Data Analytics - Relazione progetto

Elaborato sviluppato da

- **Volpato Mattia**
  - Matricola 866316
  - m.volpato4@campus.unimib.it
- **Preziosa Alessandro**
  - Matricola 866142
  - a.preziosa1@campus.unimib.it

1. Descrizione del dominio di riferimento e obiettivi dell'elaborato

Descrizione del dominio

Il dataset di riferimento consiste in una raccolta di più di 17000 tweets ottenuti tramite *scraping,* postati da presunti simpatizzanti dell'ISIS a partire dall'**attacco di Parigi** del 13/11/2015. Gli autori dei tweets sono utenti proveniente da tutto il mondo e, di conseguenza, alcuni tweet non sono in lingua inglese (e in certi casi non utilizzano neanche lettere dell'alfabeto inglese). I tweets raccolti coprono un intervallo temporale di circa un anno e non vanno oltre il 13/05/2016; maggiori informazioni vengono riportate nella sezione relativa all'esplorazione del dataset.

Obiettivi dell'elaborato

L'**obiettivo primario** che ci si pone è l'analisi della **rete** che va a costruirsi sulle **citazioni tra utenti** mediante *tagging*, al fine di determinare gli utenti più *importanti*, dove il termine *importanza* verrà definito successivamente in maniera precisa; in aggiunta a ciò, si è anche provveduto a estrarre le **entità** riportate nei tweets e il **sentiment** a loro associato, con due finalità:

- determinare quali siano le entità più *importanti* e quanto siano *polarizzanti* rispetto al sentiment suscitato negli utenti;
- determinare se gli **hashtags** possano essere utilizzati come una buona rappresentazione delle entità presenti nei tweets.

Inoltre, come **obiettivo secondario** abbiamo deciso di svolgere un'analisi temporale del sentiment espresso dai tweet di alcuni specifici **utenti rilevanti** e l'analisi temporale del sentiment associato ad alcune **entità rilevanti**. In caso di bruschi cambiamenti, siamo andati a determinare gli eventi che hanno portato a ciò.

Riassumendo:

- L'**obiettivo principale** è l'individuazione degli utenti e delle **entità** più importanti, la ricerca di eventuali relazioni tra l'**importanza di un'entità** e la **polarità** del **sentiment** che provocano nei tweets e lo stabilire se gli **hashtags** siano una buona rappresentazione delle entità;

- Individuate le entità più importanti, l'**obiettivo secondario** si compone dell'analisi dell'andamento del **sentiment a loro legato** nel corso del tempo.

2. Esplorazione del dataset

È possibile scaricare il dataset da questo link.

Struttura del dataset

Il dataset si presenta come una tabella con la seguente struttura:

name	username	description	location	followers	numberstatuses	time	tweets
str	*str*	*str*	*str*	*int*	*int*	*date*	*str*

Dove:

- L'attributo **description** rappresenta una breve descrizione dell'utente che ha postato il relativo tweet;
- L'attributo **location** rappresenta il luogo da cui è stato postato il tweet;
- L'attributo **numberstatuses** indica il numero di tweets postati dall'utente fino a quel momento.

A seguire un'*analisi uni variata* degli attributi più rilevanti.

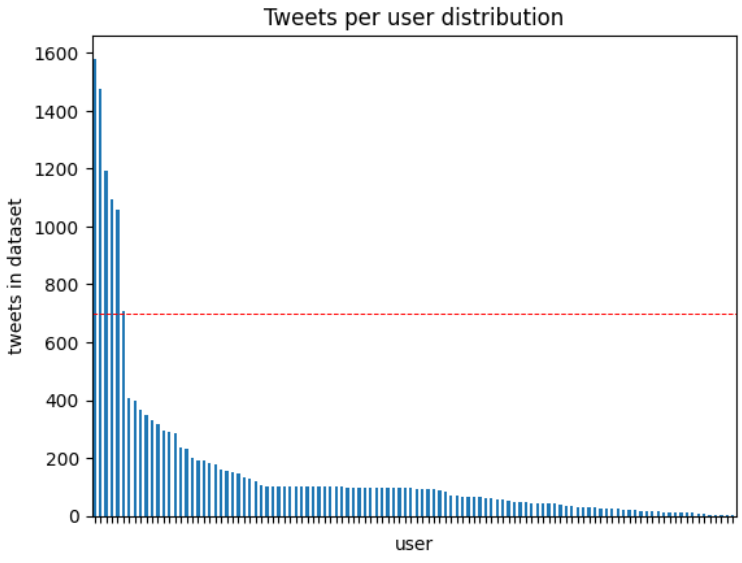
Analisi degli attributi

Utenti con più tweets all'interno del dataset

L'andamento del **numero di tweets** raccolti ha natura esponenziale, il che permette di individuare solamente **6 utenti** con più di **700 tweets**, a cui appartengono circa il $\mathbf{41\%}$ **dei tweets totali**, riportati nella tabella sottostante. Ci si aspetta che questi utenti siano elementi centrali della **rete delle citazioni** (sezione 4Grafo delle citazioni tra utenti del dataset).

Per gli attributi **location, followers** e **numberstatuses** (che possono variare nel tempo) sono stati considerati i valori aggregati:

- **location:** moda;
- **followers:** valore medio;
- **numberstatuses:** valore massimo;



Distribuzione del numero di tweets raccolti per utente

name	username	location	followers	numberstatuses	tweets	tweets %
Obi-Wan Al Coconuty	**Uncle_SamCoco**	Texas, USA	1572	4328	1580	9.08 %
Rami	**RamiAlLolah**	Unknown	31796	17411	1475	8.47 %

WarBreakingNews	**warrnews**	World	7226	6735	1191	6.84 %
Conflict Reporter	**WarReporter1**	Worldwide contributions	1659	817	1095	6.29 %
Salahuddin Ayubi	**mobi_ayubi**	Unknown	854	7555	1056	6.07 %
Ibni Haneefah	**_IshfaqAhmad**	Muslim community of Kashmir	1511	12981	709	4.07 %

Si nota subito un'inconsistenza per l'attributo **numberstatuses** (numero di tweets totali dell'utente), che chiaramente non può essere inferiore al numero di tweet raccolti nel dataset: per questo motivo non verrà considerato nel calcolo dell'*importanza* degli utenti.

Descrizione di questi sei utenti:

- **Uncle_SamCoco:**
  - **description:** here to defend the American freedom and also the freedom of coconut, Cat Lover or Hater. Kebab Fan. We're all living in America, America ist wunderbar #USA
  - **commento:** residente negli USA (Texas), ironizza sul concetto di *libertà americana*: con ogni probabilità un vero simpatizzante dell'ISIS.
- **RamiAlLolah:**
  - **description:** Real-Time News, Exclusives, Intelligence & Classified Information/Reports from the ME. Forecasted many Israeli strikes in Syria/Lebanon. Graphic content.
  - **commento:** account che riporta notizie, non necessariamente pro-ISIS. ME indica 'Middle East'.
- **warrnews:**
  - **description:** we provide fresh news from every battlefield
  - **commento:** account che riporta notizie, non necessariamente pro-ISIS.
  -
- **WarReporter1:**
  - **description:** reporting on conflicts in the MENA and Asia regions. Not affiliated to any group or movement
  - **commento:** account che riporta notizie, non necessariamente pro-ISIS. MENA indica 'Middle East and North Africa'.
- **mobi_ayubi:**
  - **description:** Journalist, specialize in ongoing war against terrorism. Retweet is not endorsement.
  - **commento:** giornalista specializzato sulla guerra al terrorismo: ci si aspetta sia anti-ISIS.
- **_IshfaqAhmad:**
  - **description:** Medico at GMC Srinagar, Pro-Khilafah, Anti-Democratic, Anti-Nationalistic, Anti-Rafidah, Innocent Bystander of the Conflict in Middle East, Cricketist
  - **commento:** medico presso il Government Medical College di Srinagar, in India. Pro-califfato (*Khilafah*), antinazionalista e antidemocratico, avverso per gli sciiti (*Rafidah*).

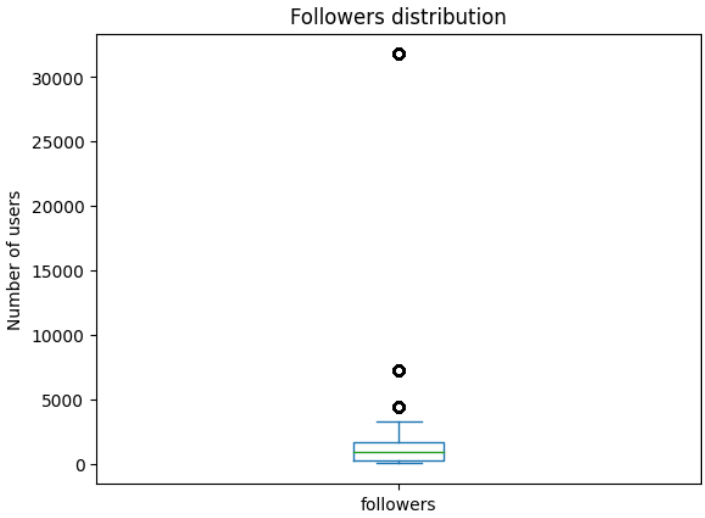
D'ora in poi si farà riferimento agli utenti attraverso il loro **username**.

