

# Hillary Clinton's Leaked Emails

Progetto Data Analytics

Gianluca Giudice - 830694  
Daniele Turra - 860350

# Tabella dei contenuti

<b>Tabella dei contenuti</b>	<b>1</b>
<b>Introduzione</b>	<b>2</b>
Domande di ricerca	2
Approccio al problema	2
Struttura del Codice	3
<b>Dataset</b>	<b>5</b>
Descrizione	5
Features selection	7
Analisi esplorativa	8
Componente temporale	8
Attività dei contatti	11
<b>Sentiment</b>	<b>13</b>
Preprocessing	13
Named Entity Recognition e Linking sulle località	16
<b>Topic Modeling</b>	<b>21</b>
Considerazioni sul dominio	21
Creazione del modello	21
Implementazione del modello	21
Preprocessing e file di input	21
Scelta degli iperparametri	23
File di output	24
Risultati	25
Topic estratti	26

# Introduzione

## Domande di ricerca

Durante il suo ruolo come Segretario di Stato degli Stati Uniti d'America, Hillary Clinton si è resa protagonista di uno scandalo riguardante l'utilizzo di server di mail privati per il trattamento di informazioni istituzionali, fra cui alcune classificate e confidenziali. Secondo quanto riportato dagli investigatori, dietro all'intrusione non autorizzata e al furto di email potrebbero esserci hacker russi a supporto degli interessi geopolitici del proprio governo<sup>1</sup>. In seguito all'istituzione delle indagini, le mail sono state rese completamente pubbliche da esponenti di WikiLeaks, aprendo la strada alla condanna di entrambi gli schieramenti politici statunitensi durante la campagna elettorale del 2016.

Tuttavia, a seguito di indagini portate avanti dal FBI, non sono emerse evidenze riguardanti la presunta azione criminale intenzionale della Clinton, ma solo una sconsiderata superficialità nel trattamento di informazioni riservate che avrebbero dovuto invece essere trattate con riservatezza su server istituzionali.

Appare rilevante quindi l'analisi di queste email per capire la natura del *leak* e la qualità delle informazioni trapelate.

In particolare, abbiamo ritenuto di rilevante importanza l'analisi delle mail per:

- Comprendere l'indirizzo di politica estera tenuto dalla Clinton e dalla sua rete di contatti
- Riconoscere i diversi topic trattati nelle email
- Comprendere il ruolo che Hillary Clinton aveva nella sua rete di contatti

## Approccio al problema

Le informazioni contenute nel dataset sono il risultato di attività umana, densa quindi di significati che spaziano dalla realizzazione dell'interazione personale, fino alla discussione di temi politici su vasta scala. L'approccio scelto per il problema fa capo alle tecniche di Natural Language Processing affrontate, in particolare *sentiment analysis* e *topic modeling*.

Vista la posizione politica ricoperta dalla Clinton, la *sentiment analysis* applicata al dataset consentirà di estrarre informazioni circa l'indirizzo di politica estera tenuto dalla Clinton e dal suo team. L'analisi testuale si confronta sempre con la natura umana e, specie se effettuata sui contenuti prodotti dai social media, può dover affrontare testi grammaticalmente scorretti che spaziano fra diversi linguaggi e spesso in tono informale. Per quanto riguarda questi scambi di email, siamo di fronte ad un grado di maggiore omogeneità di tematiche e stile, così come un ristretto numero di interlocutori, senza però poter eliminare l'ambiguità.

Per questo tipo di ricerca sono stati riconosciuti ed estratti i diversi paesi che sono presenti nelle mail tramite tecniche di Named Entity Recognition. Successivamente abbiamo

---

<sup>1</sup> <https://nypost.com/2017/11/04/this-is-how-russian-hackers-pried-into-hillary-clintons-emails/>

utilizzato DBpedia per creare un collegamento tra entità riconosciute ed entità nel mondo reale. L'utilizzo combinato di tecniche di *sentiment analysis* e di Named Entity Linking consente un elevato grado di automazione nell'analisi del dataset.

Per poter davvero valutare la rilevanza ai fini delle indagini delle informazioni contenute, si rende necessaria la creazione di un topic model capace di identificare i diversi aspetti trattati nelle email con il relativo *sentiment* associato.

Per tutta l'analisi abbiamo utilizzato Python e in particolare librerie come Pandas, NLTK, spacy, AFINN e ASUM.

## Struttura del Codice

```
└── asum          # Binari di ASUM
    ├── bin
    ├── in           # Input file asum
    │   ├── BagOfSentences.txt
    │   ├── SentiWords-0.txt
    │   ├── SentiWords-1.txt
    │   └── WordList.txt
    └── out          # Output file asum
        ├── STO2-T7-S2(2)-A0.1-B0.001,0.1,0.1-G1.0,1.0-l1000-Phi.csv
        ├── STO2-T7-S2(2)-A0.1-B0.001,0.1,0.1-G1.0,1.0-l1000-Pi.csv
        ├── STO2-T7-S2(2)-A0.1-B0.001,0.1,0.1-G1.0,1.0-l1000-ProbWords.csv
        └── STO2-T7-S2(2)-A0.1-B0.001,0.1,0.1-G1.0,1.0-l1000-Theta.csv
    └── attività_contatti.ipynb # Analisi attività contatti
    └── data          # Input Dataset
        ├── Aliases.csv
        ├── EmailReceivers.csv
        ├── Emails.csv
        ├── Persons.csv
        └── database.sqlite
    └── live_demo.py      # Flask web app (live demo)
    └── pickle          # Dataframe serializzati dopo analisi
        ├── df.pkl
        ├── df_entities.pkl
        ├── df_geo.pkl
        └── df_nations.pkl
    └── sentiment_analysis.ipynb # Sentiment analysis
    └── src             # Funzioni utili live demo
        ├── countries_sentiment.py
        ├── exploratory_analysis.py
        └── sentiment_analysis.py
    └── topic_modeling.ipynb      # Topic modelling
    └── topic_modeling_analysis.ipynb # Visualizzazione risultati topic modelling
```



# Dataset

## Descrizione

Il dataset originale è stato reso pubblico su Kaggle in seguito ad una pulizia ed analisi dei file PDF originali rilasciati dal Dipartimento di Stato americano ed è scaricabile all'URL indicato in nota<sup>2</sup>. I PDF sono stati recuperati grazie al *Freedom of Information Act* (FOIA) che consente l'accesso alle informazioni governative quando non più coperte da segreto di stato, garantendo una maggiore trasparenza verso la cittadinanza. Si tratta di documenti considerati come *unclassified* e che di conseguenza sono a *basso impatto*; non contengono cioè informazioni considerate come potenzialmente dannose se diffuse e non necessitano quindi protezioni di tipo speciale.

Come specifica il creatore del dataset sul proprio profilo GitHub<sup>3</sup>, ci sono delle imprecisioni dovute all'estrazione delle informazioni da email salvate in PDF, in particolare nelle sezioni di sender/receiver e nel corpo della email.

Il dataset acquisito è composto da 5 file:

1. **Emails.csv**: File contenente tutte le email con i relativi metadati
2. **Persons.csv**: Mapping tra ID persona e nome della persona
3. **EmailReceivers.csv**: Destinatari delle email (dal momento che una mail può essere spedita a più persone)
4. **Aliases.csv**: Associazione tra alias persona e nome della persona
5. **Database.sqlite**: Versione sqlite del dataset. I file .CSV sono un dump del database

### Emails.csv

Attributo	Descrizione	Tipo
Id	Identificatore univoco per riferimento interno	Intero
DocNumber	Numero documento FOIA	Stringa
MetadataSubject	campo SUBJECT email (dai metadati FOIA)	Stringa
MetadataTo	Campo Email TO (dai metadati FOIA)	Stringa
MetadataFrom	campo Email FROM (dai metadati FOIA)	Stringa

<sup>2</sup> <https://www.kaggle.com/kaggle/hillary-clinton-emails>

<sup>3</sup> <https://github.com/benhamner/hillary-clinton-emails/>

SenderId	PersonId del mittente dell'e-mail (collegamento alla tabella Persone)	Intero
MetadataDateSent	Data di invio dell'e-mail (dai metadati FOIA)	Data (ISO 8601)
MetadataDateReleased	Data di rilascio dell'e-mail (dai metadati FOIA)	Data (ISO 8601)
MetadataPdfLink	Link al documento PDF originale (dai metadati FOIA)	Stringa
MetadataCaseNumber	Numero del caso (dai metadati FOIA)	Stringa
MetadataDocumentClass	Classe documento (dai metadati FOIA)	Stringa
ExtractedSubject	campo SUBJECT email (estratto dal PDF)	Stringa
ExtractedTo	Campo Email TO (estratto dal PDF)	Stringa
ExtractedFrom	campo Email FROM (estratto dal PDF)	Stringa
ExtractedCc	Campo Email CC (estratto dal PDF)	Stringa
ExtractedDateSent	Data di invio dell'e-mail (estratta dal PDF)	Data (ISO 8601)
ExtractedCaseNumber	Numero del caso (estratto dal PDF)	Stringa
ExtractedDocNumber	Numero documento (estratto dal PDF)	Stringa
ExtractedDateReleased	Data di rilascio dell'e-mail (estratto dal PDF)	Data (ISO 8601)
ExtractedReleaseInPartOrFull	Se l'e-mail è stata parzialmente censurata (estratta dal PDF)	Stringa
ExtractedBodyText	Tentativo di estrarre solo il testo nel corpo che il mittente dell'e-mail ha scritto (estratto dal PDF)	Stringa

