientato la scelta.
Di seguito un esempio dei dati caricati nel database.

…
"2022-01-28": {
"Open": 831.5599975585938,
...
"Volume": 44929700,
"post": {
"author": "merlinsbeers",
...
"num_comments": "6",
"permalink": "/r/stocks/comments/sey18y/the_future_value_of_tsla_is_not_computable/",
"subreddit": "stocks",
"title": "The future value of TSLA is not computable.",
...
"symbol": "TSLA",
"text": "You can't justify calculating TSLA's future value from the current numbers.
You have to be expecting growth in sectors
and markets not yet accessed by the company..."
}
…

Nella generazione del database si è partiti dalla lista di riferimento dei tickers. I 4459 item presenti nella lista sono stati utilizzati per creare i 4459 documenti e popolare la collezione "***collect_local***".

Integrazione dei dataset

Dopo la creazione dei documenti presenti nella collezione si è passati alle due fasi di arricchimento dei dati.
La prima fase è consistita nel caricare i valori di borsa su base giornaliera e relativa alla finestra temporale del primo trimestre del 2022.
In particolare i dati aggiunti sono relativi ai valori in dollari di apertura, chiusura, minimo e massimo ed anche al volume delle azioni scambiate del giorno in analisi.
Successivamente si è passati ad un ulteriore arricchimento aggiungendo i post comparsi in un determinato giorno e relativi ad uno specifico titolo di borsa.
Tutte le operazioni, dalla creazione del database, manipolazioni, fino all' arricchimento dei dati, sono avvenute utilizzando la libreria ***pymongo***.
Il file finale ottenuto dopo l'arricchimento è il seguente:
collezione_arrichita_valori_post_reddit.json

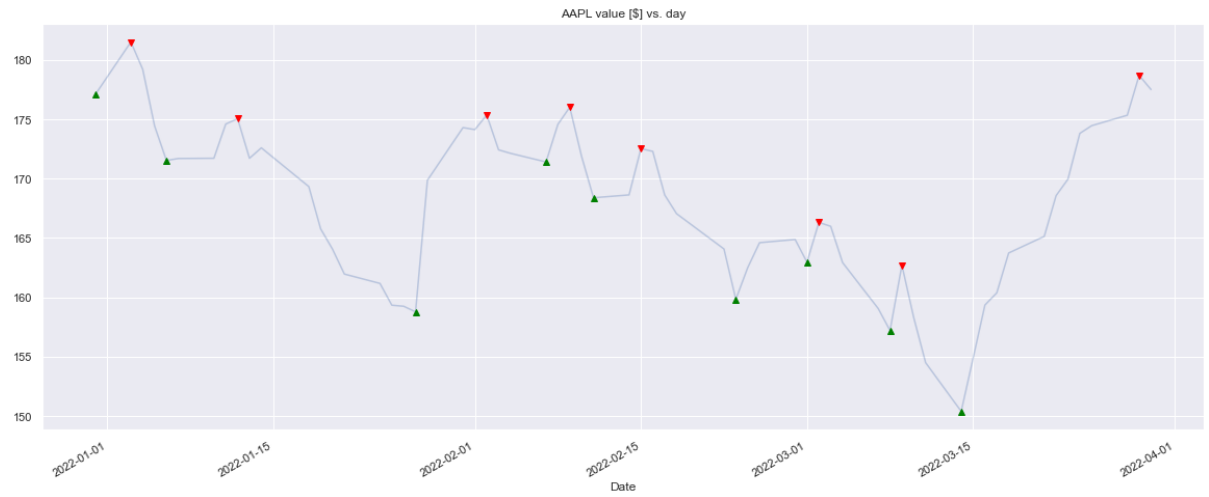
Possibili sviluppi futuri: correlazione tra post e valori di borsa di un titolo

Al fine di monitorare la correlazione tra i post ed i valori sono stati analizzati separatamente i due flussi di dati (post di Reddit e valori di borsa associati ai tickers). L'analisi dei valori di borsa si è rivelata più semplice rispetto al flusso Reddit poiché si è cominciato ad analizzare una serie storica riferita a valori di chiusura ("Close") di un titolo azionario.

Trend dei valori di borsa

Al fine di individuare i picchi positivi e negativi all'interno di una finestra temporale è stata utilizzata la funzione ***argrelextrema*** di ***scipy.signal*** [4].
Questa funzione ha permesso di valutare i minimi ed i massimi relativi utilizzando i valori di una finestra mobile estratti dall' API **yfinance** ; l'estensione della finestra mobile, scelta per analizzare i dati, è una settimana pari al valore di "***order=3"***.
Grazie ad essa è stato possibile cominciare ad impostare il codice per estrarre due vettori, uno con i minimi ed uno con i massimi in previsione di un confronto con le informazioni provenienti dai post di Reddit.

Di seguito l'andamento del valore di borsa a fine giornata del titolo Apple (ticker "AAPL") sulla finestra di analisi presa in considerazione in tutto il presente lavoro (Q1 2022)



Trend dei post di Reddit

Uno dei possibili sviluppi del progetto potrebbe essere tracciare l'andamento dei dati di borsa in funzione della pubblicazione dei post più influenti.
Pushshift ha annunciato come in futuro i metadati dei post verranno aggiornati periodicamente facilitando dunque la ricerca dei post più commentati e votati.

Non solo, si potrebbe monitorare il numero di post pubblicati giornalmente in ciascuna subreddit per determinare quale sia più attiva in merito a determinati ticker.

Riteniamo comunque necessario un lavoro più preciso nella fase di identificazione dei ticker dai titoli dei post, magari implementando l'utilizzo di text mining o anche text recognition per poter scartare eventuali errori di selezione.

Riferimenti

[1] https://www.nasdaq.com/market-activity/stocks/screener

[2]https://medium.com/nerd-for-tech/all-you-need-to-know-about-yfinance-yahoo-finance-library-fa4c6e48f08e

[3] GitHub - pushshift/api: Pushshift API

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.argrelextrema.html