Numero di followers

Il *boxplot* relativo al **numero di followers** risulta essere molto schiacciato, con la quasi totalità degli utenti che risulta avere meno di 2000 followers e con solo tre utenti con un numero veramente alto di questi:

- **RamiAlLolah:** 31796 followers
- **warrnews:** 7226 followers
- **Nidalgazaui**: 4455 followers

L'utente **Nidalgazaui** non era stato individuato tra gli utenti con più tweets:

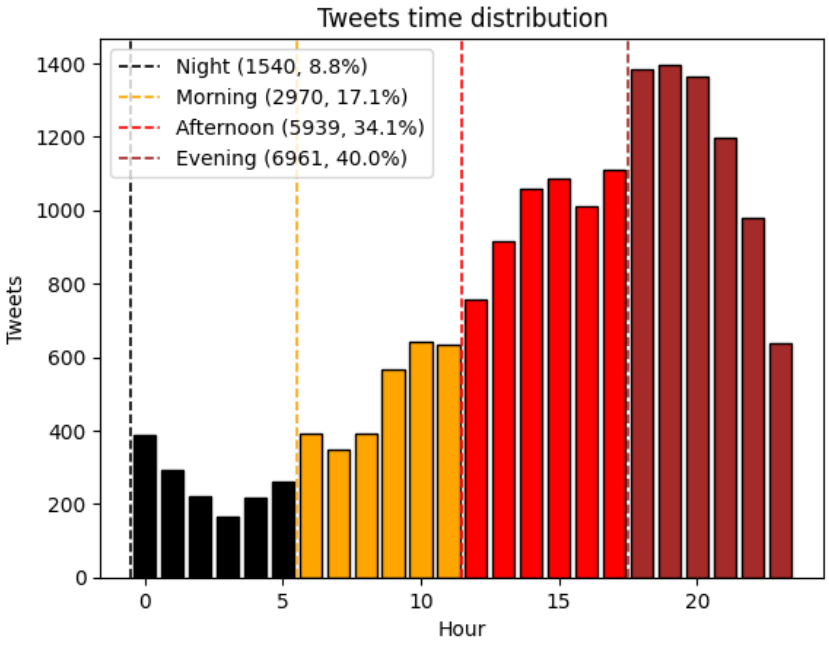


name	username	location	followers	numberstatuses	tweets	tweets %
Nidal	**Nidalgazaui**	Germany	4455	5703	397	2.23 %

Descrizione: 17yr. old Freedom Activist, Correspondence of NGNA, Terror Expert, Middle East Expert. Daily News about Syria, Iraq, Yemen, Russia, Middle East.

Orario dei tweets

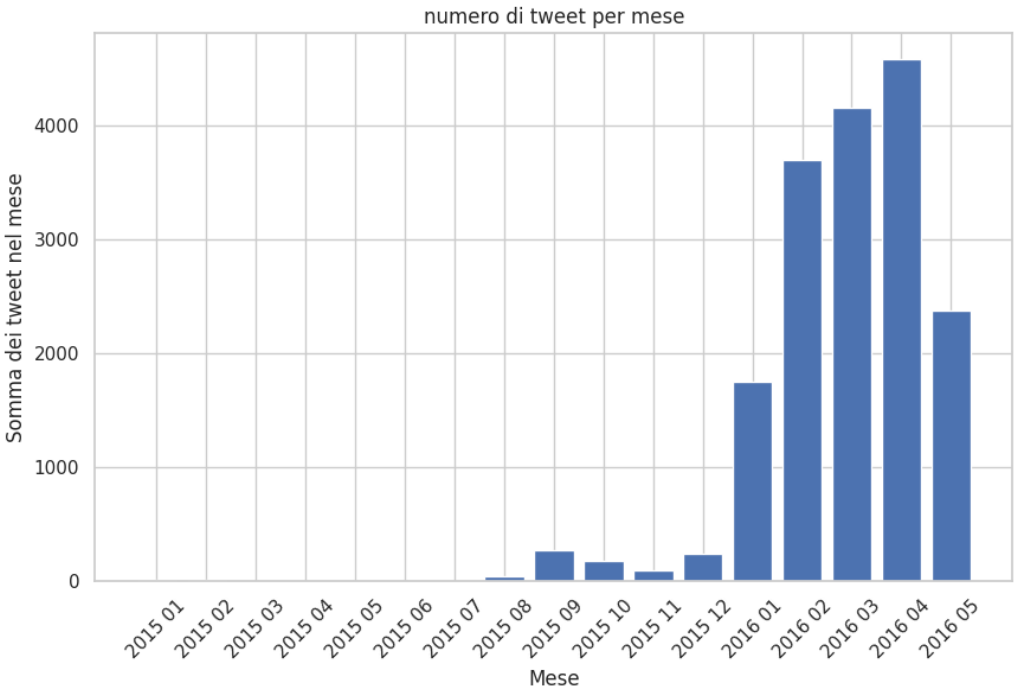
Dal grafico riportato accanto si nota una preferenza degli utenti nel postare tweets di **pomeriggio** e di **sera** (quasi il 75% del totale), con un picco nell'intervallo di tempo che va dalle 18:00 alle 21:00 (25% dei tweets).



Distribuzione dell'orario dei tweets

Andamento temporale dei tweets

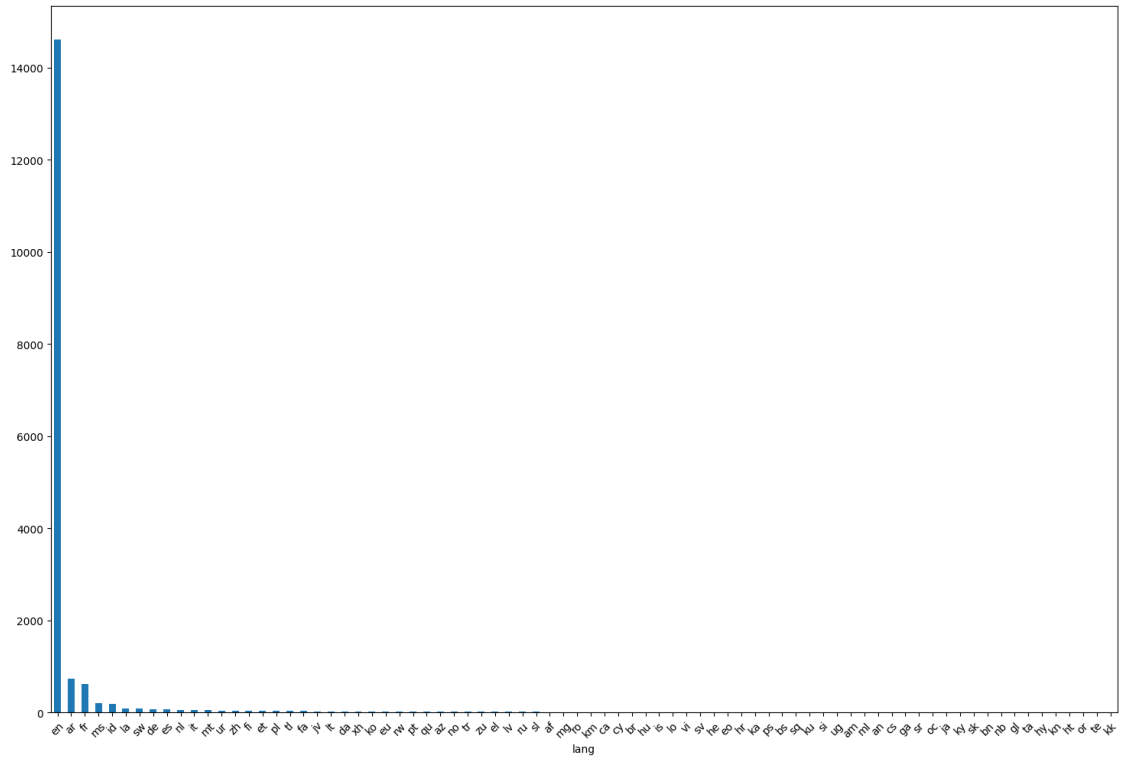
Il numero di tweet raccolti nel dataset per mese non è costante bensì varia. Qui vediamo il conteggio dei tweet mese per mese:



Lingua

Non tutti i tweet sono in lingua inglese. Tramite la libreria ***python*** Langid abbiamo classificato i tweet in base alla lingua con sono più probabilmente scritti.

