Sentiment Analysis of Twitter data before 2016 presidential election

Name(s): Chongyang Zhu

Problem or Question

- (1) Is there difference between the sentiment of tweets talking about Clinton VS Trump?
- (2) Does users have more followers post tweets with stronger sentiment? (or the opposite)
  - (a). If so, positive or negative?
  - (b). Is there differences for Clinton VS Trump?
- (3) Do the topics for Clinton VS Trump show any insights on the positive or negative sentiment?

Related Background

- Social media may contribute to heightened levels of political extremism
  - Hong, S., Kim S. H. (2015 ,October 7). *Political polarization on twitter: Implications for the use of social media in digital governments*. Government Information Quarterly. Volume 33, Issue 4, Pages 777-782.
- Topic modeling and sentiment analysis of twitter data on climate change
  - Dahal, B., Kumar, S. A. P., Li, Z. (2019, June 10). *Topic modeling and sentiment analysis of global climate change tweets*. Social Network Analysis and Mining. **9,** 24. https://doi.org/10.1007/s13278-019-0568-8

Data

- About the data
  - Twitter data before 2016 presidential election.
  - Data time range from 2016-10-05 to 2016-10-12
  - 82647 tweets random samples after cleaning

	Keyword	Handle	TimeStamp	TweetID	Text	Followers	WC
0	Clinton	CrabbyAbbey1	Fri Oct 07 00:00:06 +0000 2016	7.841815e+17	@SarahKSilverman If you have to lie to win you...	1	16
1	Clinton	Marianne_Summer	Fri Oct 07 00:00:26 +0000 2016	7.841816e+17	Here's how much Hillary Clinton paid in taxes ...	2	11
2	Trump	SAcurrent	Fri Oct 07 00:00:32 +0000 2016	7.841816e+17	Protesters at Trump's SA Fundraiser Were Passi...	78082	11