RawText	Testo e-mail non elaborato (estratto dal PDF)	Stringa
---------	--	---------

### Persons.csv

Attributo	Descrizione	Tipo
Id	Identificatore univoco per riferimento interno	Intero
Name	Nome della persona	Stringa

## Features selection

Come si può notare Emails.csv ha 22 attributi; per l'analisi sono stati considerati solo:

- SenderPersonId
- MetadataDateSent
- ExtractedBodyText

Tutte le analisi condotte considerano solo il mittente della mail in quanto questo viene riconosciuto come "proprietario" della mail. Inoltre vengono considerate tutte le email del dataset e non solo quelle inviate da Hillary Clinton in quanto per tutte le conversazioni vale che o Hillary Clinton è il mittente della mail o il destinatario (o uno dei destinatari nel caso di più riceventi).

Persons.csv è stato selezionato per poter accoppiare in modo univoco, preciso e facilmente interpretabile gli IDs associati a ciascuna persona ad un nome preciso, a prescindere dagli alias possibilmente utilizzati. Infatti, molti interlocutori hanno vari alias che rendono difficile l'identificazione di ciascuno, raccolti e identificati nel file *Aliases.csv*. Ad esempio, Cheryl Mills ha i seguenti alias dovuti a variazioni nelle stringhe associate: *c:mills*; *cherylccheryl*; *millscheryl*; *mills*; *cosmill*; *cherylmills*; *cherlyl dmills*; *chery dmills*; *cherylmills cheryl*; *dmillscheryl dmills*; *cherl dmills*; *cheryl*; *millscd@state.gov*.

## Analisi esplorativa

Il dataset di partenza è composto da 7945 email.

Dopo aver associato l'identità univoca di ogni persona al relativo ID, abbiamo provveduto all'analisi esplorativa del dataset con l'intenzione di estrarre una prima descrizione e ricerca degli elementi salienti del dataset.

Per prima cosa, abbiamo deciso di concentrarci sugli attributi *SenderId* e *MetadataDateSent* con l'intenzione di esplorare le attività dei vari contatti del dataset. Tuttavia, concentrandoci sul file *Emails.csv*, sono emerse alcune celle mancanti proprio per questi due attributi, rendendo necessaria una strategia per gestire queste istanze incomplete. Riteniamo che la mancanza dell'attributo relativo alla data non dipende da altri attributi dell'istanza, ma piuttosto sia una mancanza completamente casuale (Missing value MCAR). Per questo motivo abbiamo deciso di ignorare le mancanza e di cancellare le 190 osservazioni imputate, riducendo il dataset del 3%. La dimensione del dataset viene di poco ridotta in quanto si ha una buona completezza del dataset sulla dimensione della data.

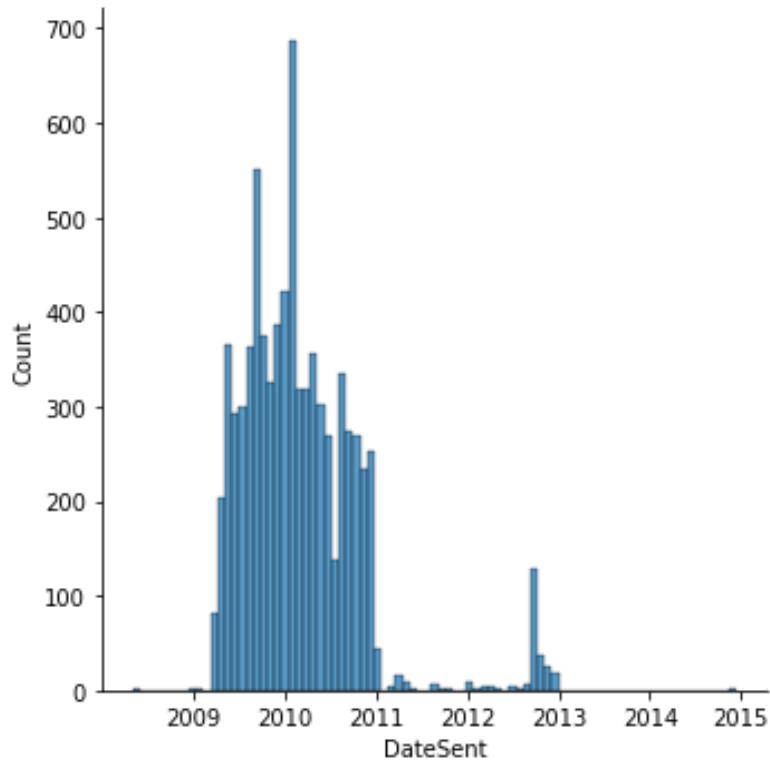
Dopo questi fasi di pulizia del dataset e filtraggio si passa da 7945 email iniziali a 6737 email che verranno analizzate.

### Componente temporale

Quindi abbiamo analizzato le date lavorando sull'attributo “*MetadataDateSent*” e abbiamo notato come non fossero presenti informazioni utili riguardo l'orario di spedizione, in quanto sempre alternativamente fra le 4:00 AM o le 5:00 AM. Ogni orario era in ISO 8601 e seguiva la seguente formattazione.

*yyyy-mm-ddThh:mm:ss+hh:ss*

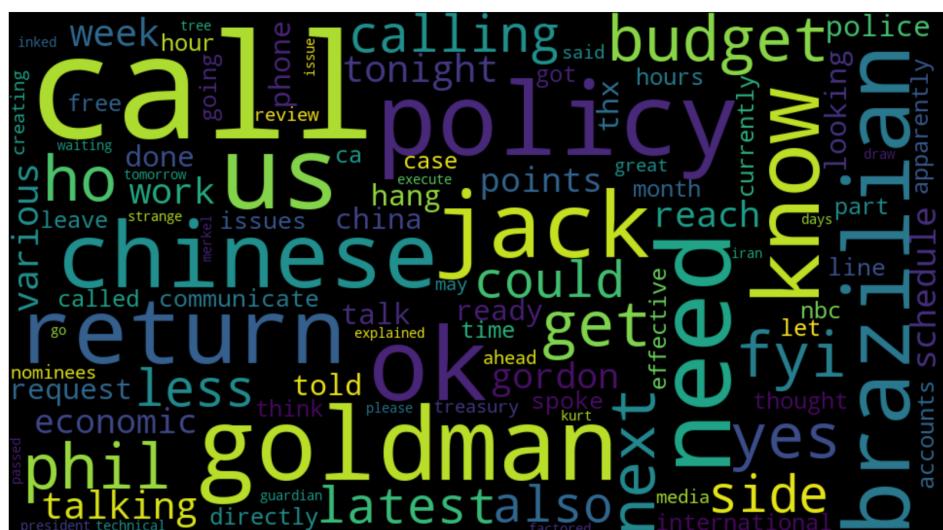
Abbiamo quindi provveduto alla pulizia del timestamp dagli orari, tenendo in considerazione solo giorno, mese e anno. In tal modo è stato semplice notare come la prima email sia stata spedita in data 2008-05-01, mentre l'ultima sia stata spedita in data 2014-12-14, consentendoci di affermare che il dataset copra 2418 giorni, ovvero più di 6 anni.



Dalla distribuzione del dataset appare evidente come la maggior parte delle mail si concentri tra il 2009 e il 2011, con comunque il 90% delle mail concentrato nel 65% del totale dei giorni.