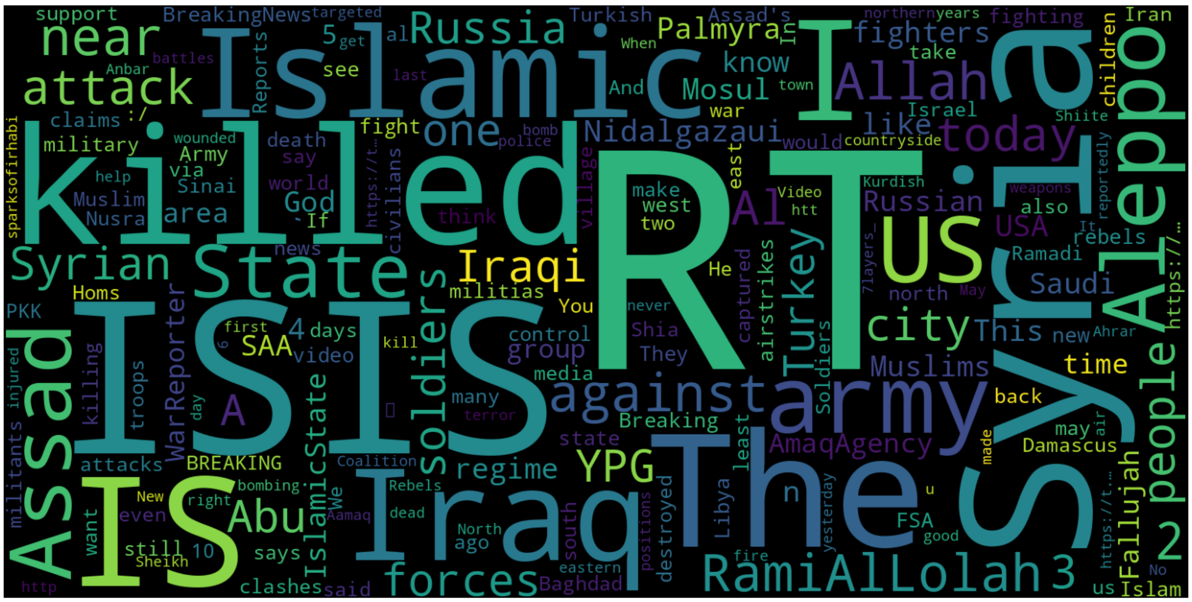
Istogramma delle frequenze delle lingue:



Dopo un veloce **pre-processing** del testo (approfondito nella sezione NER), abbiamo tradotto automaticamente i tweet non in inglese grazie alla libreria googletrans. D'ora in poi saranno analizzati nella parte di ***NLP*** i tweet tradotti e i tweet originalmente in inglese.

Data Visualization

Data la natura del dataset (insieme di tweet e metadati) viene naturale visualizzare la frequenza delle parole con una ***wordcloud***:



Molte *stopword* e segni di punteggiatura sono state rimosse prima di calcolare le frequenze delle parole e visualizzare questa *wordcloud*; in seguito, verrà illustrato come. RT nel gergo di Twitter è un invito a retwittare il post.

3. Grafo delle citazioni tra utenti

Struttura del grafo

A partire dalle **menzioni** effettuate tramite *tagging* contenute nei tweets, è stato costruito un **grafo delle citazioni** tra utenti, in cui due nodi sono in relazione se appunto esiste una citazione in un tweet del nodo sorgente verso il nodo destinatario.

Formalmente, si tratta di un ***grafo pesato e orientato***

$$G = (V, E, \delta)$$

dove:

- $v \in V$ rappresenta lo **username** di un utente;
- $\forall v_1, v_2 \in V: (v_1, v_2) \in E \Leftrightarrow v_1$ **menziona** in un tweet v_2;
- $\delta: E \rightarrow \mathbb{N}$ rappresenta il **numero** di **citazioni** tra due utenti:

$\forall e = (v_1, v_2) \in E: \delta(e) = k \Leftrightarrow k$ è il **numero totale di citazioni** di v_1 verso v_2 (considerando tutti i tweet di v_1).

In particolare, nel grafo saranno presenti due tipi di nodi:

- **Nodi** con un **grado di uscita maggiore di 0**, che rappresentano gli utenti presenti nel dataset;
- **Nodi pozzo**, che rappresentano utenti di *Twitter* citati nei tweets, ma non presenti nel dataset.

Statistiche del **grafo delle citazioni**:

$\|V\|$	$\|E\|$	$v \in dataset$	$v \notin dataset$	*Freeman's centrality*	$< v >$	*Average path length*	*Diametro*	*Densità*	*Reciprocità*
3334	5368	112 (3.4%)	3222 (97.6%)	0.0146	1.61	3.46	9	0.00048	0.0093

Osservando questi numeri si nota come la maggioranza dei nodi siano utenti non appartenenti al dataset; di conseguenza, anche la maggioranza delle **menzioni** sarà rivolta verso questo tipo di nodi, che però non potranno essere usate al fine di calcolare *l'importanza* degli utenti e, conseguentemente, delle **entità**. Inoltre, dato il grande numero di nodi pozzo, anche il *coefficiente di centralità di Freeman* indica l'assenza di utenti veramente centrali nella rete. Nel continuo delle analisi si è deciso quindi di procedere con la rimozione degli utenti non appartenente al dataset.

Tuttavia, prima di effettuare la rimozione di questa tipologia di nodi, si procede con un'analisi più approfondita sulla **distribuzione del grado**, che distingua tra **ingresso** e **uscita** e tenga conto anche del **peso** degli archi.

Distribuzione del grado pesato dei nodi

Si definisce ***grado pesato (entrante o uscente)*** di un nodo come la *somma* dei *pesi degli archi* (entranti o uscenti):

$$in\_strength: \quad s_{in}(v) = \sum_{e=(w,v)\in E} \delta(e), \quad \forall v \in V$$

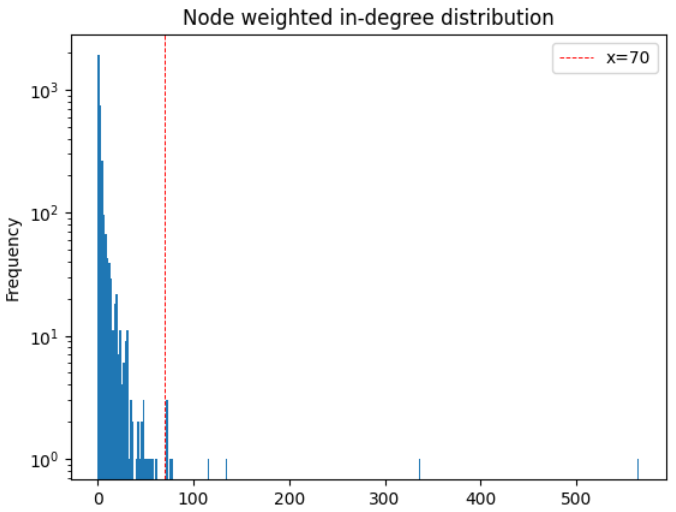
$$out\_strength: \quad s_{out}(v) = \sum_{e=(v,w)\in E} \delta(e), \quad \forall v \in V$$

$$strength: \quad s(v) = s_{in}(v) + s_{out}(v), \quad \forall v \in V$$

Grado pesato in ingresso (numero di menzioni ricevute)

La distribuzione del *grado pesato in ingresso* mostra la presenza di un numero limitato di ***hubs,*** corrispondenti ai seguenti utenti

username	followers	$s_{in}(v)$	$s_{out}(v)$
RamiAlLolah	31796	565	647
Nidalgazau	4455	336	89
WarReporter1	1659	135	592
7layers_	-	116	0
ScotsmanInfidel	-	79	0
sparksofirhabi3	-	76	0
MaghrebiQM	1559	72	96
Conflicts	-	72	0
DidyouknowVS	-	72	0



Grado pesato in ingresso dei nodi (scala logaritmica sull'asse y)

Tra i utenti presenti nel dataset, ne ritroviamo tre (RamiAlLolah, Nidalgazau, WarReporter1) già incontrati come utenti con il numero maggiore di tweet e followers; al contrario, **MaghrebiQM** è un utente con *un numero limitato followers* che non è ancora stato incontrato:

name	username	location	followers	numberstatuses	tweets	tweets %
Maghrebi	**MaghrebiQM**	Unknown	1559	343	158	0.91 %

Descrizione: A Moroccan engineer in nanotechnology commenting on Middle East politics, mainly Iraq/Syria. Not representing any group. RTs and Favs aren't endorsements.

Considerando invece i cinque utenti non appartenenti al dataset, l'unico account a risultare ancora attivo è Conflicts, il quale tratta notizie di guerra.

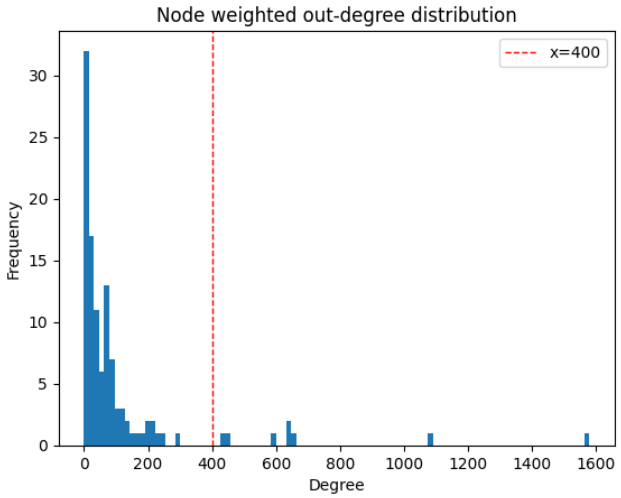
Infine, si riportano i valori medi dei **gradi pesati in ingresso (numero medio di menzioni ricevute):**

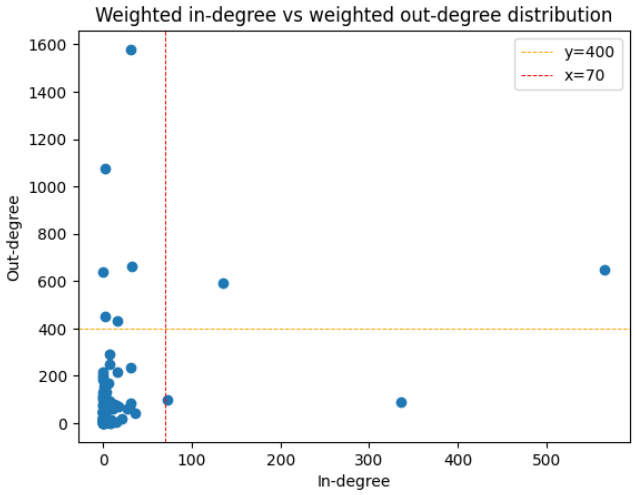
- Complessivo: $< s_{in} >= 3.59$;
- Utenti presenti nel dataset: $< s_{in} >= 14.46$;
- Utenti non presenti nel dataset: $< s_{in} >= 3.22$.