- Methods to use in this study
  - Sentiment analysis
  - Topic model

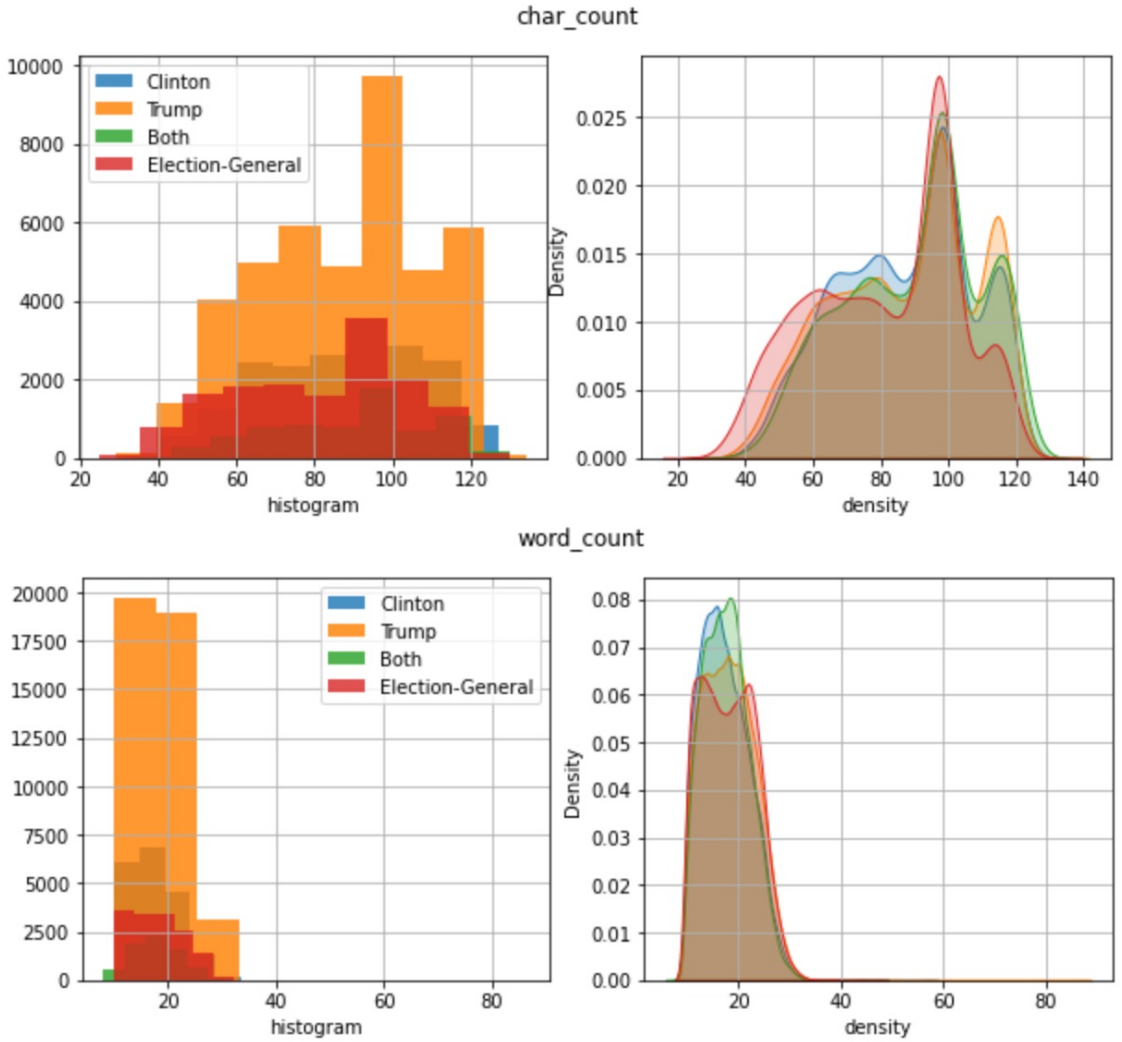
Preliminary Analysis

- Baseline comparison
  - Count of tweets shows that the number of tweets about Trump is much more than the number of tweets about Clinton.
  - The distribution of character counts and word counts are not too different, so it's comparable.

```
df.groupby('Keyword')['TweetID'].count()
✓ 0.2s

Keyword
Both                 7052
Clinton             18999
Election-General    14734
Trump               41862
Name: TweetID, dtype: int64
```

Count of tweets

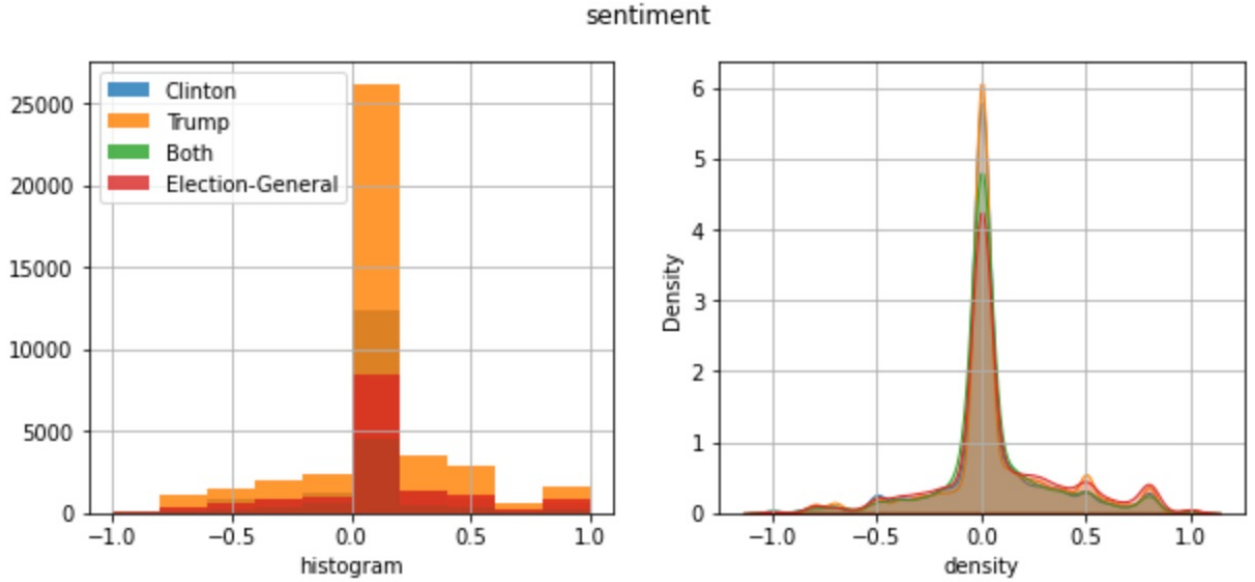


Distribution of character/word counts

Preliminary Analysis

Distribution of sentiments by keywords

- Sentiment analysis
  - All groups shows positive sentiment in general.
  - Answered Question (1): There is difference between the sentiment of tweets talking about Clinton VS Trump. Sentiment for Trump shows stronger positive sentiment (p-value<0.05) than for Clinton.