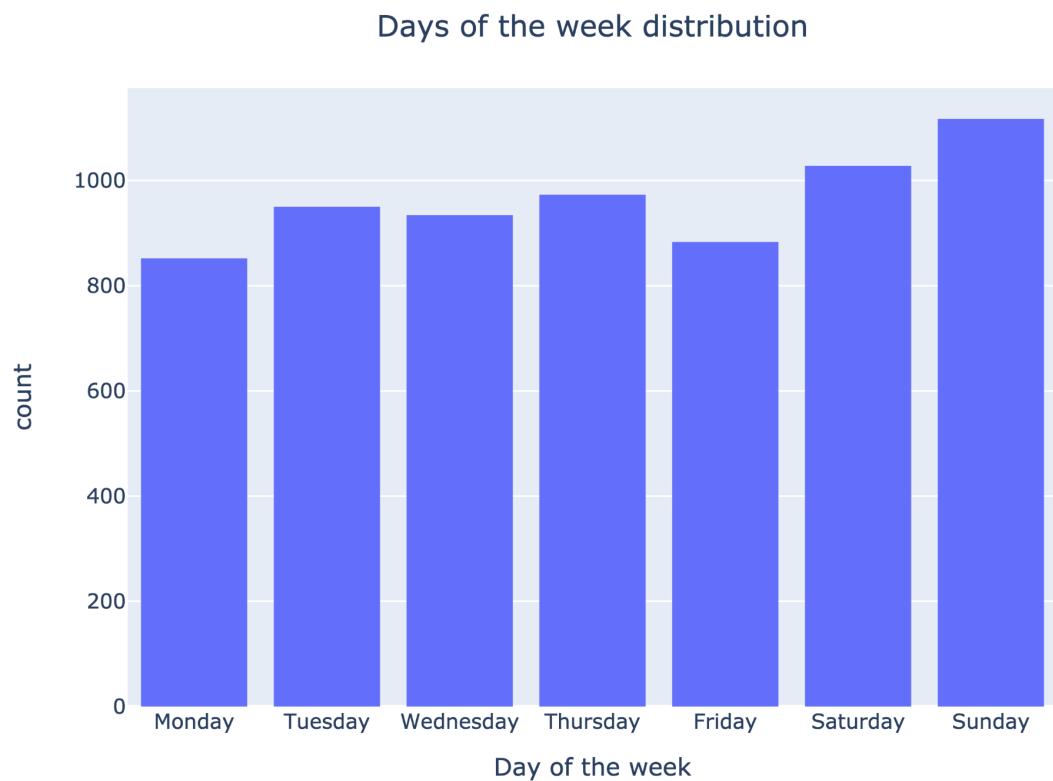
Il giorno più attivo si è rivelato essere il 23 dicembre 2009 con 47 email fra spedite e ricevute, portandoci a supporre inizialmente che si trattasse di auguri natalizi, e quindi tematiche marcatamente personali. Tuttavia analizzando la word cloud relativa a questo giorno possiamo confutare la tesi iniziale.

Word cloud of the 23/12/2009

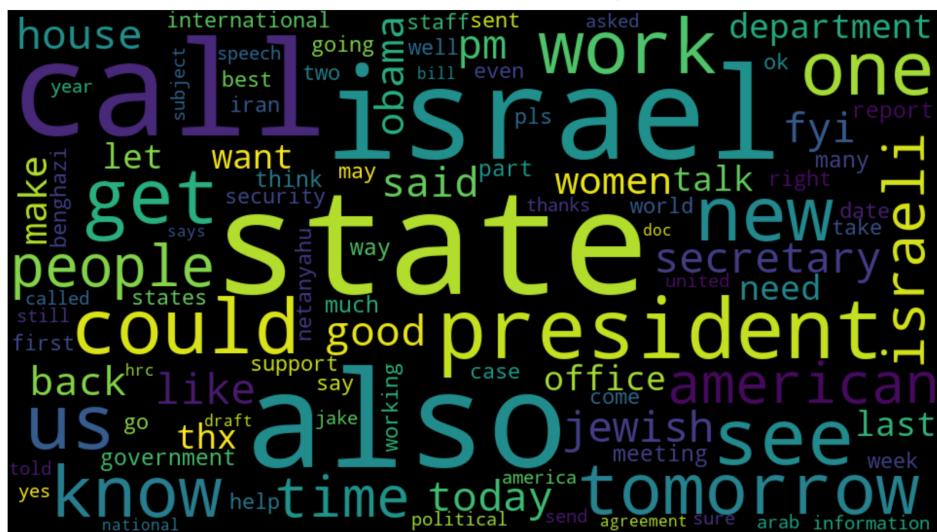


Questa attività sembra essere dovuta a rapporti diplomatici fra gli Stati Uniti e altri paesi, si può infatti notare “chinese” e “brazilian”, probabilmente causato da una particolare fibrillazione del Congresso. Inoltre ad una lettura manuale delle mail si può confermare quanto detto.

L'analisi delle frequenze sulla base del giorno della settimana porta al primo posto la domenica, avvalorando l'ipotesi che l'account fosse utilizzato principalmente per scambio di mail personali e non riguardanti il lavoro. Ovviamente, essendo questo un dataset ridotto rispetto ai dati originali del mail server della Clinton dal quale sono state rimosse le mail contenenti materiale classificato, a questo punto dell'analisi non possiamo affermare con certezza questa ipotesi.



## Word cloud of Sundays



## Attività dei contatti

Viene ora analizzata l'attività dei contatti per capire quali sono quelle persone che più hanno interazioni con Hillary Clinton.

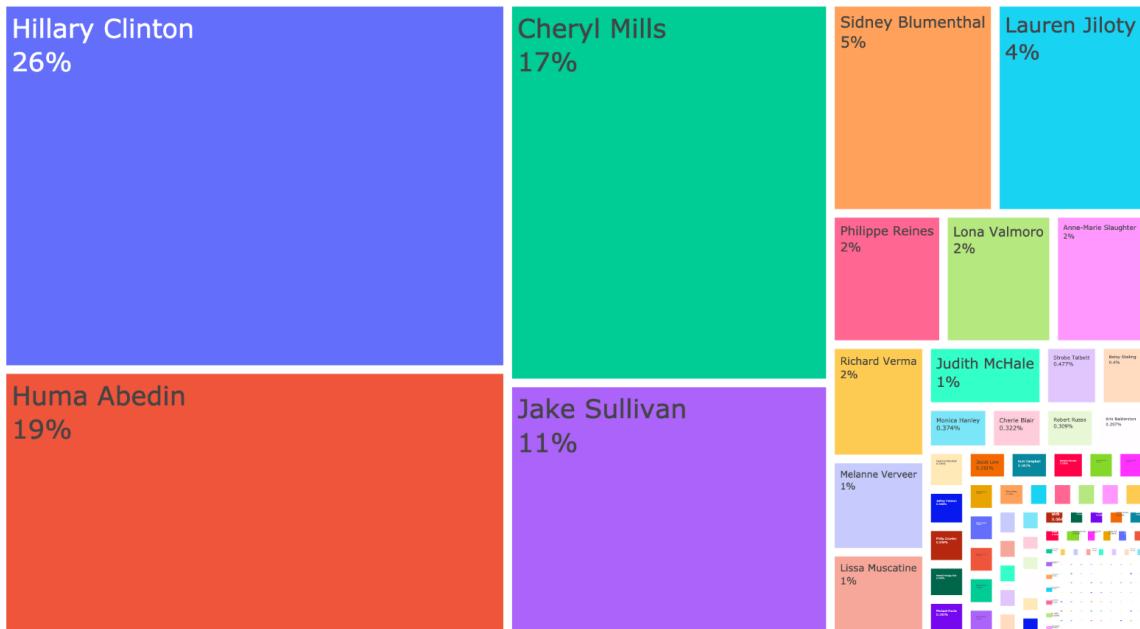
Viene sotto riportata la tabella con le dieci persone con frequenza maggiore per campo “mittente” delle mail. In particolare, Wikipedia definisce i primi 4 contatti come appartenenti al partito Democratico e partecipanti al **team di supporto della campagna elettorale presidenziale della Clinton nel 2016**. Il quinto contatto per frequenze, Sidney Blumenthal, è stato accusato più volte di particolarismo nella propria attività da giornalista, “erodendo sempre più il confine tra giornalismo di parte e giornalismo indipendente”<sup>4</sup>.

SenderId	Counts	RelativePercentage
Hillary Clinton	1991	26%
Huma Abedin	1437	19%
Cheryl Mills	1317	17%
Jake Sullivan	871	11%
Sidney Blumenthal	373	5%
Lauren Jiloyt	341	4%
Philippe Reines	159	2%
Lona Valmoro	155	2%

<sup>4</sup> [https://en.wikipedia.org/wiki/Sidney\\_Blumenthal](https://en.wikipedia.org/wiki/Sidney_Blumenthal)

Anne-Marie Slaughter	130	~2%
Richard Verma	118	~2%
Others	863	~10%

### Origine delle mail in percentuale



Mostrando l'attività dei contatti cumulata si evince che le prime dieci persone per frequenza di attività concentrano circa il 90% del flusso di mail.



# Sentiment

## Preprocessing

Partendo da questi dati si utilizzano tecniche di pre-processing che consistono di pulire il testo di partenza e prepararlo per essere analizzato.

Per l'analisi del *sentiment* abbiamo innanzitutto estratto i dati dal dataset originario selezionando i valori non nulli degli attributi *ExtractedBodyText* e *MetadataDateSent*. Nonostante le righe imputate rappresentassero circa il 15% del dataset originale, non riteniamo si sia inserito un bias nell'analisi successiva andando a rimuovere queste istanze. Infatti, in questo caso, abbiamo supposto che una mail con il corpo del testo vuoto non potesse giovare per la successiva analisi del *sentiment*. Inoltre, come già affermato precedentemente, il ridotto numero di mail con campo *MetadataDateSent* nullo rende l'effetto della loro rimozione trascurabile.