Grado pesato in uscita (numero di menzioni fatte)

La distribuzione del *grado pesato in uscita* (che rappresenta quanto un utente è attivo) mostra la presenza di un numero limitato di ***hubs***, corrispondenti ai seguenti utenti

username	followers	$s_{in}(v)$	$s_{out}(v)$
Uncle_SamCoco	1572	31	1579
mobi_ayubi	854	2	1078
warrnews	7266	32	660
RamiAlLolah	31796	565	647
melvynlion	58	0	638
WarReporter1	1659	135	592
MaghrabiArabi	225	3	450
_IshfaqAhmad	1511	16	432





È possibile notare immediatamente come non ci sia necessariamente *correlazione* tra numero di **menzioni ricevute** (*importanza nella rete*) e numero di **menzioni fatte** (*attività nella rete*), come confermato dallo *scatterplot accanto*.

La correlazione tra **grado pesato in ingresso** e **in uscita** è: 0.2818.

Infine, si riporta il valore medio del **grado pesato in uscita (numero medio di menzioni fatte):**

- Utenti presenti nel dataset:
  $< s_{out} >= 106.97$

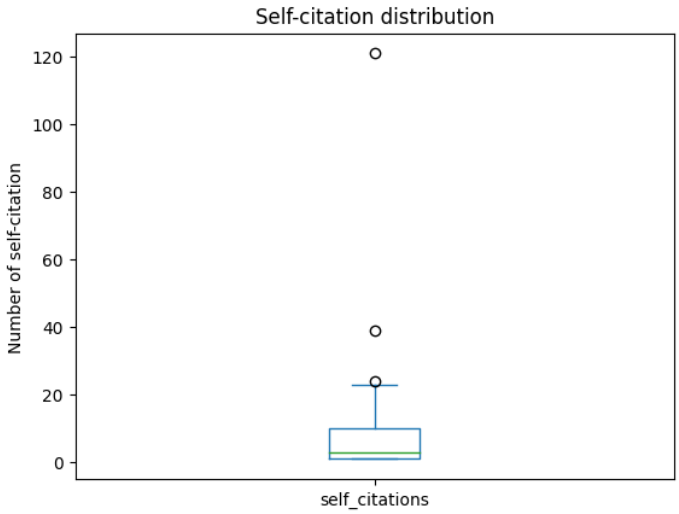
Auto-menzioni

Un'usanza inaspettata che abbiamo trovato molto utilizzata è quella delle **auto-menzioni,** ovvero la pratica di *taggare* sé stessi all'interno di un proprio tweet.

L'obiettivo è probabilmente quello di aumentare la visibilità del proprio account sfruttando il funzionamento del sistema di *tagging* di Twitter; tuttavia, per i nostri scopi non ha senso che un nodo possa attribuirsi importanza da solo: di conseguenza, le **auto-menzioni** (rappresentate da *self-loop* all'interno del grafo) sono state rimosse (e non appaiono nelle statistiche riportate in precedenza).

Gli utenti con più auto-citazioni sono:

username	Auto-citazioni
WarReporter1	121
Uncle_SamCoco	39
Fidaee_Fulaani	24
ismailmahsud	23
Maisaraghereeb	21



4. Grafo delle citazioni tra utenti del dataset

A questo punto ci siamo concentrati unicamente sugli utenti presenti nel dataset, estraendo il **sottografo indotto** a partire dai **nodi rappresentanti utenti presenti nel dataset**. È emerso un grafo *non connesso* composto da una serie di nodi singoli disconessi e da una **componente gigante,** di cui si riportano le statistiche:

$\|V\|$	$\|E\|$	*Freeman's Centrality*	$< v >$	*Average path length*	*Diametro*	*Densità*	*Reciprocità*
88	303	0.548	3.44	3.17	8	0.04	0.158

Nonostante il ridotto numero di nodi, le misure di *average path length* e *diametro* restano molto simili al grafo precedente; al contrario le misure di *densità* e *reciprocità* aumentano di molto, al punto che il 16% degli utenti si **menziona** in maniera reciproca in uno o più tweets. Anche il *coefficiente di Freeman* diventa significativo, segnalando la presenza di un **nodo molto centrale** nella rete.

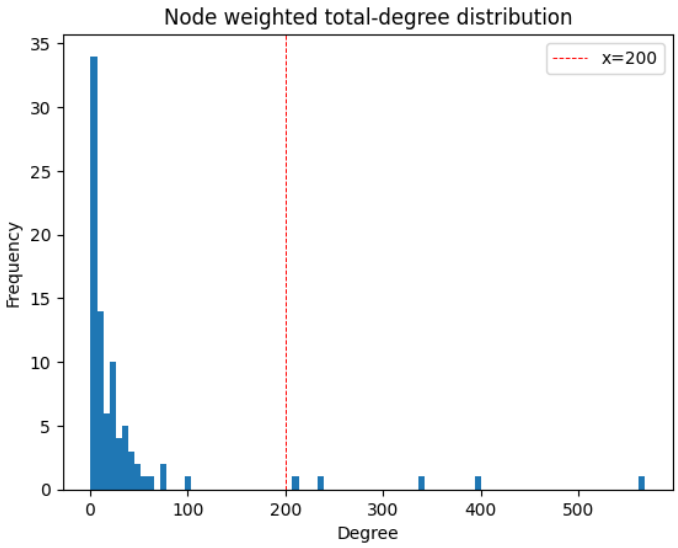
Distribuzione del grado pesato dei nodi

Si riportano in questa sezione le più importanti misure di centralità ottenute dal grafo, che verranno poi utilizzate nel calcolo dell'importanza di un utente.

Grado pesato totale

L'effetto che si ottiene dall'eliminazione dei nodi pozzo è la forte diminuzione del **grado pesato in uscita** di alcuni degli ***hubs;*** tuttavia, questo permette di catturare a pieno l'importanza dei pochi utenti rimasti citati da essi, il che giustifica la nostra scelta di ridurre il grafo.

username	followers	$s_{in}(v)$	$s_{out}(v)$	$s(v)$
RamiAlLolah	31796	565	3	568
mobi_ayubi	854	2	393	395
Nidalgazau	4455	336	3	339
warrnews	7266	32	203	235
WarReporter1	1659	135	74	209



Grado pesato totale dei nodi

Il **grado pesato totale medio** (citazioni medie ricevute/fatte per utente) del grafo è: $< s > = 36.7$.

Importanza di un utente

Al fine di attribuire un indice di **importanza** a ogni utente, abbiamo cercato una misura che tenesse in considerazione tutte le *metriche di centralità*, ma non necessariamente con lo stesso peso. Infatti:

- Siamo **molto interessati** alla capacità di *diffondere informazioni rapidamente* attraverso la rete, misurata dalla ***closeness centrality***;
- Siamo **molto interessati** alla *qualità delle menzioni,* ovvero dall'essere menzionati da nodi importanti, misura ottenibile dalla ***eigenvector centrality***;
- Siamo **meno interessati** alla capacità di fungere da ponte tra comunità diverse, misurata dalla ***betweenness centrality***, in quanto consideriamo il nostro grafo come una comunità unica, come conseguenza di quanto da emerso dalla ***community detection***, che non ha permesso di individuare clusters significativi;
- Non vengono considerate le misure dei **gradi in ingresso e in uscita,** in quanto sarebbero poco significative rispetto a quanto già ottenuto mediante la ***eigenvector centrality.***

Inoltre, abbiamo ritenuto importante anche la capacità di **raggiungere utenti** *al di fuori* della rete che stiamo trattando (al fine di reclutarne altri, per esempio); perciò, nel calcolo dell'importanza è stato considerato anche il **numero di followers** di un utente, valore indipendente dalla struttura della rete trattata. Per le misure di **betweenness** e **closeness** sono state usate le versioni pesate.

Formalmente, l'importanza è stata definita come la **somma pesata** delle metriche elencate, secondo i seguenti pesi:

$$\boldsymbol{importance}(v) = \boldsymbol{followers}(v) + \boldsymbol{closeness}(v) + \boldsymbol{eigenvector}(v) + \frac{1}{2} \cdot \boldsymbol{betweenness}(v), \quad \forall \mathbf{v} \in \mathbf{V}$$

dove le metriche considerate sono state tutte *normalizzate* nell'intervallo $[0, 1]$. Si ottiene così un coefficiente di importanza per ogni utente, che viene a sua volta normalizzato nell'intervallo $[0, 1]$, ottenendo il seguente ranking dei dieci utenti **più rilevanti**:

Rank	**username**	followers	$< s_{in} >$	$< s_{out} >$	$< s >$	eigenvector	betweennes	closeness	**importance**
1	RamiAlLolah	31796	565	3	568	0.892	0.573	1	0.908
2	Nidalgazaui	4455	336	3	339	1	0.106	0.97	0.618
3	Uncle_SamCoco	1572	31	70	101	0.5	1	0.978	0.579
4	warrnews	7266	32	203	235	0.609	0.17	0.865	0.511
5	WarReporter1	1659	135	74	209	0.404	0.304	0.894	0.429
6	__alfresco__	542	16	47	63	0.511	0.161	0.879	0.425
7	NaseemAhmed50	2157	7	45	52	0.189	0.425	0.923	0.398
8	st3erer	986	8	11	19	0.304	0.233	0.926	0.393
9	wayf44rerr	1093	31	11	42	0.36	0.164	0.712	0.34
10	MagherebiQM	1559	72	4	76	0.31	0.043	0.799	0.337

Nel ranking ottenuto ritroviamo molti degli utenti che già incontrati nella sezione □, il che rispecchia le *aspettative* che avevamo a partire dalle analisi fatte in precedenza e giustifica le *scelte* adottate sulle **metriche di centralità**.

