TASK 2
TRUMP TWEETS
TEXT ANALYSIS

GROUP #4

GOTTA CAMILLA – IERARDI ANDREA – LAZZARA FRANCESCO – LENI THOMAS – LEPEK ALEKSANDRA

PROJECT GOAL

The objective of the project is to analyze Trump tweets in order to define the most used words and to classify them in categories thanks to SVD diagrams and clustering methods

DATASET USED

The dataset is composed by 4212 tweets, with only one column for each one, containing the text of the tweets.

DATA CLEANING

- ➢Remove punctuation, numbers and isolate words;
- ➢Remove all the words without an utility for the project;
- ➢Cutoff all the words below an arbitrary thresholds, in this case study the threshold is 30;
- ➢Combine and insert some words particurarly relevant for the topic (e.g. “make America great again”);

Numero di termini	Numero di casi	Token totali	Token per maiuscola/minuscola	Numero di casi non vuoti	Suddividi non vuoti per caso
286	4212	132564	31,4729	3717	0,8825

Figure 1- Number of terms after data cleaning

MOST COMMON WORDS

In Figure 2 and Figure 3 there is a representation of the most used words in the tweets such as «great», «people» and «president».

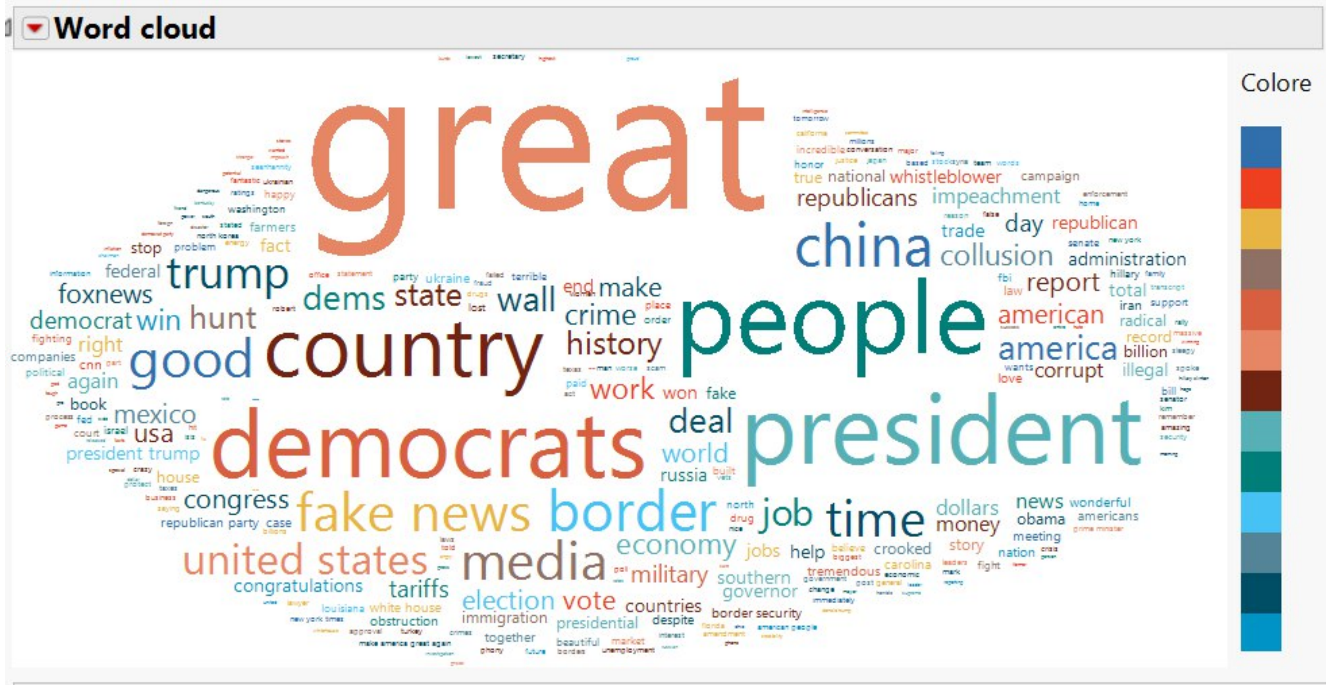
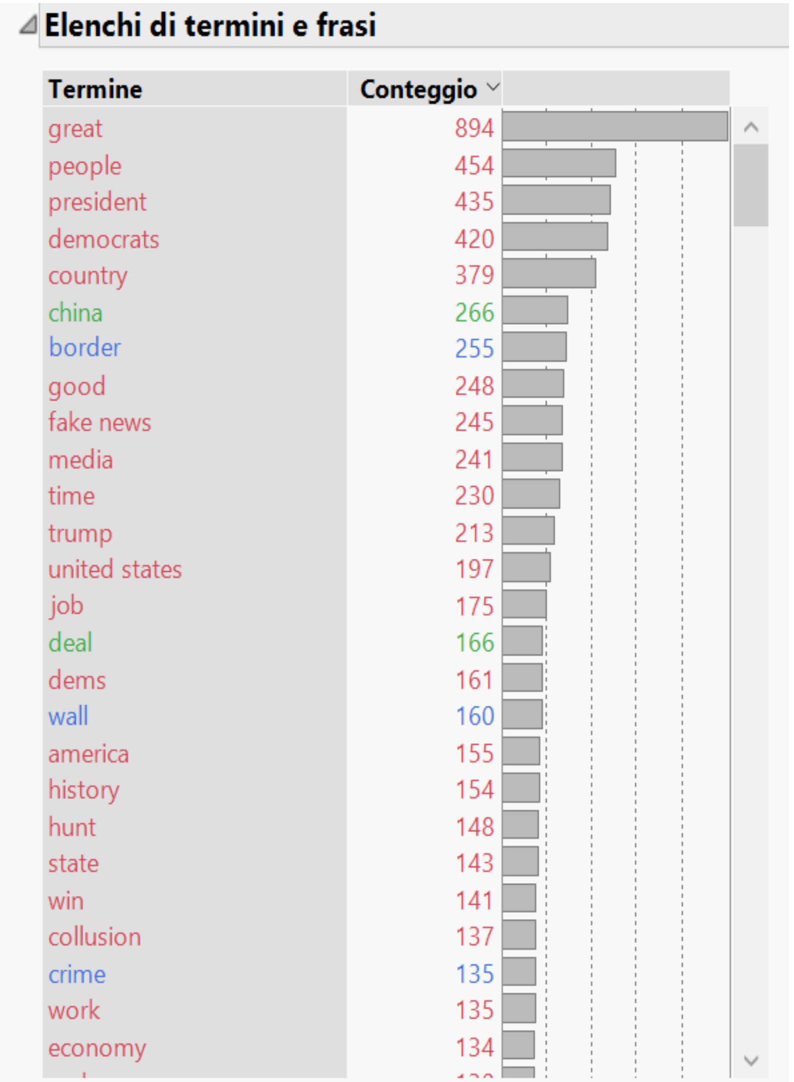