Successivamente, abbiamo reso le stringhe estratte più adatte alla tokenizzazione operando un filtraggio tramite regular expression andando a rimuovere i seguenti elementi:

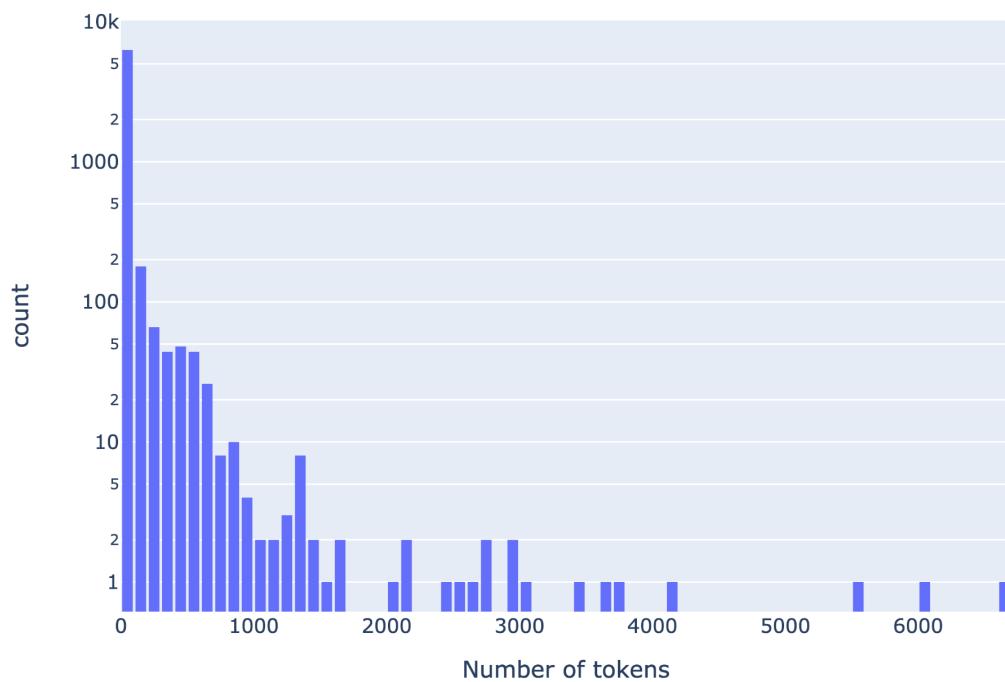
- **Header della mail**
  - Mittente della mail
  - Destinatario della mail
  - Data della mail
  - RE:
  - FW:
- **Firma della mail**
  - “U.S. Department of State”
  - “SUBJECT TO AGREEMENT ON SENSITIVE INFORMATION & REDACTIONS. NO FOIA WAIVER.”

Si noti come la lunghezza media del corpo del testo estratto si abbassa a quota 496.15 caratteri per mail, contro i 525.25 originari.

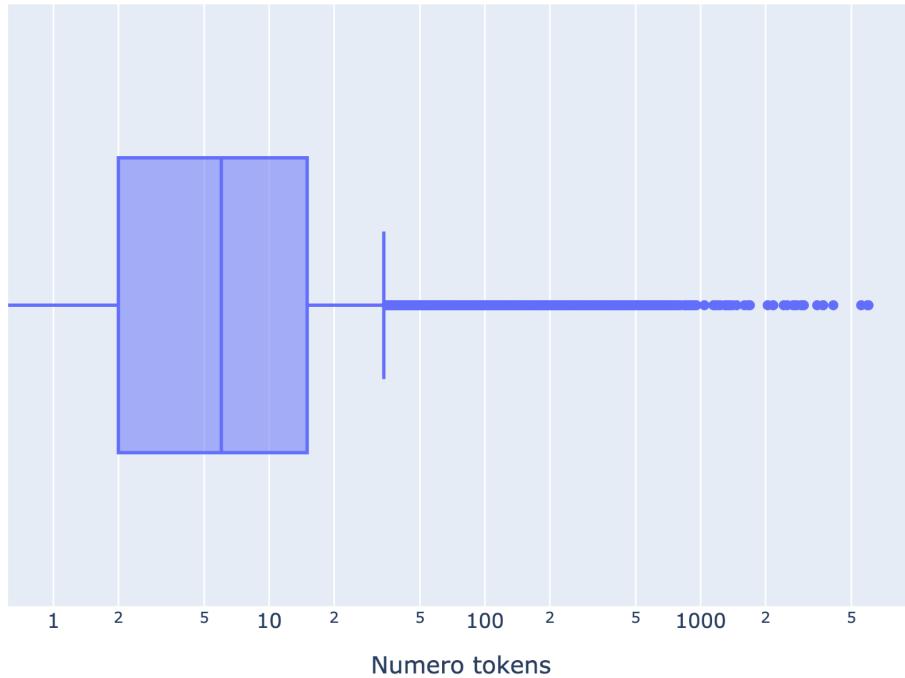
Grazie all'ausilio della libreria NLTK, abbiamo creato i token utili alla successiva analisi andando a rimuovere punteggiatura, numeri, caratteri singoli e le cosiddette stop words, ovvero le parole più usate nella lingua target che non contengono di per sé un significato se non legate ad un contesto. Importante notare come non è stato effettuato stemming. Questa scelta è stata fatta avendo l'obiettivo di fare *sentiment analysis* utilizzando la libreria AFINN. Infatti questa libreria approccia la *sentiment analysis* sfruttando un lessico. In questo lessico le parole sono contenute nella versione non stemmatizzata, è quindi di fondamentale importanza non stemmatizzare i tokens così da avere dei match all'interno del lessico.

Alla fine del processo di tokenizzazione, la media dei token è di 43 per email, con una deviazione standard pari a 221.

Number of tokens per email (bin size = 100)



Numero di tokens per email



Mean	Std	Median	min	25%	50%	75%	max
43.476	221.163	6	0	2	6	15	6611

Analizzando la distribuzione del numero di tokens per email, emerge la natura skewed della distribuzione verso le email più corte. Infatti il terzo quartile indica come il 75% delle email siano molto corte, con numero di tokens minore o uguale a 15.

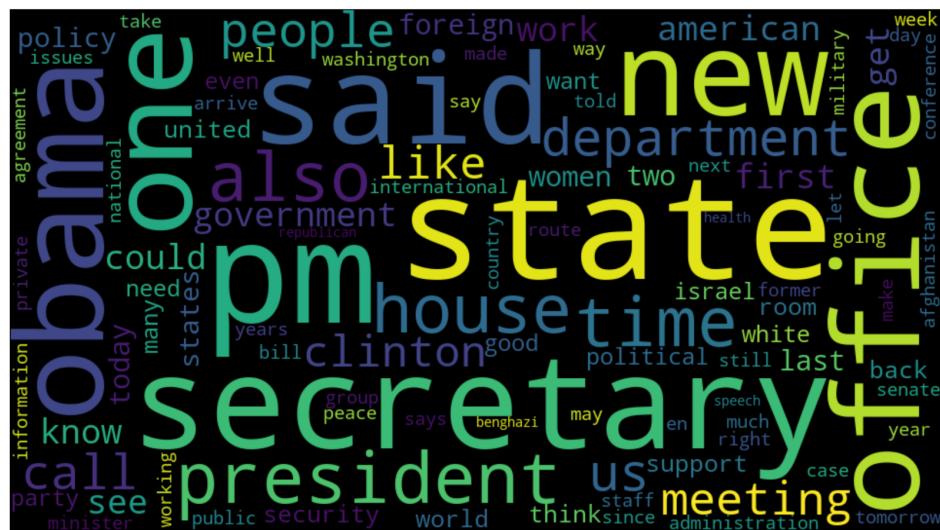
Nonostante non sia semplice definire gli outliers, sopra i 3500 token per email abbiamo poche osservazioni, rendendo questi valori potenzialmente interessanti per la possibile peculiarità dei temi trattati nelle mail a cui appartengono.

Di seguito sono mostrate le word cloud dei 100 token più rappresentativi contenuti rispettivamente in email con meno del valore mediano e almeno il valore mediano.

