TEXT MINING

SIGNORI ELENA –
MATR: 843017

È stato analizzato l'insieme dei discorsi di Donald Trump da giugno 2015 a novembre 2016 nei vari stati degli Stati Uniti.
L'obiettivo dell'analisi è stato quello di osservare quali argomenti sono stati toccati dal candidato nel corso della sua campagna presidenziale per convincere gli elettori.
È stata comunque svolta anche una breve analisi sui discorsi, per osservare se in alcuni stati si siano tenuti discorsi simili.

Dopo aver importato i discorsi è stato creato un oggetto di tipo «corpus» con le seguenti caratteristiche:
<<SimpleCorpus>>
Metadata: corpus specific: 1, document level (indexed): 0
Content: documents: 56

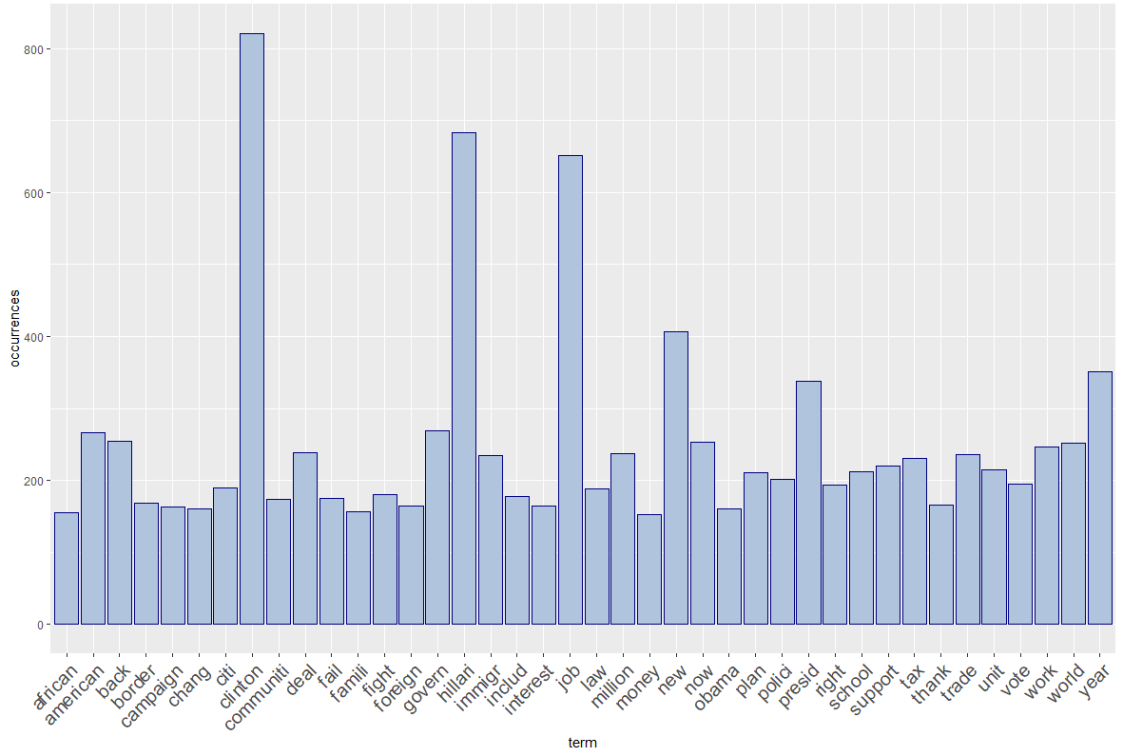
Dove ogni documento ha:
Length: 2
Class: PlainTextDocument
Mode: list

PRE-PROCESSING

Sono state effettuate le operazioni di pre-processing e data cleaning dei dati quali: rimozione di simboli, rimozione della punteggiatura, rimozione dei numeri, rimozione di spazi bianchi, rimozione delle stopwords e stemming.
È stato poi eseguito il completamento delle parole tramite la funzione stemCompletion, ma tale passaggio non è stato utilizzato per l'analisi sui discorsi poiché restituisce un unico testo (PlainTextDocument).