Figure 2- Word Cloud



Elenchi di termini e frasi

Termine	Conteggio
great	894
people	454
president	435
democrats	420
country	379
china	266
border	255
good	248
fake news	245
media	241
time	230
trump	213
united states	197
job	175
deal	166
dems	161
wall	160
america	155
history	154
hunt	148
state	143
win	141
collusion	137
crime	135
work	135
economy	134

Figure 3- List of words

SVD PLOTS

- ➢ The words appearing close to each other appear together frequently in documents in the corpus;
- ➢ In the Figure 4 the two plots have a similar shape, even if there is a difference in the concentration deriving from the number of points in each plot;
- ➢ In the first plot there are some outliers, the clearest is reported beside:

Getting ready to land in Louisiana to do Rally with Great Republican @EddieRispone for Governor. He will
get your taxes and auto insurance (highest in Country!) way down. Loves our Military & Vets. Will
protect your 2nd Amendment. VOTE SATURDAY! [81]

Figure 5- Outliers doc

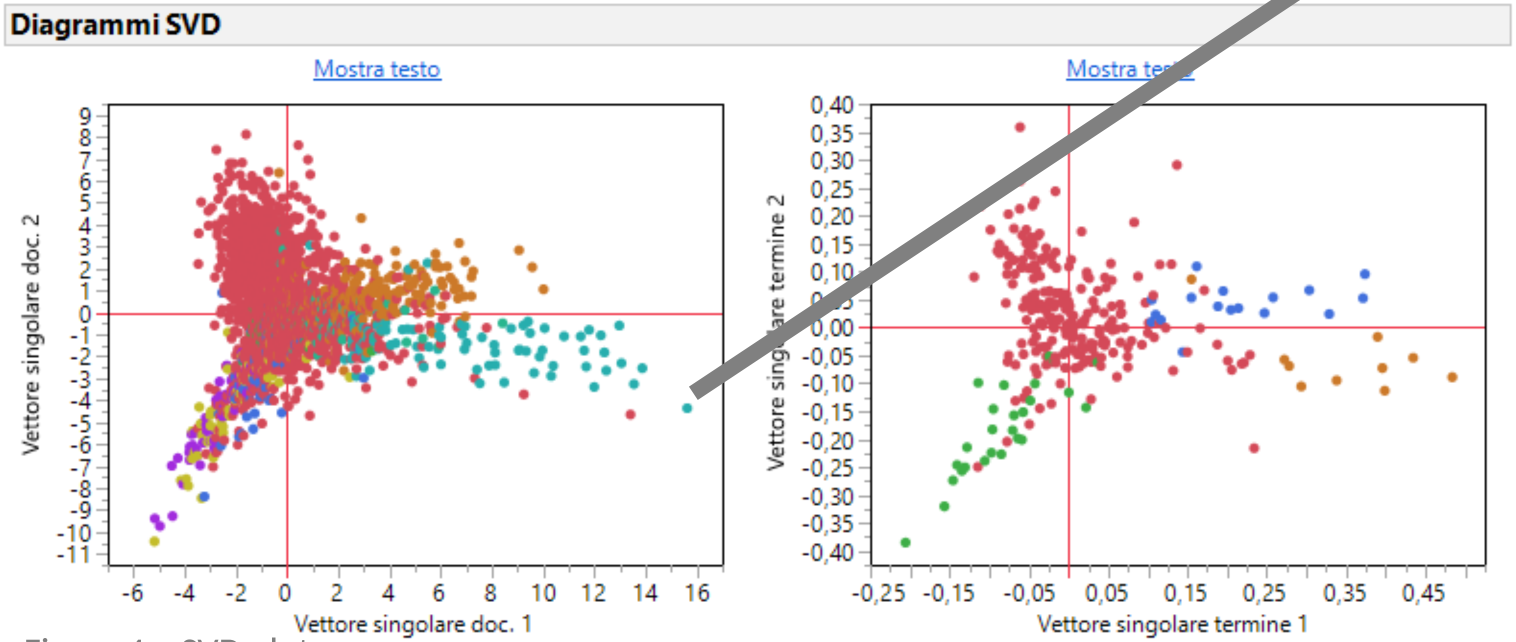


Figure 4 – SVD plots

TOP LOADINGS BY TOPIC

This fuction matches the words, according to the different topics. In particular they are divided in 10 groups:

1. Election campaign;
2. Immigration and Mexico;
3. Trump’s opponents;
4. Import-export;
5. North Korea troubles;
6. Mass media;
7. Impeachment case;
8. Economy;
9. Domestic economy;
10. Middle east conlicts:

Pesi principali per argomento

Argomento 1		Argomento 2		Argomento 3		Argomento 4		Argomento 5	
Termine	Caricamento in corso	Termine	Caricamento in corso	Termine	Caricamento in corso	Termine	Caricamento in corso	Termine	Caricamento in corso
amendment	0,60192	border	0,65366	collusion	0,57636	dollars	0,63248	north korea	0,74222
governor	0,53740	southern	0,61894	crooked	0,48231	china	0,62040	kim	0,72570
vote	0,52517	immigration	0,48710	hillary	0,42930	tariffs	0,60712	chairman	0,57238
vets	0,49576	laws	0,46176	obstruction	0,39702	billion	0,53970	korea	0,53670
republican	0,47895	mexico	0,41224	report	0,36354	billions	0,46402	south	0,49246
louisiana	0,46703	drugs	0,40894	campaign	0,29530	deal	0,34515	meeting	0,37920
taxes	0,35827	wall	0,40089	democrats	0,28750	trade	0,31270	economic	0,29958
military	0,35721	fix	0,31663	hillary clinton	0,28144	farmers	0,29195	potential	0,29531
protect	0,34195	illegal	0,30007	russian	0,27721	usa	0,28774	japan	0,22036
north	0,33836	crisis	0,29303	trump	0,26168	paid	0,26484		
carolina	0,33579	democrats	0,28214	russia	0,25702	companies	0,26457		
crime	0,31811	stop	0,27975	angry	0,25188				
rally	0,29108	built	0,27909	fbi	0,24775				
		change	0,27908	intelligence	0,24682				
				hunt	0,24670				