Word cloud of the short emails (Num. of tokens < 6)



### Word cloud of the long emails (Num. of tokens >= 6)

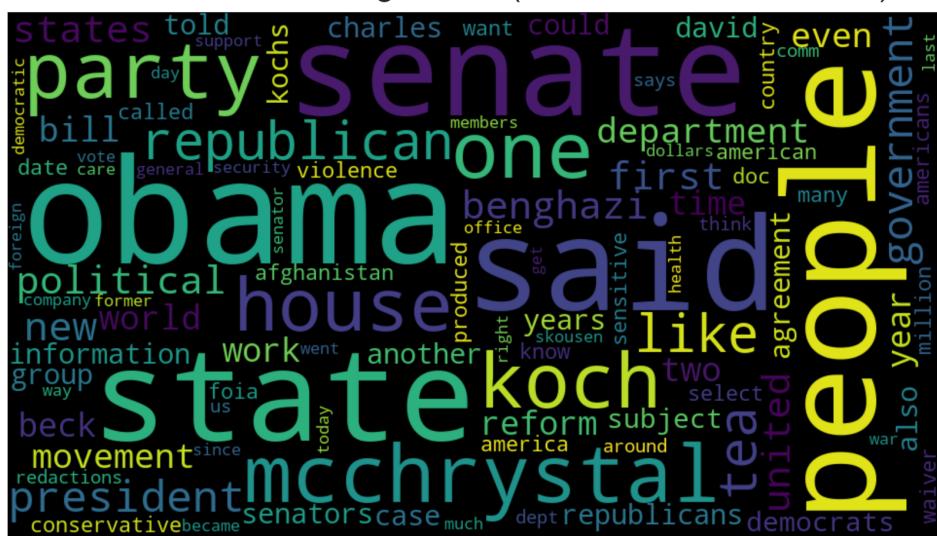


Com'è facilmente intuibile, le email più corte riguardano **situazioni d'ufficio e organizzative**, così come il più evidente "fyi" (*for your information*) che indica lo scambio e

l'inoltro di informazioni originatesi al di fuori con il semplice scopo dell'avviso. Si intravedono anche parole associate al mondo della politica ("state", "secretary", "obama", "department") che renderebbero associabile ad essa la wordcloud anche in assenza di altre informazioni sull'origine del dataset.

Le email più lunghe, invece, sono dichiaratamente politiche, con "state" come parola più utilizzata. "Obama" e "secretary" appaiono pure rilevanti, essendo stata la Clinton appunto segretaria di Stato durante il mandato Obama. In particolare, nella wordcloud sottostante sono rappresentati i 100 token più frequenti nelle email con più di 3500 token, appartenenti alla coda della distribuzione. In queste email, appaiono soggetti interessanti e non visti precedentemente, come "benghazi", "mcchrystal", "koch" e "beck". Ognuno di questi 4 token rappresenta un punto di interesse per la politica americana di quegli anni.

### Word cloud of the long emails (Num. of tokens >= 3500)



## Named Entity Recognition e Linking sulle località

L'accuratezza della fase di preprocessing si è resa di vitale importanza data la successiva intenzione di utilizzare tecniche di *Named Entity Recognition* e *Linking*. Ciò nonostante, le intrinseche caratteristiche dei canali di comunicazione online rendono possibili errori di identificazione da parte dei modelli utilizzati per NER, probabilmente dovuto all'apprendimento dei modelli stessi.

L'obiettivo della fase di Named Entity Recognition è di classificare un singolo token o una composizione di token come un particolare tipo di entità tra quelle definite a priori (su cui è stato trainato il modello). Nel nostro caso l'obiettivo della fase di Named Entity Linking è l'associazione della singola entità alla sua controparte nel mondo reale sfruttando come base di conoscenza DBpedia. Questo ci permette di arricchire l'informazione iniziale andando a reperire delle informazioni relative alle entità, con lo scopo di combinare questi dati per estrarre aspetti rilevanti.

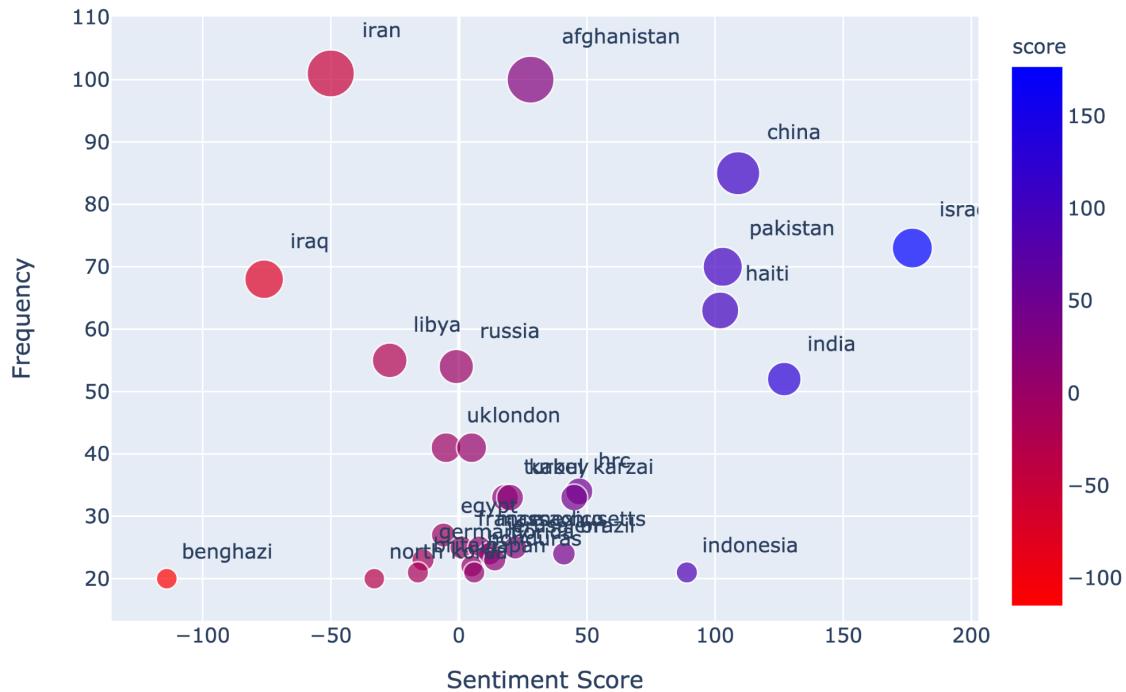
Per la prima fase di riconoscimento delle entità, ci siamo avvalsi dell'utilizzo della libreria Python SpaCy, un potente strumento *open source* che integra vari aspetti di NLP e con un ampio bacino di contributori, e di AFINN, una consolidata libreria per la *sentiment analysis* basata su lessico.

L'algoritmo di estrazione delle entità va a ricercare le entità alle quali può essere assegnata la label "GPE", ovvero nazioni, città o stati. Per ogni corpo del testo vengono rilevati i token che possono appartenere all'etichetta GPE, viene estratta la frase che contiene quella particolare entità e da lì viene assegnato il relativo *sentiment* calcolato in base alle altre parole presenti nella frase. In quanto troppo ricorrenti e indicanti lo stesso concetto, le entità più frequenti legate agli Stati Uniti sono state rimosse, preservando le entità legate agli altri paesi stranieri e agli altri stati americani al di fuori della capitale. È importante notare come la libreria SpaCy non sia perfetta e come possano apparire episodi di falsi positivi, cioè entità non appartenenti alla label selezionata ma classificate come tali. È per questo che è stato necessario togliere manualmente alcune entità errate che non sono state correttamente rilevate durante la fase di named *entity recognition*.

[...] The United States GPE should immediately ask the Security Council  
to authorize a no-flight zone and make clear to Russia GPE and China GPE that if they  
block the resolution, the blood of the Libyan NORP opposition will be on their hands. [...]

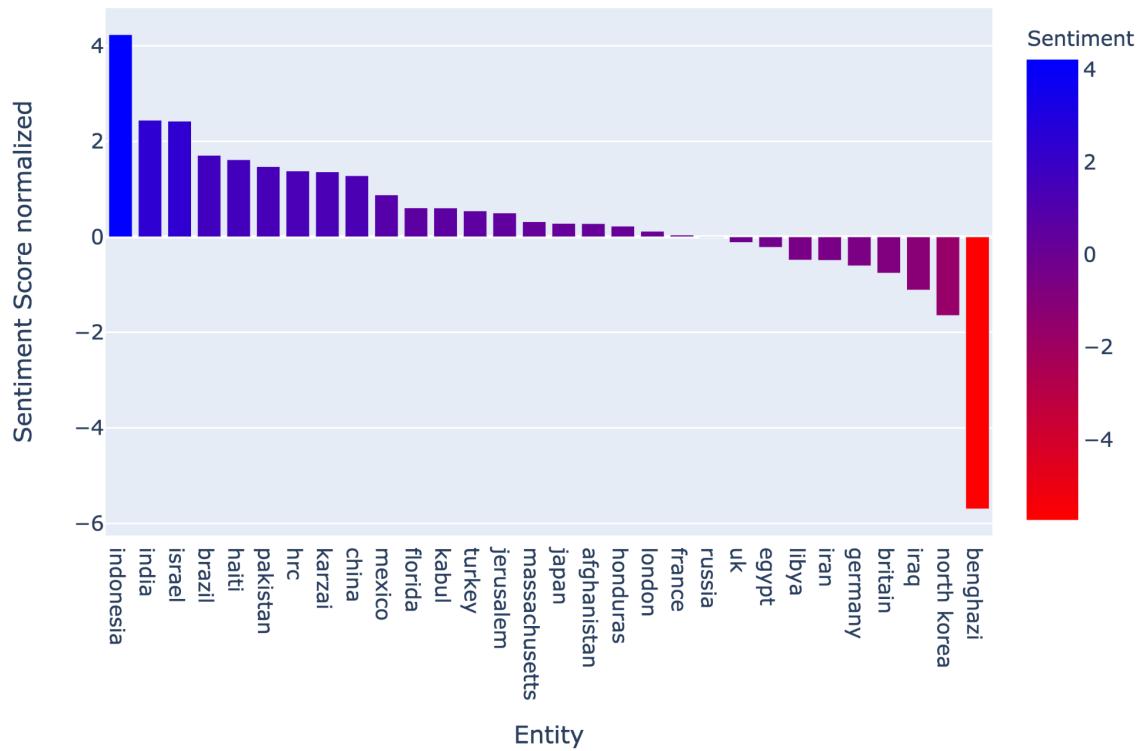
Successivamente, abbiamo sommato i valori del *sentiment* aggregandoli per entità. Per questo passaggio, abbiamo utilizzato l'assunto che se un'entità appare più di una volta all'interno della email, il suo contributo sia pari ad 1. Nel grafico seguente viene riportato il *sentiment* associato ad ogni entità e la relativa frequenza all'interno delle email.