Visualizzazione del grafo

Date le dimensioni contenute del grafo ottenuto, ne riportiamo una visualizzazione, nella quale:

- La **dimensione** dei nodi è determinata dalla loro *importanza*;
- Lo **spessore degli archi** è determinato dal numero di citazioni che rappresenta;
- I cinque utenti **più importanti** sono riportati in rosso.

Osservando il *plot* è possibile confermare:

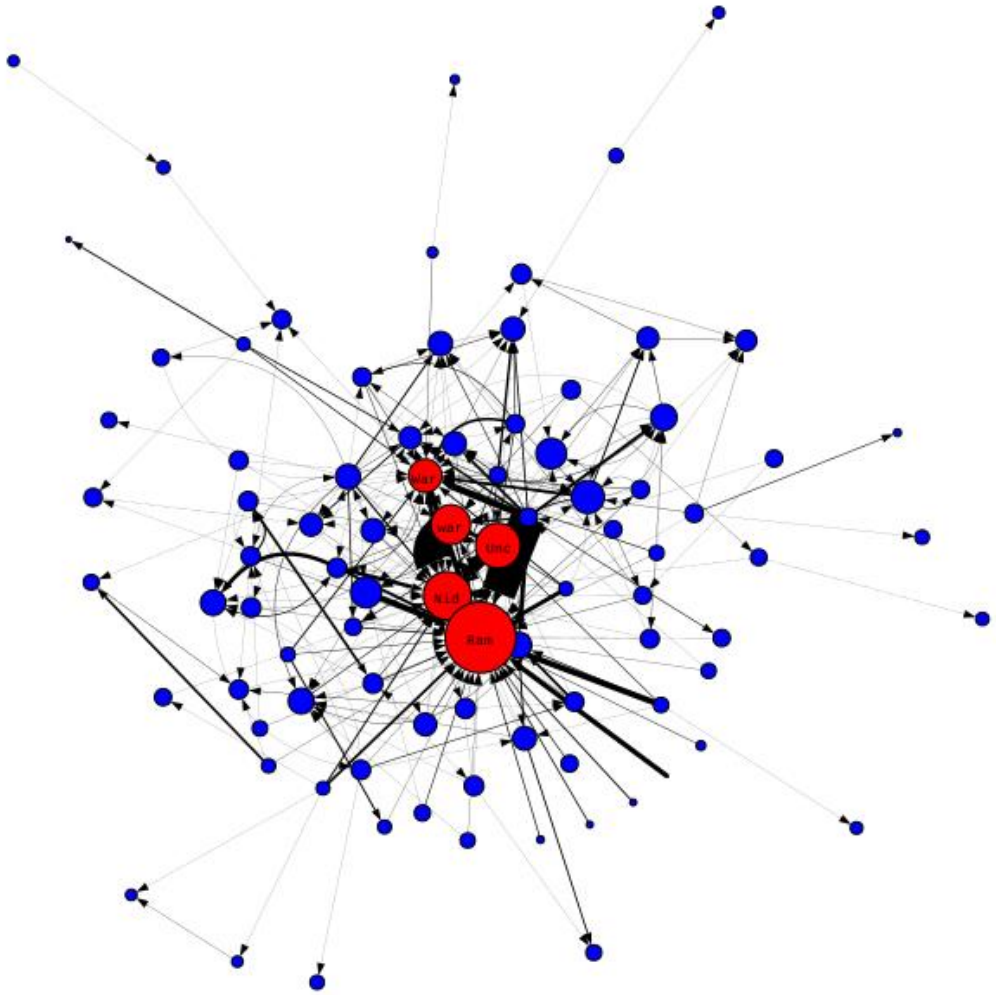
- l'**assenza** di **comunità significative**;
- l'alto **grado in ingresso** dei nodi più importanti, nonostante non si sia tenuto conto di questa metrica nel calcolo dell'importanza;

fattori che confermando ulteriormente le scelte prese in precedenza.

Inoltre, si notano particolarmente il **grande numero** di **menzioni** degli utenti:

- *warrnews* nei confronti di *Nidalgazaui;*
- *mobi_ayubi* nei confronti di *RamiAlLolah*.

I cui account però risultano sospesi.



Visualizzazione della rete di menzioni degli utenti

5. *NER* - Name Entity Recognition

Pre-processing base del testo dei tweet

Ai fini di NER e di sentiment analysis è utile rimuovere o comunque modificare i tweet da come sono presenti nel dataset originale. Le operazioni di pre-processing che verranno descritte saranno effettuate prima di sottoporre i tweet a traduzione automatica.

Le operazioni sono:

- rimozione hyperlinks;
- rimozioni di # (interferiscono con la sentiment analysis fatta da nltk e con la NER);
- rimozione di frasi come “ENGLISH TRANSLATION” o “ENGLISH TRANSCRIPT”;
- rimozione di “a capo” o di “&”.

Queste operazioni non sono effettuate prima di task quali la ricerca degli hashtag o la ricerca delle citazioni tra utenti.

Pipeline fino alla NER

Sono stati sperimentate due *pipeline* che portano ad estrarre sia le **entità** menzionate nei tweets sia il **sentiment** rilevato in tutto il tweet sia il sentiment associato alle varie entità (trovare entità diverse nelle pipeline implica associare poi sentiment diverso a queste).

In entrambe ci siamo avvalsi del modello ***en_core_web_lg*** di ***Spacy*** per fare **NER** (e le operazioni preliminari a questa). Si noti come il motore NER di Spacy utilizzi eventuali maiuscole per individuare meglio le entità, e quindi tutti i tweet che gli passeremo rispetteranno la capitalizzazione originale del tweet.

Pipeline 1:

Corrisponde alla pipeline di base di Spacy per fare NER. Internamente viene effettuata la ***Tokenization*** tramite un tokenizer rule-based specifico per l'inglese. Successivamente, applica un ***Tagger*** che effettua part-of-speech tagging. Applica la tecnica di ***Dependency Parsing*** (che riunisce anche parole over-segmentate in diversi token). Fa ***Lemmatization***. Viene effettuata ***Named Entity Recognition***.

Restituisce, per ogni tweet a cui la applichiamo, le sentences in cui divide il tweet e, per ognuna di queste, anche le entità che vi individua internamente. Sulle sentences e su tutto il tweet verrà fatto poi la sentiment analysis.

Pipeline 2:

Corrisponde alle operazioni di:

- *Stopword removal*: rimuoviamo le stopword presenti in una apposita lista della lingua inglese. Non rimuoviamo le negazioni perché possono essere rilevanti durante la sentiment analysis;
- *Punctuation removal*: togliamo la punteggiatura;
- *Stemming*: stemming tramite il *LancasterStemmer* della libreria NLTK.

A questo punto abbiamo un vettore di token per ogni tweet. Da questo si fa il join di tutti i token (con uno spazio in mezzo); e si ottiene per ogni tweet una stringa. Si usa **Spacy** per fare: **Tokenization**, **Tagging**, **Dependency Parsing**, **Named Entity Recognition**.

Quindi la differenza tra le due pipeline è che la prima non gestisce esplicitamente le stopword e la punctuation e fa la lemmatizzazione, mentre la seconda pipeline non fa lemmatization e non toglie stopword e la punteggiatura.

Sfortunatamente questo dataset **non contiene annotazioni** su quali dovrebbero essere le entità rilevanti citate nei tweets, quindi, risulta impossibile stimare oggettivamente se un risultato, in termini di entità estratte, di una pipeline sia *buono* o no (e anche compararle tra di loro quindi).

Abbiamo svolto un confronto soggettivo sulle entità che le due pipeline estraggono sui primi 20 tweet.

Confrontando i risultati delle due pipeline di NER otteniamo risultati che, in termini di entità estratte, sembrano **migliori** applicando la **prima pipeline** ai tweet.

Considereremo d'ora in poi le entità estratte dalla prima pipeline e il sentiment che verrà calcolato considerando queste.

6. Grafo delle citazioni a entità

Struttura del grafo

Una volta estratte le entità per ogni **tweet**, si è passati alla costruzione del **grafo di citazioni a entità**, un *grafo bipartito* nel quale un utente è in relazione con un'entità se quell'entità appare una o più volte nei suoi tweet.