Keyword	count	mean	std	min	25%	50%	75%	max
Both	7052.0	0.040446	0.276128	-1.0	0.0	0.0	0.100000	1.0
Clinton	18999.0	0.035836	0.279659	-1.0	0.0	0.0	0.068182	1.0
Election-General	14734.0	0.073691	0.318641	-1.0	0.0	0.0	0.200000	1.0
Trump	41862.0	0.058738	0.296486	-1.0	0.0	0.0	0.136364	1.0

```
# t test on whether there's difference on the sentiment of Trump VS Clinton
print(stats.ttest_ind(df2.loc[df2['Keyword']=='Trump','sentiment'], df2.loc[
print(stats.ttest_ind(df2.loc[df2['Keyword']=='Trump','sentiment'], df2.loc[
```

✓ 0.5s

```
Ttest_indResult(statistic=8.986334585573342, pvalue=2.6271988646916943e-19)
Ttest_indResult(statistic=9.185570777439535, pvalue=4.2902677047367086e-20)
```

Preliminary Analysis

- Sentiment analysis
  - Answered Question (2)(a): There is no correlation shown to sentiment and #Followers.
  - Answered Question (2)(b):
    - Clinton seems to have more strong negative sentiment tweets from large influencers than Trump.

```
df2[['Followers','sentiment']].corr()
```

✓ 0.1s

	Followers	sentiment
Followers	1.000000	0.001611
sentiment	0.001611	1.000000

```
df2['sentiment_abs'] = abs(df2['sentiment'])
df2[['Followers','sentiment_abs']].corr()
```

✓ 0.7s

	Followers	sentiment_abs
Followers	1.000000	-0.003919
sentiment_abs	-0.003919	1.000000