### Sentiment delle entità (top 30 per frequenza)



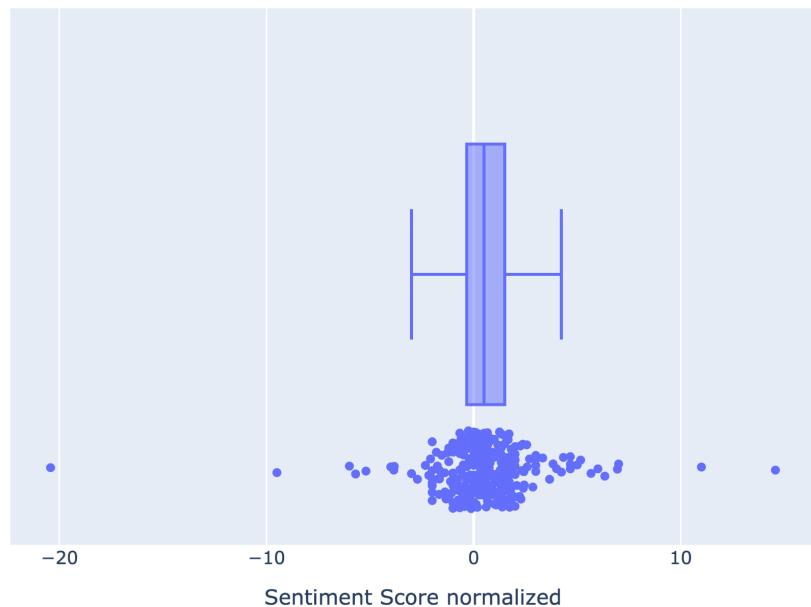
Da questo grafico emerge un aspetto importante: Il nostro approccio aggrega le informazioni sul *sentiment*, questo però porta a una distorsione del valore di *sentiment* per quelle entità che sono molto frequenti nelle email, in un certo senso il *sentiment* associato viene scalato in base alla frequenza. Per questo motivo si ritiene che normalizzare il *sentiment* per la frequenza associata possa meglio rappresentare come un' entità viene realmente considerata (positiva o negativa). L'operazione di normalizzazione permette quindi di ponderare eventuali squilibri, restituendo importanza relativa ad ogni entità analizzata. Pertanto per ogni entità il valore del *sentiment* viene diviso per la frequenza. Nel grafico sottostante viene mostrato il *sentiment* normalizzato ottenuto.

### Sentiment normalizzato delle entità (top 30 per frequenza)



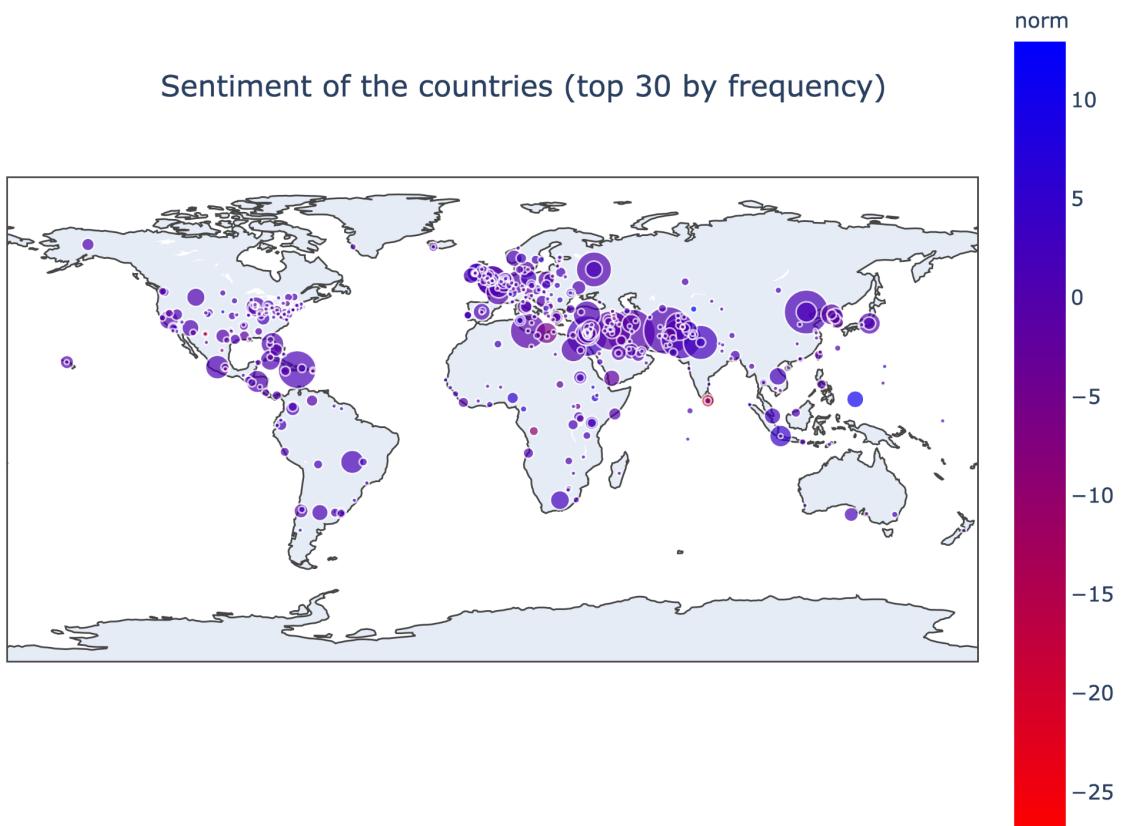
Nel grafico vengono riportate solo le prime 30 entità basandosi sulla frequenza. Per avere un'idea del *sentiment* generale associato viene qui riportato il boxplot del *sentiment* calcolato.

Boxplot of the normalized sentiment of the countries (frequency threshold 2)



Dopo aver determinato quali espressioni sono menzioni di entità, siamo passati alla fase di *entity linking*. Dopo aver confrontato i pro e i contro delle varie soluzioni per NEL, abbiamo optato per la gratuità DBpedia, vista la sua facilità di utilizzo. DBpedia è un database *open source* che contiene le informazioni presenti su Wikipedia ma organizzate in maniera strutturata, per questo motivo è permessa la formulazione di query per l'estrazione di queste informazioni, rendendo di conseguenza possibile una nuova integrazione fra pagine di wikipedia, data analytics e l'arricchimento di risorse sul web. Dall'altro lato, DBpedia Spotlight è uno strumento da noi utilizzato per formulare le query e fornisce le API sfruttabili con python per interrogare DBpedia.

Lo scopo di questa fase è associare ogni località e il relativo *sentiment* alle sue coordinate geografiche in modo da poter tracciare su mappa i risultati. Per trasmettere queste informazioni con un grafico, ogni paese, città o luogo viene riportato nella relativa posizione sulla mappa, inoltre il *sentiment* viene codificato con il colore e la frequenza per mezzo della dimensione del cerchio associato.



# Topic Modeling

## Considerazioni sul dominio

Un'ulteriore analisi che è stata condotta riguarda il topic modeling. Lo scopo di questa indagine è cercare di identificare i diversi topic all'interno delle email, distinguendo tra topic con *sentiment* positivo e negativo. Per topic si intende un insieme di parole (i tokens) ciascuna delle quali ha una certa probabilità di appartenere al del topic stesso. I topic sono pertanto i diversi aspetti che vengono trattati nelle email e sono composti da un insieme di parole del corpus. L'assunzione di partenza è che nelle diverse email vengono trattati alcuni topic, lo scopo è di ottenere la distribuzione di probabilità per ogni token di essere presente in un topic, pertanto ci aspettiamo che i token più probabili per uno specifico topic meglio caratterizzino il topic stesso così da dedurre i diversi aspetti trattati nelle email.