Formalmente, si tratta di un grafo **diretto**, **bipartito** e **pesato**

$$G = (V, U, E, \delta, \gamma)$$

Nel quale

- $V = \{v_1, v_2, \dots, v_m\}$ è l'insieme degli **utenti presenti** nel **dataset**;
- $U = \{u_1, u_2, \dots, u_m\}$ è l'insieme delle **entità** estratte dei **tweets**;
- $E \subseteq V \times U$ è l'insieme degli archi, che rappresentano le citazioni di un utente verso una entità: $\forall v \in V, \forall u \in U: (v,u) \in E \Leftrightarrow$ l'**utente** v **cita** in (almeno) un tweet l'**entità** u;
- $\delta: E \to \mathbb{N}$ rappresenta il numero di citazioni di un utente verso un'entità: $\forall e = (v,u) \in E: \delta(e) = k \Leftrightarrow k$ è il **numero totale di citazioni** di v verso u (considerando tutti i tweet di v);
- $\gamma: V \to [0,1]$ rappresenta l'importanza di un utente, come definita nella sezione 4 e calcolata attraverso il **grafo delle citazioni**.

Il grafo sarà quindi composto da due categorie di nodi:

- **Nodi sorgente**, corrispondenti agli **utenti** del dataset;
- **Nodi pozzo**, corrispondenti alla **entità** estratte dai tweets.

Statistiche del **grafo delle entità**:

$\lvert V \rvert$	$\lvert U \rvert$	$\lvert E \rvert$	$< s_{out}(v) >$
88	7227	19598	520.5

Importanza di un'entità

Analogamente a quanto fatto per gli utenti, si è definita una misura di **importanza** per le **entità**, che tenesse conto di due aspetti principali:

- Il **numero** totale di volte che un'entità è **citata** nei tweets;
- La **qualità** delle **citazioni**, ovvero l'*importanza* dell'utente che cita l'entità.

Da un punto di vista formale, l'**importanza di un'entità** è stata definita come

$$importance(u) = \sum_{e=(v,u)\in E} \delta(e) \cdot \gamma(v)$$

Da cui, una volta normalizzata, è stato possibile estrarre un **ranking** delle **entità più importanti**, di cui riportiamo le prime dieci nella tabella che segue.

Rank	Entity	Category	Citations	importance
1	ISIS	ORG	1987	1
2	Syria	GPE	1457	0.868
3	RT	GPE	1465	0.476
4	Assad	PERSON	741	0.368
5	Iraq	GPE	601	0.308
6	Aleppo	GPE	530	0.255
7	US	GPE	667	0.248
8	Turkey	GPE	376	0.238
9	USA	GPE	265	0.165
10	Syrian	NORP	367	0.160

RT nel gergo di Twitter è un invito a re-twittare; tuttavia, con ogni probabilità è stato riconosciuto dalla **NER** come **Russia Today** (notiziario controllato dallo stato russo, ***G**eo**P**olitcal **E**ntity*).

Tutte le prime entità dieci entità sono in linea con importanti argomenti che sappiamo riguardare direttamente l'ISIS, il che valida la definizione di *importanza* data.

Ranking degli hashtags

Allo scopo di verificare se sia possibile utilizzare gli **hashtags** come una *buona rappresentazione* delle **entità** contenute nei **tweets**, ma ottenibili con un *costo computazionale* decisamente ridotto, abbiamo ripetuto la stessa procedura utilizzando gli **hashtags** estratti al posto delle **entità**.

Il grafo costruito risulta essere molto più contenuto:

	$\|V\|$	$\|U\|$	$\|E\|$	$< s_{out}(v) >$
	88	1872	4500	170.1
% su *Entity Graph*	100%	25.9%	22.96%	32.68%

E, applicando la stessa definizione di **importanza**, si ottiene il seguente ranking:

Rank	Hashtag	Citations	importance
1	Syria	1403	1
2	ISIS	1554	0.955
3	Iraq	633	0.410
4	IS	643	0.262
5	Aleppo	421	0.238
6	USA	272	0.210
7	Turkey	248	0.196
8	BreakingNews	252	0.194
9	Russia	274	0.179
10	Assad	247	0.156

In cui si ritrovano la maggioranza delle entità estratte anche con l'utilizzo della ***NER***.

Sicuramente questo approccio presenta due svantaggi:

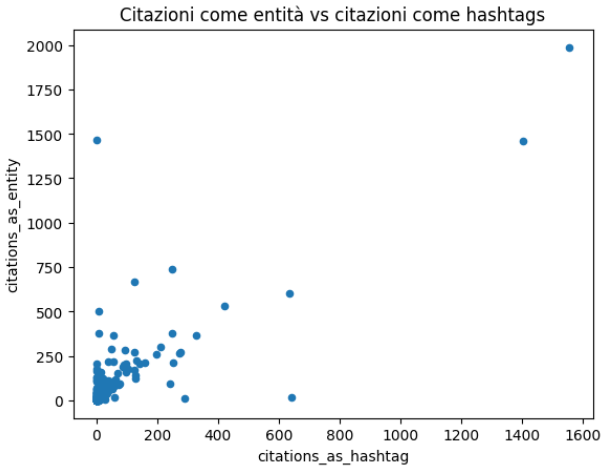
- il non poter riunire parole diverse che rappresentano la stessa **entità** (come succede per *ISIS* e *IS,* ad esempio);
- la mancata individuazione di alcune **entità**.

Tuttavia, almeno i primi dieci elementi derivanti da questa procedura risultano essere molto simili a quelli ottenuti tramite la ***Name Entity Recognition***, ma sono ottenuti dovendo pagare un *costo computazionale* molto ridotto. Inoltre, un altro vantaggio che emerge dal non dover utilizzare parametri di alcun tipo (come succede nella ***NER***), ottenendo sempre gli stessi risultati.

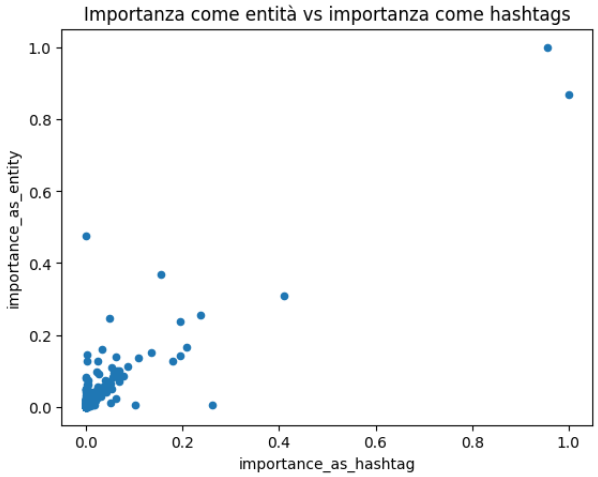
Una conferma numerica di quanto detto si ottiene considerando la **correlazione** tra gli elementi comuni delle due classifiche; in particolare:

- la ***NER* individua 7227 entità**, mentre gli **hashtags** sono solo **1872**; gli **elementi in comune** (a meno di maiuscole/minuscole/spazi) sono **489**, che però corrispondono alle entità più importanti citate.
- La **correlazione** tra **numero di citazioni** come entità e come hashtag è 0.790 (*correlazione medio-forte*);

- **correlazione** tra **importanza** come entità e come hashtag è 0.890 (*correlazione forte*).



Correlazione tra numero di citazioni come hashtag e come entità: 0.79



Correlazione tra importanza come hashtag e come entità: 0.89

7. Sentiment Analysis

Il linguaggio naturale può produrre frasi anche complicate, che possono menzionare più entità, e possono associare ad esse una certa emozione (o nessuna).

Osservando i dati abbiamo notato come in molti tweet venissero menzionate più di un entità.

Volendo estrarre il sentiment associato ad una specifica entità all'interno di una frase con più di una, abbiamo fatto un'assunzione: il sentiment associato ad un entità sarà dato dal sentiment della frase in cui si trova. Quindi il problema diventa calcolare il sentiment delle frasi.

Spacy, nel fare le operazioni necessarie per la NER, effettua anche un partizionamento della frase (o del tweet, in questo caso) in varie sentences. Per ogni sentence individua una o più entità.

L'algoritmo che eseguiamo è:

- Per ogni tweet, calcola il sentiment di tutto il tweet e, per ogni sentence di questo tweet, considera ogni entità: il sentiment associato all'entità è pari a quello della sentence in cui si trova.

Vogliamo quindi ottenere il sentiment legato a ogni tweet e quello legato ad ogni entità.

Sentiment Score

Assegniamo ad una stringa un sentiment score pari alla media tra

- l'***AFFIN score*** normalizzato su di essa
- lo score assegnato da ***NLTK***.

Combiniamo i metodi perché usando un metodo solo spesso il sentiment estratto risulta poco plausibile; usando questa combinazione ci pare di assegnare score più veritieri.

Per quest'ultimo abbiamo usato un ***SentimentIntensityAnalyzer*** della libreria **NLTK**: esso produce risultati nell'intervallo $[-1, +1]$, che poi vengono mappati in $[-5, +5]$.

AFFIN è un metodo *wordlist-based* per assegnare un sentiment tra -5 (molto negativo) e +5 (molto positivo) ad ogni parola. Noi lo usiamo come segue:

- l'***AFFIN score*** normalizzato di una stringa è la media degli score ***AFFIN score*** non-nulli di ogni parola che la compone.

Osservando i dati si nota anche che un'entità può essere citata più volte nello stesso tweet e in tweets diversi; di conseguenza occorre creare una struttura dati dove salvare ogni sentiment relativo ad ogni sua menzione.

Quindi, ad ogni entità associo diversi score di sentiment, uno per ogni volta che viene citata. Per calcolare il sentiment relativo alle entità poi si farà la media di questi score.