Argomento 6		Argomento 7		Argomento 8		Argomento 9		Argomento 10	
Termine	Caricamento in corso	Termine	Caricamento in corso	Termine	Caricamento in corso	Termine	Caricamento in corso	Termine	Caricamento in corso
fake news	0,52877	ukrainian	0,5154	interest	0,70297	market	0,6117	kurds	0,76609
media	0,48130	conversation	0,4542	rates	0,65723	stock	0,6012	turkey	0,71581
post	0,45755	ukraine	0,4431	inflation	0,56335	unemployment	0,4078	syria	0,61736
washington	0,41521	phone	0,4296	federal	0,54365	history	0,3965	isis	0,49777
new york times	0,37590	whistleblower	0,4202	fed	0,49691	economy	0,3507	fight	0,37003
story	0,37563	transcript	0,4017	dollar	0,42718	lowest	0,2805	fighting	0,27629
fake	0,36678	president	0,3997	rate	0,34122	impeach	0,2727	protect	0,19499
corrupt	0,35700	scam	0,3334	countries	0,31436	country	0,2124		
failing	0,33712	democrat	0,2553			record	0,2117		
stories	0,29421	impeachment	0,2443			enforcement	-0,2047		
cnn	0,29065	great	-0,2253			law	-0,2045		
reporting	0,29049					biggest	0,2025		
phony	0,26399								

Figure 6- Top Loading Topic

CLUSTERING

The dataset is divided in 4 cluster, in this way is possible to delete non significant groups. It's evident that the biggest group is the red one, since the most common terms belong to this group.

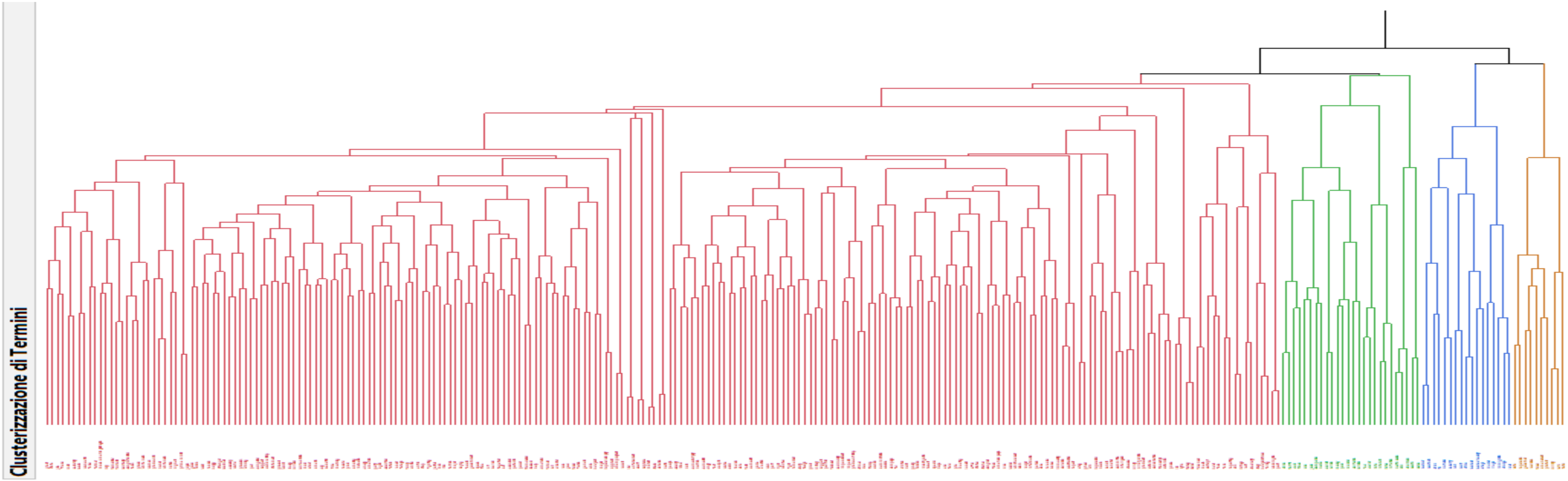


Figure 7- Clustering Terms