ANALISI PRELIMINARI

Dopo aver creato la matrice TermDocument, considerando solo termini composti da 3 a 27 lettere, è stato creato un grafico per visualizzare le parole più frequenti.



È stata poi calcolata l'associazione tra le parole più frequenti e tutte le altre, scegliendo tra le parole presenti nel grafico quelle più significative per il contesto che stiamo analizzando.

N.B: le associazioni sono state calcolate prima di applicare la funzione stemCompletion, perché altrimenti la funzione findAssocs non funziona, e le parole sono state completate ai fini della presentazione. Lo stesso vale per la costruzione del grafico a barre.

Associazioni:

- Obama: president, world, worse, aggressive, short, weak, bomb, nato, enemy, isis.
- Tax: property, overregulation, shut, collapse, growth, depress, scale, economy.
- Military: union, disarray, easy, smaller, tension, base, nuclear, weapon, combat, adversary, asia.
- Job: trade, regulation, million, manufactur, model, currency, repeal, tax.
- Border: illegal, open, deport, sanctuary, cheat, criminal, gang.
- Policy: ally, vision, war, aimless, atrophy, awake, blaze, caution, chaotic, cheapest

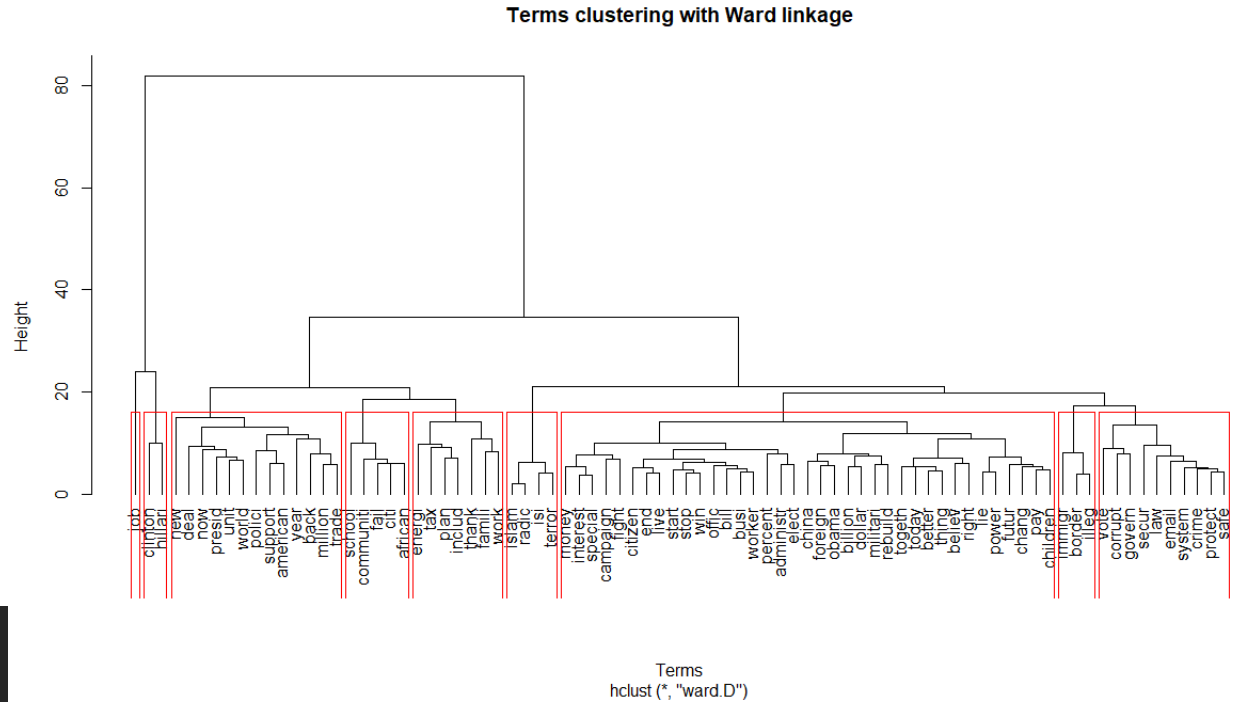
CLUSTERING PER TERMINI

Per rappresentare il dendogramma e i gruppi, è stata calcolata una matrice delle distanze euclidee e i clusters sono stati individuati usando il linkage di Ward. Alcuni clusters contengono solo uno o due termini; è stato scelto di non eliminare quei termini dall'analisi poiché rientrano tra quelli più frequenti.