Questo problema viene approcciato tramite tecniche non supervisionate, andando a creare modelli generativi con lo scopo di campionare distribuzioni di probabilità per associare ad ogni parola la probabilità di comparire in un topic il quale ha associato a sua volta una specifica polarità del *sentiment* (nel nostro caso positivo o negativo).

Esistono due principali modelli per fare topic modeling: JST o ASUM. La differenza principale tra questi due modelli è che JST assume che un intero documento (nel caso preso in esame un mail) sia associato ad un solo topic. ASUM invece risulta essere un'estensione di JST, dove per ogni documento vengono considerate le diverse frasi e per ognuna di queste viene associato un topic. Pertanto un documento può trattare diversi topic.

Come visto in precedenza dalla distribuzione dei tokens per email, la maggior parte di queste è composta da un numero di tokens molto basso. Tuttavia questo aspetto non è vero in generale, infatti nel corpus sono presenti delle mail piuttosto lunghe, nelle quali ci aspettiamo che vengano trattati diversi aspetti. Sulla base di ciò viene giustificato l'utilizzo di ASUM rispetto a JST.

## Creazione del modello

### Implementazione del modello

La creazione del modello avviene tramite l'utilizzo di un programma scritto in Java. Come mostrato nella [Struttura del Codice](#), la specifica implementazione di ASUM che è stata utilizzata viene distribuita sotto forma di codice java compilato. Andremo quindi ad eseguire la classe che ha la funzione di entry point per il programma per mezzo della libreria python subprocess.

### Preprocessing e file di input

Il modello da creare ha bisogno dei relativi file di input corrispondenti alla codifica delle email in un formato gestibile dal programma. Questa specifica implementazione richiede di creare 4 diversi file:

1. **WordList.txt:** Dizionario del corpus (insieme delle parole univoche che costituiscono le email)
  - a. Ogni riga contiene esattamente un token
  - b. L'ordine dei token all'interno di questo file è molto importante
  - c. Viene creato indirettamente un mapping da token a numeri interi, il numero associato al token rappresenta l'ID del token
  - d. L'associazione token-ID avviene in questo modo:
    - i. [RIGA 1 file]: {token\_A} → ID 0
    - ii. [RIGA 2 file]: {token\_B} → ID 1
    - iii. [RIGA 3 file]: {token\_C} → ID 2
    - iv. ecc..
2. **SentiWords-0.txt:** Insieme di token con funzione di seed per il *sentiment* con polarità positiva
  - a. Un token per riga
3. **SentiWords-1.txt:** Come sopra ma per polarità negativa
  - a. Un token per riga
4. **BagOfSentences.txt:** Codifica delle email per mezzo del dizionario costruito nel file WordList.txt
  - a. All'interno di questo file viene codificato ogni documento
  - b. Per ogni documento si ha a disposizione la lista delle  $N$  frasi che compongono il documento
    - i. Ogni frase è rappresentata come una tupla andando a rappresentare i token sfruttando il dizionario costruito a partire dal file WordsList.txt
  - c. La codifica del singolo documento viene rappresentata in questo modo:
    - i. [Riga 1 file]:  $N$  (numero intero che rappresenta il numero di frasi nel documento)
    - ii. [Riga 2 file]: token\_id\_1 token\_id\_2 token\_id\_3 ... (tupla composta dagli IDs dei token che compongono la prima frase del documento originale)
    - iii. [Riga 2 file] token\_id\_4 token\_id\_5 token\_id\_6 .... (tupla composta dagli IDs dei token che compongono la seconda frase nel documento originale)
    - iv. ...
      - v. [Riga N + 1 file] ...
        - vi. [Riga N + 2 file] Codifica nuovo documento: vedi passo 4.c.i.

Viene qui discussa la creazione del file **WordList.txt**. A partire dalle email dopo che è stato effettuata una prima fase di cleaning delle email (ovvero la rimozione degli header spiegata qui [Preprocessing](#)) si passa alla tokenizzazione. Successivamente è stato necessario riconoscere le diverse frasi all'interno di ogni documento. Per questo passo abbiamo utilizzato la funzione `sent_tokenize` della libreria NLTK. Il funzionamento è un banale split del documento basandosi sui caratteri di punteggiatura qui riportati (';', ',', '.', '!', '?'), facilmente implementabile manualmente.

Una volta identificate le frasi di ogni documento, si passa alla tokenizzazione di ognuna di queste per ogni documento. La tokenizzazione è del tutto analoga a quanto spiegato nel

capitolo di [Preprocessing](#). Tuttavia viene effettuato come passo aggiuntivo l'operazione di *stemmatizzazione*. Questa scelta viene fatta in quanto non si è più vincolati, a differenza dell'approccio basato su lessico, ad un insieme di parole (appunto il lessico di AFINN) che non sono state stemmatizzate, al contrario si vogliono mappare diverse forme e declinazioni della stessa parola ad una scrittura unica, si noti come la forma finale stemmatizzata non necessariamente deve essere una reale parola della lingua di partenza in quanto il fattore importante è il simbolo rappresentato dal token stemmatizzato. Questo viene fatto per evitare che all'interno di un topic, a seguito del campionamento della distribuzione di probabilità congiunta del modello, vengano inserite le diverse parole che in realtà fanno riferimento ad un unico termine declinato in modi differenti, o ancora peggio, per evitare che declinazione diverse della stessa parola vengano inserite in topic differenti.

Completate queste operazione vengono creati i file di input per il modello così come specificato a inizio capitolo.

Un aspetto molto importante è che andando a stemmatizzare i tokens delle email, questa operazione deve essere coerente per i token che hanno funzione di seed per distinguere tra *sentiment* positivo e negativo. Per questo motivo tutti i token presenti nei file “SentiWords-0.txt” e “SentiWords-1.txt” vengono anch'essi stemmatizzati.

## Scelta degli iperparametri

La creazione del modello dipende da alcuni iperparametri che vanno specificati manualmente. Vengono qui elencati gli iperparametri e giustificate le loro scelte:

- **Numero di topic: 5**
  - Questo parametro indica il numero di topic per ogni *sentiment* che si vuole specificare. Viene scelto 5 come numero di topic in quanto un numero troppo elevato non permette di ben caratterizzare i diversi aspetti rilevati nelle email.
- **Numero di sentiment: 2**
  - Numero di *sentiment* considerati, questo valore deve essere coerente con il numero di file “SentiWords-X.txt”, dove X indica l'i-esimo *sentiment* e quindi i token e i topic associati a quello specifico *sentiment*. Viene scelto 2 in quanto si vogliono considerare topic associati a *sentiment* con polarità positiva o negativa, senza considerare altre sfumature.
- **Numero di iterazioni: 1000**
  - Numero di iterazioni utilizzate per il processo di inferenza sulla rete bayesiana (si pensi ad esempio ad una iterazione dell'algoritmo “*Markov chain Monte Carlo*” per effettuare inferenza approssimata). Viene scelto 1000 in quanto risulta essere un buon trade-off tra convergenza al valore esatto e tempo di calcolo.
- **Alpha: 0.1**
  - Questo parametro permette di specificare la probabilità a priori relativa ai topic. Modificare questo parametro permette di indicare quanto è probabile che in una frase si parli di un determinato aspetto. Dal momento che assumiamo che le mail siano piuttosto generali e non tutte relative ad uno

specifico aspetto, ciò si riflette anche sulle singole frasi del documento. Per questo motivo si sceglie un valore di alpha relativamente basso.

- **Beta:**

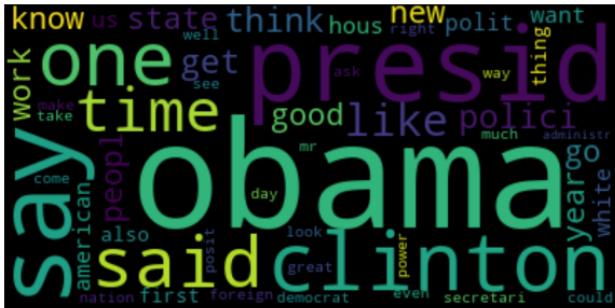
- Probabilità a priori che una singola parola all'interno di una frase, a sua volta all'interno di un documento, faccia parte di un topic con associato uno specifico *sentiment*. Per come è definita la rete bayesiana su cui è stato progettato ASUM, ogni topic è condizionalmente dipendente dal *sentiment*. Per questo motivo infatti è necessario specificare  $N + 1$  valori, dove  $N$  indica il numero di *sentiment* specificati.
  - In questo caso quindi è necessario specificare 3 valori, ognuno di questi regola uno specifico comportamento:
    - **Parametro 1** (0.001): Probabilità a priori che una parola appartenente a un seed positivo sia presente in un aspetto negativo
    - **Parametro 2** (0.001): Probabilità a priori che una parola appartenente a un seed negativo sia presente in un aspetto positivo
    - **Parametro 3** (0.1): Probabilità a priori di tutte le altre parole
  - Per il parametro 1 e 2 vengono scelti valori bassi in quanto ci aspettiamo che una parola con seed negativo è molto poco probabile che sia presente in un aspetto con polarità positiva, ragionamento analogo viene fatto per la polarità opposta.
- **Gamma: 1 / 1**
- Probabilità a priori di uno specifico *sentiment*. I valori per questo parametro devono essere tanti quanto il numero di *sentiment* considerati. In questo caso vengono utilizzati i valori 1, 1 dal momento che non si ha una probabilità a priori del *sentiment* all'interno dei documenti.