Abbiamo provato a salvare il tutto in una struttura dati che distinguesse entità diverse dal loro nome (nome=stringa con cui appaiono citate nei tweet); la procedura ha fatto emergere alcuni problemi:

1. Entità che nel mondo reale sono le stesse ma sono menzionate nei tweet in molti modi, come (USA, US) o (ISIS, ISIL, IS);
2. Presenza di entità spurie, cioè di casi in cui la NER non performa ottimamente e individua entità vere con caratteri aggiuntivi errati, e.g. "heretic kurds".
3. Individuazione di entità errate dove non presenti.

Per mitigare i primi due problemi abbiamo fatto in modo di creare strutture simili a "classi di equivalenza" tra stringhe: quando la NER individua un'entità, si cerca nella struttura dati un'entità con lo stesso nome o con una ***distanza di edit*** più bassa possibile. La soglia è pari a:

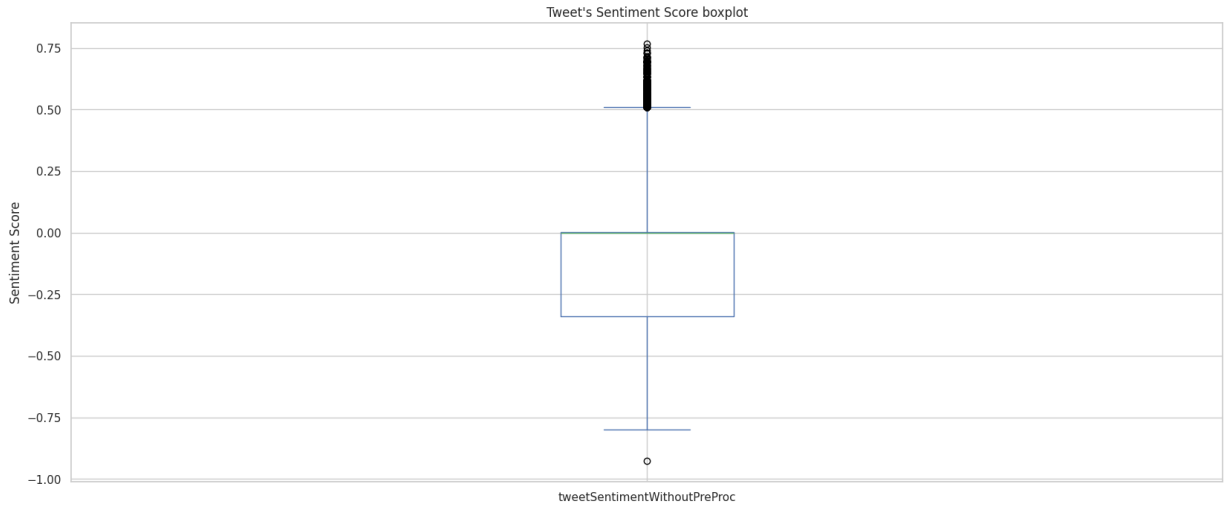
$$0.4 \cdot max\,\{lughezza\ del\ nome\ dell'entità\ da\ inserire;\ nome\ entità\ più\ vicino\ al\ nome\ della\ nostra\ entità\}$$

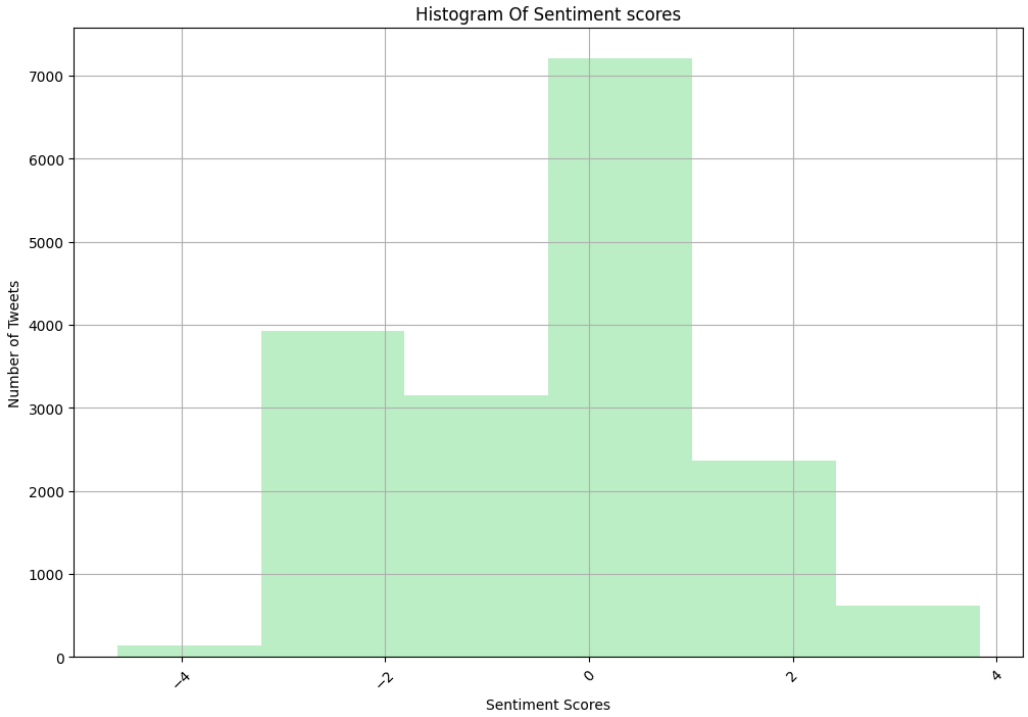
Questa soglia "dinamica", ovvero dettata dalla lunghezza dei nomi delle entità, è pensata come alternativa ad una soglia hard-coded come "3", ad esempio, che farebbe finire tutte le stringhe abbastanza corte in unica classe di equivalenza.

Se presente, si registra il sentiment associato all'entità da inserire nel "posto" di questa entità (dal nome uguale o simile).

Se nessun nome di entità è abbastanza simile, si crea una nuova entry per questa entità e si registra il suo sentiment.

Per migliorare ancora abbiamo fatto in modo che, prima del processing dei tweets, si partisse con la struttura dati già inizializzata da dei nomi di entità presenti. Essi non hanno ancora un sentiment associato a questo punto. Abbiamo scelto che questi "rappresentanti" delle nostre classi di equivalenza fossero i **cinquanta hashtags più rilevanti**, giustificato da quanto ottenuto nella sezione precedente. Avere dei buoni rappresentanti aiuta a fare in modo che le citazioni ad entità presenti nei tweet finiscano nelle classi di equivalenza "giuste" e non in classi create da menzioni spurie ma abbastanza vicine ai nomi di molte entità citate.

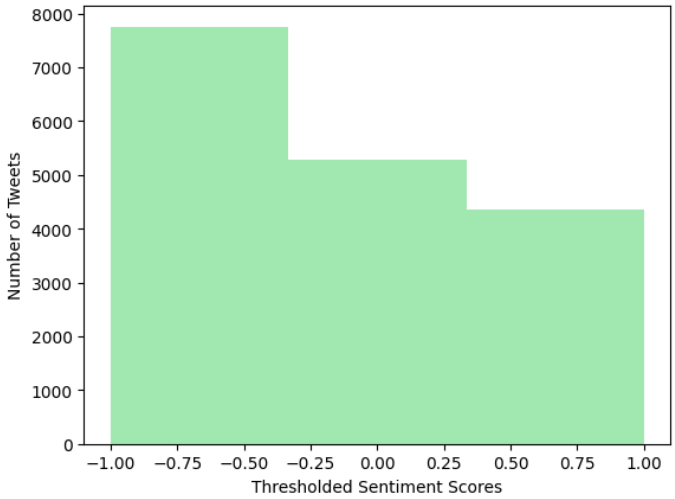




Nell'istogramma il sentiment è nel range [-5, 5]. Nel boxplot il sentiment score è riportato tra [-1, +1]

Notiamo come la moda sia il sentiment neutrale, ma che quei tweet che non lo sono, è più probabile che siano negativi che positivi, anche se abbiamo più outlier positivi che negativi.

Restringendo il valore del sentiment in $\{-1, 0, +1\}$ si ottiene il seguente diagramma.



Concludiamo che il dataset è sbilanciato verso l'avere tweets con **sentiment negativo**.

8. Relazione tra *polarità del sentiment* e *importanza* di un'entità

Per concludere, siamo andati a verificare la veridicità dell'ultima delle ipotesi iniziali, ovvero che ci sia una relazione di qualche tipo tra **importanza** e **polarità del sentiment** associati a un'**entità**, dove la **polarità del sentiment** è definito come il *valore assoluto* del sentiment associato a un'entità.

Al fine di ottenere valori in scala con la misura di **importanza**, il **sentiment** è stato portato nell'intervallo $[-1, 1]$.

