**SNA PROJECT**

La nostra analisi mira a costruire una rete ispirata alla diffusione del dibattito (tweet e retweet) sullo scoppio del conflitto in Ucraina.

Abbiamo iniziato ad analizzare i tweet in inglese dal 24 febbraio 2022 al 20(?) maggio.

L'idea era quella di cercare di classificare i tweet degli utenti in 4 gruppi:

* Pro Ucraina
* Pro Russia
* "Pax Romana
* "I don’t care"

L'obiettivo è categorizzare gli utenti esaminando i loro tweet.

L'analisi che effettueremo sui tweet comporterà, oltre all'osservazione degli hashtag in essi contenuti, anche l'analisi dell'andamento degli stessi grazie all'ausilio di tecniche NLP (Natural Language Processing) di Text analytics.

**1 CAT TREE**

Come individuiamo set di hashtags per categorizzare i vari utenti in base ai tweets che postano?

Il metodo del cat tree consiste nel:

- creare una lista per ogni categoria

- individuare un hashtag padre per ogni categoria, il quale andrà inserito nella lista appartenente

- selezionare i figli di ogni padre che verranno inseriti nella lista appartenente

**1.1. Hashtags-padre**

Nel nostro caso abbiamo individuato 3 principali categorie con i seguenti hastag padre:

1. Pro Ucraina - #slavaukraini
2. Pro Russia - #istandwithputin
3. Pax Romana - #stopwarinukraine

Gli hashtag padre sono stati selezionati utilizzando la funzione count; tramite ricerche su Twitter e utilizzando la libreria Twarc abbiamo individuato dei possibili candidati- padre per ogni categoria e per ciascuno abbiamo preso il più “twittato”. (fare tabella)

**1.2. Hashtags- figli**

Dopo aver individuato il padre relativo a ogni categoria abbiamo ricercato tutti i tweet contenenti tali hashtag padre (#slavaukraini, #istandwithputin, #stopwarinukraine) e se un tweet comprendeva un hashtag padre + altri hashtag, questi ultimi venivano inseriti automaticamente nella lista appartenete del padre.

(fare esempio)

Dopo aver ordinato le liste in base al numero di volte in cui sono comparsi(primo= più twittato), ripetere l’ago per i primi k figli di ogni lista (k=3), in modo da trovare hashtags provenienti non solo da quello madre.

Dopo aver creato le liste, il lavoro successivo consisterà in:

1. duplicate removing
2. list cleaning
3. individuare un metedo per selezionare una quarta categoria: I don’t care category

***1.3. Duplicate removing***

Un primo problema riscontato è stato che un figlio potesse apparire in più di una lista, quindi non avere una categorizzazione ben definita. Per risolverlo abbiamo deciso di assegnare uno score a ogni figlio che è presente in più di una lista.

*Score-cat-method*:

Viene effettuata una seconda ricerca per ogni figlio appartenete a più di una categoria fj:

- ricercare tutti i (recent?) tweet che contengo il figlio ripetuto + i 3 hashtags padre e salvarli in 3 liste differenti

- fare il count del numero totale di tweet (figlio +ogni hastags padre) trovati: c\_tot= Σ i=1 Ci

- fare il count del numero di volte in cui il figlio appare insieme a un determinato hashag padre, per ogni categoria i: Ci

Lo score per ogni figlio ripetuto fj è dato da: Ci / Σ i=1 Ci

Assegniamo un figlio ripetuto fj a una determinata categoria se:

- score maggiore

- se score>threshold

***1.4 List cleaning***

Dopo aver creato un algoritmo per eliminare in modo automatico i duplicati, occorre effettuare una valutazione umana (arbitraria) sull’appartenenza o meno di un hasHtag a una particolare categoria.

Ad esempio, tra i vari hashtags emersi, appaiono più volte #russia o #ucraina. Questi non appartengono a nessuna categoria, per questo motivo sono stati eliminati dalle liste.

Successivamente, abbiamo deciso di rimuovere hashtags irrilevanti ai fini della categorizzazione, come ad esempio tutti gli hashtags con un solo count o con un count inferiore a un certo thershold (ad esempio 10).

***1.5. Individuazione 4 categoria: I don’t care***

La categoria rimanente, denominata "I don't care", mira a raccogliere i tweet degli utenti disinteressati all'argomento. Sfortunatamente, l’approccio usato in precedenza non può essere utilizzato per questa quarta categoria ("I dont care"), poiché non è stato possibile identificare un vero e proprio hashtag che rappresentasse questa linea di pensiero.

Abbiamo quindi optato per un meccanismo di selezione "per esclusione": osserviamo gli hashtag utilizzati da persone che hanno twittato su focus diversi dalla guerra, ad esempio l’eurovision contest, utilizzando semplicemente l'hashtag #eurovision sulla query di ricerca dei tweet.

Ci concentreremo solo sugli utenti "attivi" sulla piattaforma. Se gli utenti più attivi individuati non hanno usato uno degli hashtag appartenenti a uno dei 3 gruppi già identificati, classificheremo questi utenti nella categoria "I don’t care".

* Per identificare gli utenti più attivi utilizzziamo l’attributo ['author.public\_metrics.tweet\_count']
* Invece per individuare gli hastags usati da un utente, prenderemo in considerazione l’attributo ['author.entities.description.hashtags']