## File di output

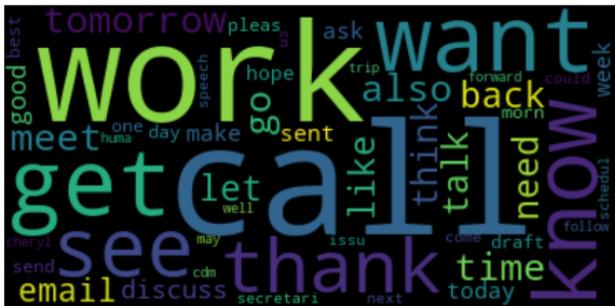
Una volta aver creato il modello, viene prodotto in output il file “*ProbWords.csv*”. Questo file contiene le probabilità delle parole di appartenere ad uno specifico aspetto relativo ad uno specifico *sentiment*. Sfruttando questo file abbiamo creato tante word cloud quanti sono i topic relativi ad ogni *sentiment*. Nelle diverse word cloud vengono riportate le prime 50 parole più probabili dove la dimensione di ogni parola rappresenta la probabilità che un token sia presente in uno specifico topic, più la parola è grande maggiore è la probabilità.

## Risultati

Positive Sentiment - Topic 0



Positive Sentiment - Topic 1



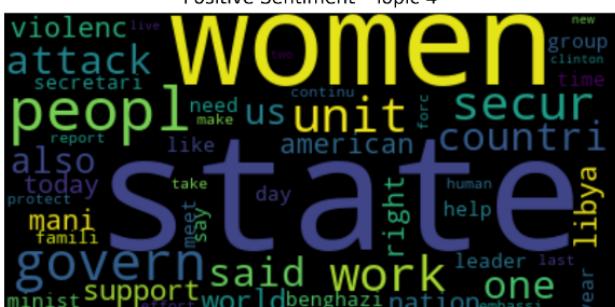
Positive Sentiment - Topic 2



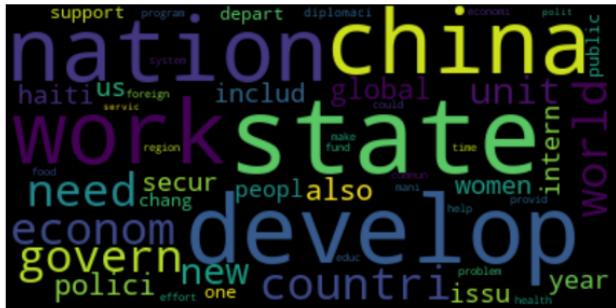
Positive Sentiment - Topic 3



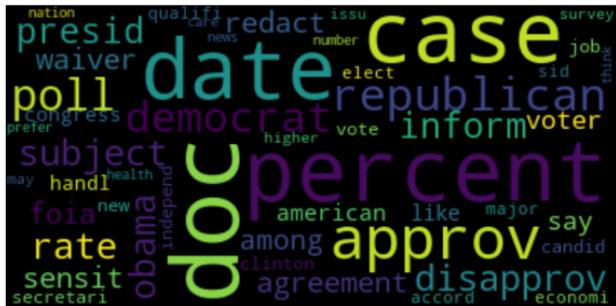
Positive Sentiment - Topic 4



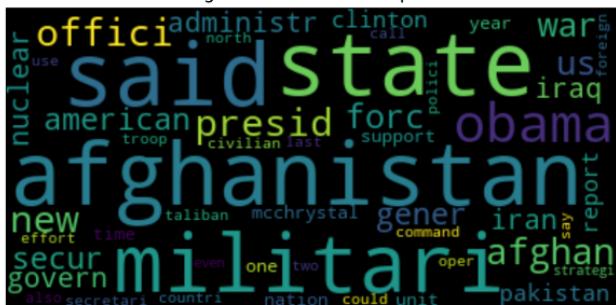
## Negative Sentiment - Topic 0



Negative Sentiment - Topic 1



Negative Sentiment - Topic 2



Negative Sentiment - Topic 3



Negative Sentiment - Topic 4



## Topic estratti

Positive Sentiment		Negative Sentiment	
<b>Topic 0</b>	Internal relations of the party	<b>Topic 0</b>	China
<b>Topic 1</b>	Office work	<b>Topic 1</b>	Industrial support
<b>Topic 2</b>	Press	<b>Topic 2</b>	Afghanistan Conflict
<b>Topic 3</b>	Israeli-Palestinian Conflict	<b>Topic 3</b>	British relations
<b>Topic 4</b>	Women	<b>Topic 4</b>	Against Republicans

Interessante notare come il topic "Women" venga associato ad un *sentiment* positivo pur essendoci "violence" o "attack". Questo avviene in quanto ASUM identifica i due token inseriti all'interno di un contesto che ha un tono generalmente positivo, calcolando la probabilità che questi appartengano al Topic 4. Perciò, tutte le volte che nelle mail si parla di sicurezza e di diritti delle donne, compare anche la parola "violence", come ad esempio (mi sembra strano ma non saprei darmi altra giustificazione). Infatti nello stesso topic troviamo "work", "secure" e "right" (versione stemmatizzata di rights), rendendo positivo il *sentiment* finale associato. I due token "benghazi" ed "embassi" (versione stemmatizzata di embassy) sono nominati relativamente poche volte rispetto agli altri token presenti nel contesto, ma sono comunque vicini al tema "violence". Questo porta ASUM ad identificare come "sufficiente" la probabilità che questi ultimi due token possano essere inseriti nel Topic 4, pur presentando un significato non riconducibile al positivo. Supponiamo inoltre che questi token vengano ampiamente utilizzati per identificare sia tematiche positive legate ai diritti delle donne (con *sentiment* maggiormente positivo), sia alle attività di violenza fisica tipiche della guerra (con *sentiment* negativo). Calcolate le probabilità relative e pesata l'importanza dei token, ASUM assegna un *sentiment* positivo finale al Topic 4.

# Conclusioni

In seguito all'analisi effettuata, è stato possibile rispondere con efficacia a tutte e tre domande di ricerca poste in fase di progettazione. In merito alla comprensione dell'indirizzo di politica estera tenuto dalla Clinton e dalla sua rete di contatti, è stato possibile identificare le località maggiormente citate e l'atteggiamento verso esse rivolto. In particolare, sono stati identificati interessi militari variegati concentrati maggiormente in Medio Oriente, con Israele, Palestina, Pakistan, Afghanistan, Iran e Iraq come paesi più importanti. Siano essi associati a *sentiment* positivo o *sentiment* negativo, gli interessi nella zona sono spalmati su un'ampia fascia temporale, andando ad interessare anche l'ultimo periodo della carica di Segretario di Stato, con particolare interesse per l'attentato di Benghazi del 2012. I rapporti con la Cina appaiono controversi e un rapido confronto con la stampa dell'epoca conferma che le tensioni sono durature, andandosi a consolidare anche nel periodo successivo all'amministrazione Obama. Per quanto riguarda paesi minori, o per i quali non è presente un diretto interesse militare da parte degli Stati Uniti, dalle email emerge come vi sia una solida presenza in questioni diplomatiche di altri paesi minoritari esterni al Medio Oriente, come Haiti, Sri Lanka, Palau e altri.

Analizzando i risultati del topic modeling, i temi trattati dalla Clinton riguardano appieno il suo ruolo come Segretario di Stato, concentrandosi maggiormente sull'immagine pubblica degli Stati Uniti, sia all'estero che in patria.

Infine, quanto emerge sia dalle word clouds che dalle statistiche generali sul dataset conferma che il server fosse utilizzato per necessità lavorative, indipendentemente dal giorno della settimana e dai temi trattati.

L'idea generale che emerge dal dataset è che la Clinton abbia mantenuto rapporti stretti con i propri fidati per la discussione e l'organizzazione di tematiche prettamente istituzionali. Il dataset non consente ulteriori speculazioni sulla natura riservate delle informazioni omesse.