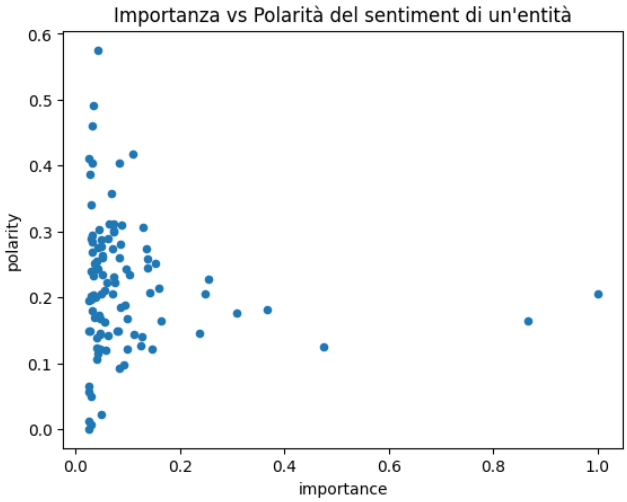
Si riportano i risultati ottenuti per le prime dieci **entità più importanti**:

Rank	Entity	Category	importance	sentiment
1	ISIS	ORG	1	-0.205
2	Syria	GPE	0.868	-0.164
3	RT	GPE	0.476	-0.126
4	Assad	PERSON	0.368	-0.182
5	Iraq	GPE	0.308	-0.176
6	Aleppo	GPE	0.255	-0.227
7	US	GPE	0.248	-0.206
8	Turkey	GPE	0.238	-0.146
9	USA	GPE	0.165	-0.164
10	Syrian	NORP	0.160	-0.214
11	YPG	ORG	0.152	-0.251

RT nel gergo di Twitter è un invito a re-twittare; tuttavia, con ogni probabilità è stato riconosciuto dalla **NER** come **Russia Today** (notiziario controllato dallo stato russo, ***G**eo**P**olitcal **E**ntity*).

Nonostante i valori di sentiment non siano troppo piatti verso la neutralità, non risulta possibile confermare l'ipotesi della corrispondenza tra **importanza** e **polarità del sentiment** associati a un'entità.

Anche escludendo le entità irrilevanti e considerando solo le prime cento in ordine di importanza, si ottiene infatti una correlazione quasi nulla, come si nota dal grafico a fianco.



Correlazione tra importanza e polarità del sentiment di un'entità: -0.088

Tuttavia, possiamo affermare di aver ottenuto un altro risultato: tutte quante le prime dieci entità riportate riportano un **sentiment negativo** e questa affermazione continua a valere per le prime **85 entità** in ordine di importanza. In particolare, considerando le prime **cento entità:**

- **98 entità** riportano sentiment negativo, solo **2** sentiment positivo (il cui valore è praticamente pari 0 e che quindi corrispondono più a un **sentiment neutro**);
- Il **valore medio** del sentiment è $-\mathbf{0.217}$ (su un minimo di -1);

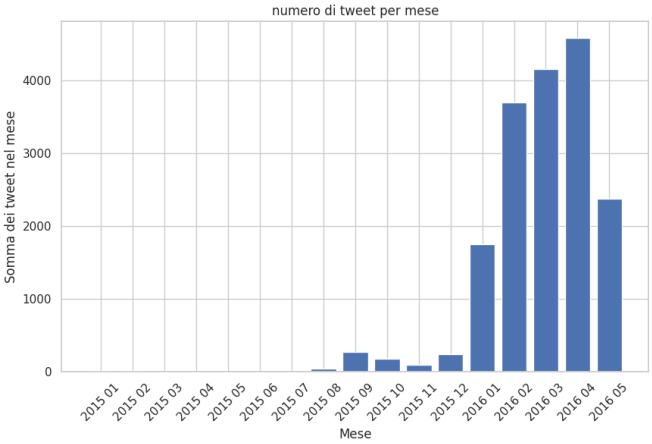
Possiamo quindi concludere che, in linea con l'andamento generale, anche alle entità più importanti è associato un sentiment negativo, fenomeno probabile molto diffuso nel mondo dei social network, e in particolare considerando un dominio applicativo come questo.

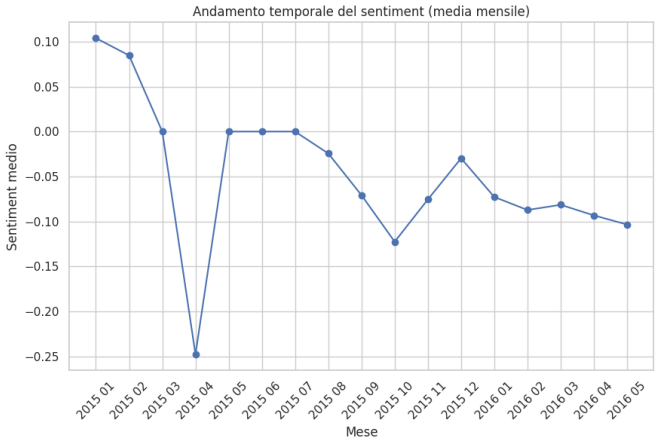
9. Task secondario: variazione del sentiment nel tempo

Un secondo task su cui ci siamo focalizzati è andare a studiare il nesso tra tempo e variazione del sentiment associato a certe entità.

Siamo partiti con l'analizzare la **variazione temporale del sentiment**, da ora in poi scalato nell'intervallo $[-1, 1]$, dei tweet considerando tutto il dataset.

Riportiamo i risultati ottenuti nei due diagrammi che seguono.





Notiamo che il numero di tweet inizia ad essere non trascurabile solo a partire da *settembre 2015*; da quel momento il **sentiment medio** dei tweets resta abbastanza **stabile** (fluttuazioni precedenti a questa data sono date dalla presenza di pochi tweet). A seguire considereremo quindi tweets a partire da questa data.

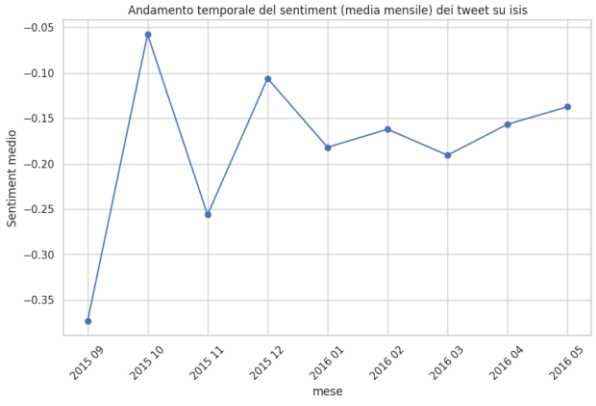
Distribuzione temporale del sentiment relativo a varie entità

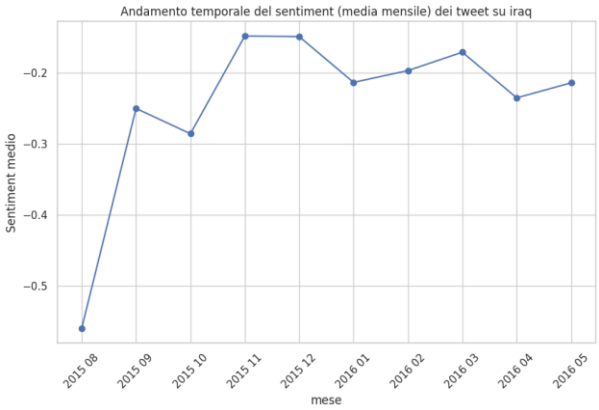
Selezionate quattro tra le **entità più importanti**

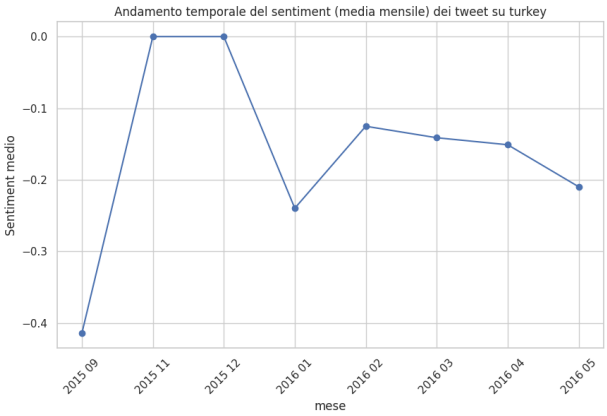
1. **ISIS**
2. **Turkey**
3. **Assad**
4. **Iraq**

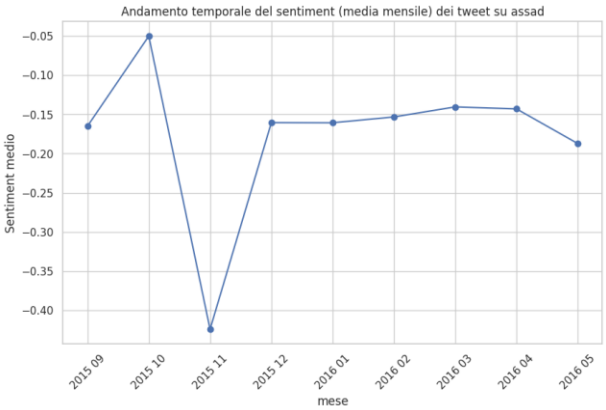
Ora analizziamo come cambia il sentiment a loro associato evolve nel corso del tempo.

Riportiamo i risultati ottenuti mese per mese nei diagrammi che seguono.









E’ possibile notare come dall’inizio del 2016 tutte le curve si stabilizzino (prima del 2016 ci sono relativamente pochi tweet quindi il sentiment fluttua di più).

Focalizzandosi sull’andamento del sentiment relativo ai all’entità **ISIS**, risulta evidente una repentina caduta nei mesi **ottobre e novembre** del **2015**.
Se selezioniamo i tweet con sentiment negativi di questo periodo e che menzionano **ISIS**, otteniamo dei tweets che parlano di recenti fatti di cronaca negativi per l’ISIS (come sconfitte o attacchi militari subiti) e notizie riguardo all'attacco a Parigi del 13/11/2015.