Infatti, come si può vedere dal dendogramma, il primo cluster contiene solo la parola «job» e il secondo «hillary» e «clinton». A partire dal terzo i cluster sembrano avere più senso.

- Cluster 3: sembra contenere termini riguardanti gli americani, il senso di unità e il supporto reciproco e il loro ruolo nel mondo (world, unit, american, president, support, now).
- Cluster 4: contiene solo tre termini legati all'immigrazione e ai confini (border, illegal, immigration),
- Cluster 5: si potrebbe dire che riassume in pochi termini alcuni temi generali toccati da Trump durante la campagna presidenziale (plan, energy, tax, include).
- Cluster 6: riguarda l'argomento del terrorismo (islam, radical, isis, terror).
- Cluster 7: è il cluster di maggiori dimensioni; sintetizza in modo più dettagliato rispetto al cluster 5 i discorsi di Trump. Si può dire che il candidato ha promesso un cambiamento, sicurezza, protezione. All'interno del gruppo la parola china è vicina a termini che permettono di ipotizzare che Trump possa aver parlato di nuove politiche economiche nei confronti di questo paese (policy, foreign, china, billion), ma anche di interventi militari (military).
- Cluster 8: si tratta di un gruppo di difficile interpretazione. Contiene termini legati alla scuola, alla comunità, agli africani ma anche al fallimento (school, community, african, fail, city).
- Cluster 9: l'ultimo cluster contiene parole che sembrano incitare gli americani a fare il loro sostegno a Trump. Sembrano essere parole che potrebbero essere usate per concludere il discorso (together, better, today, believe, right, thank).

È stato scelto di fare 9 clusters invece che solo 5 (come suggerito dalla funzione pamk) in quanto sarebbero stati molto difficili da interpretare. Inoltre, anche con 5 clusters, due sarebbero stati costituiti dalle parole «job» e «clinton» e «hillary.



I gruppi di termini sono stati rappresentati anche attraverso una rappresentazione wordcloud dato che questo tipo di grafico tiene conto anche della frequenza di ogni parola. Escludendo i primi due clusters, la cui rappresentazione non avrebbe molto senso, si sono ottenuti i seguenti grafici:

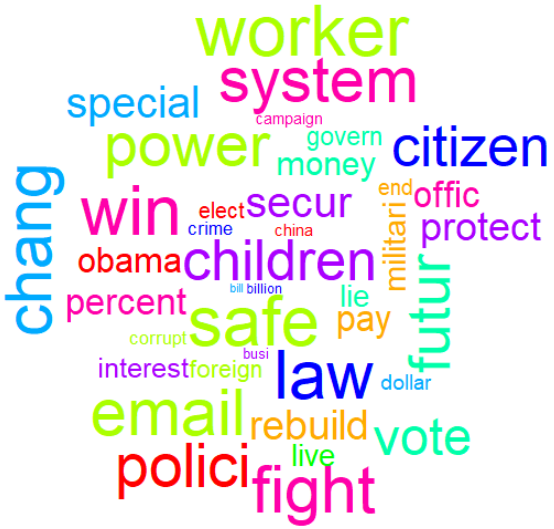
CLUSTER 3

CLUSTER 4

CLUSTER 5

CLUSTER 6

CLUSTER 7



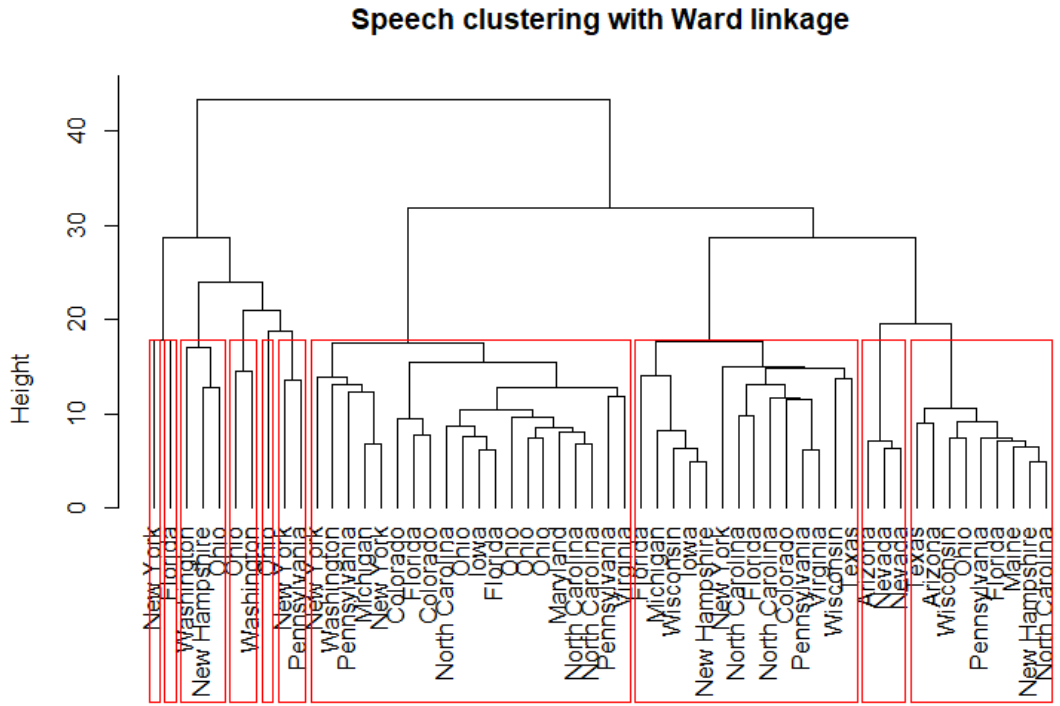
CLUSTER 8

CLUSTER 9

CLUSTERING PER DISCORSI

Per completezza si è svolto anche il clustering per i discorsi fatti da Trump nei vari stati. I discorsi sono prima stati rinominati, attraverso una funzione, in modo che ad ognuno fosse assegnato il nome dello stato in cui è stato pronunciato.
È stato ottenuto il dendogramma presente in questa slide e il numero di gruppi è stato scelto attraverso la funzione pamk.
Si sono ottenuti 10 clusters.

Saltando le possibili considerazioni sui primi sei gruppi che sono di dimensioni minime, si può osservare che il cluster 7, quello di dimensioni maggiori, contiene il nome di stati che si trovano nella zona est degli USA (ad eccezione del Colorado e Washington). Questo significa che Trump ha pronunciato discorsi simili in stati vicini.
Anche il cluster seguente contiene stati che si trovano ad est, ad eccezione di Texas e Colorado. Per fare considerazioni più approfondite sarebbe utile riuscire ad attribuire anche una data ad ogni discorso per poter osservare se, nel corso della campagna presidenziale, il candidato abbia adattato i propri discorsi in base all'andamento del sostegno degli elettori.
Il cluster 9 contiene solo Arizona e Nevada; si tratta di due stati confinanti che potrebbero essere stati interessati da problemi simili e, per questo, sono stati trattati gli stessi argomenti durante i discorsi.
L'ultimo cluster contiene stati che si trovano in zone diverse degli USA quindi, per poter trarre conclusioni più approfondite sarebbe necessario indagare ulteriormente.



NOTA: la funzione stemCompletion non è stata applicata in nessuna parte dell'analisi perché, nonostante completi correttamente le parole, ne moltiplica la frequenza rendendo inutilizzabili i termini (alcuni arrivano ad avere frequenza superiore a 40,000)