Correlation of sentiment and #Followers

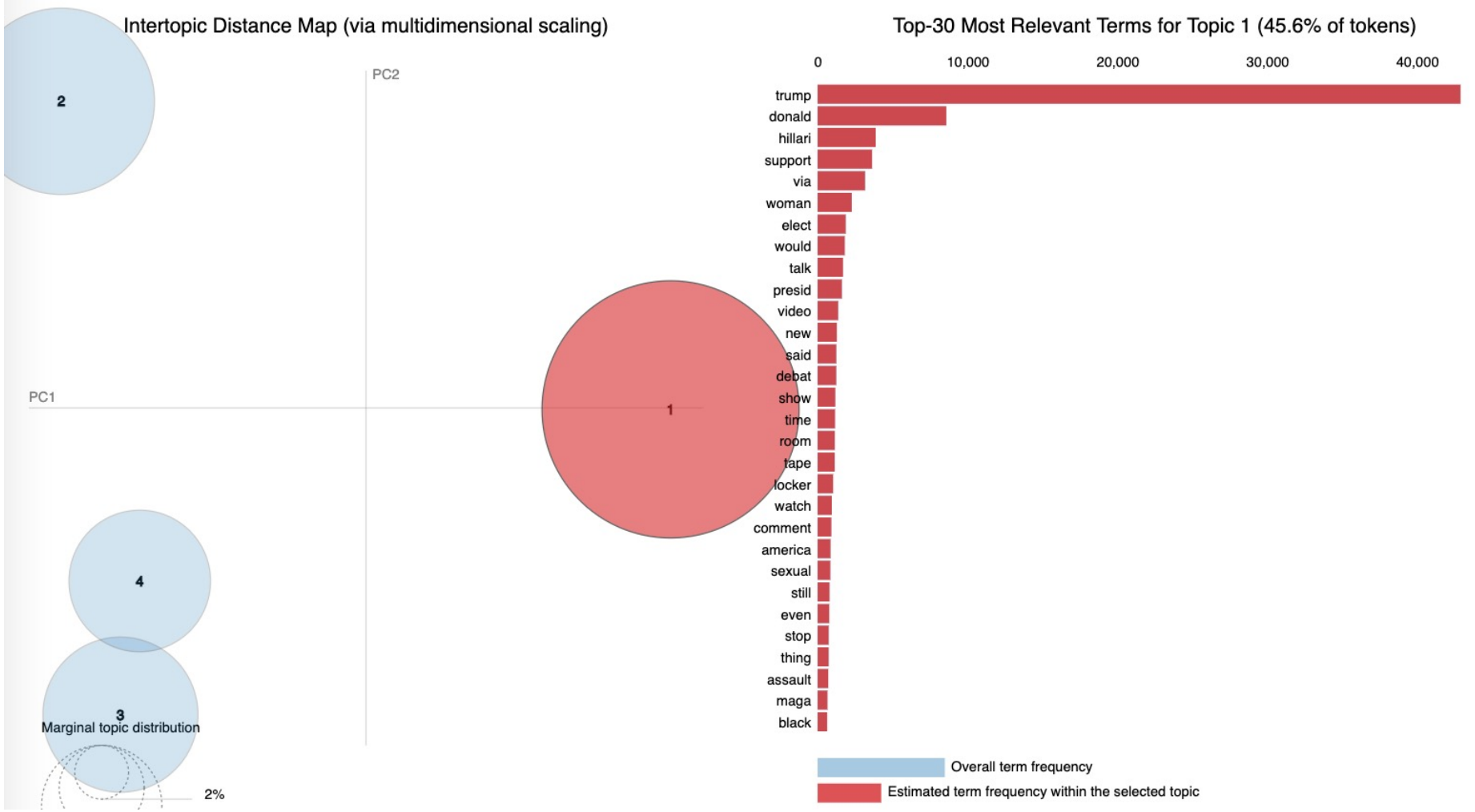
	Keyword	group	count	mean	std	min	25%	50%	75%	max
1	Both	<1k followers	4571.0	0.038824	0.275737	-1.000	0.0	0.0	0.100000	1.0
0	Both	1k~100k followers	2378.0	0.042811	0.276874	-1.000	0.0	0.0	0.098214	1.0
2	Both	>=100k followers	103.0	0.057808	0.277904	-1.000	0.0	0.0	0.023485	1.0
4	Clinton	<1k followers	11843.0	0.037669	0.285110	-1.000	0.0	0.0	0.100000	1.0
3	Clinton	1k~100k followers	7023.0	0.033216	0.269730	-1.000	0.0	0.0	0.050000	1.0
5	Clinton	>=100k followers	133.0	0.010902	0.303418	-0.875	0.0	0.0	0.100000	0.8
7	Election-General	<1k followers	10085.0	0.072006	0.319887	-1.000	0.0	0.0	0.200000	1.0
6	Election-General	1k~100k followers	4586.0	0.076458	0.315850	-1.000	0.0	0.0	0.200000	1.0
8	Election-General	>=100k followers	63.0	0.142063	0.317093	-0.800	0.0	0.0	0.283333	1.0
10	Trump	<1k followers	27531.0	0.060671	0.299818	-1.000	0.0	0.0	0.136364	1.0
9	Trump	1k~100k followers	13938.0	0.054671	0.290317	-1.000	0.0	0.0	0.136364	1.0
11	Trump	>=100k followers	393.0	0.067571	0.276775	-1.000	0.0	0.0	0.125000	1.0

Mean of sentiment value

	Keyword	group	count	mean	std	min	25%	50%	75%	max
1	Both	<1k followers	4571.0	0.153582	0.232263	0.0	0.0	0.000000	0.250000	1.000
0	Both	1k~100k followers	2378.0	0.151556	0.235614	0.0	0.0	0.000000	0.212500	1.000
2	Both	>=100k followers	103.0	0.140009	0.246598	0.0	0.0	0.000000	0.162500	1.000
4	Clinton	<1k followers	11843.0	0.159737	0.239142	0.0	0.0	0.000000	0.250000	1.000
3	Clinton	1k~100k followers	7023.0	0.144614	0.230091	0.0	0.0	0.000000	0.200000	1.000
5	Clinton	>=100k followers	133.0	0.174561	0.247951	0.0	0.0	0.000000	0.325000	0.875
7	Election-General	<1k followers	10085.0	0.200349	0.259555	0.0	0.0	0.066667	0.350000	1.000
6	Election-General	1k~100k followers	4586.0	0.195201	0.259801	0.0	0.0	0.050000	0.336458	1.000
8	Election-General	>=100k followers	63.0	0.200718	0.283051	0.0	0.0	0.033333	0.318750	1.000
10	Trump	<1k followers	27531.0	0.176779	0.249640	0.0	0.0	0.000000	0.300000	1.000
9	Trump	1k~100k followers	13938.0	0.167683	0.243215	0.0	0.0	0.000000	0.285714	1.000
11	Trump	>=100k followers	393.0	0.151669	0.241080	0.0	0.0	0.000000	0.250000	1.000

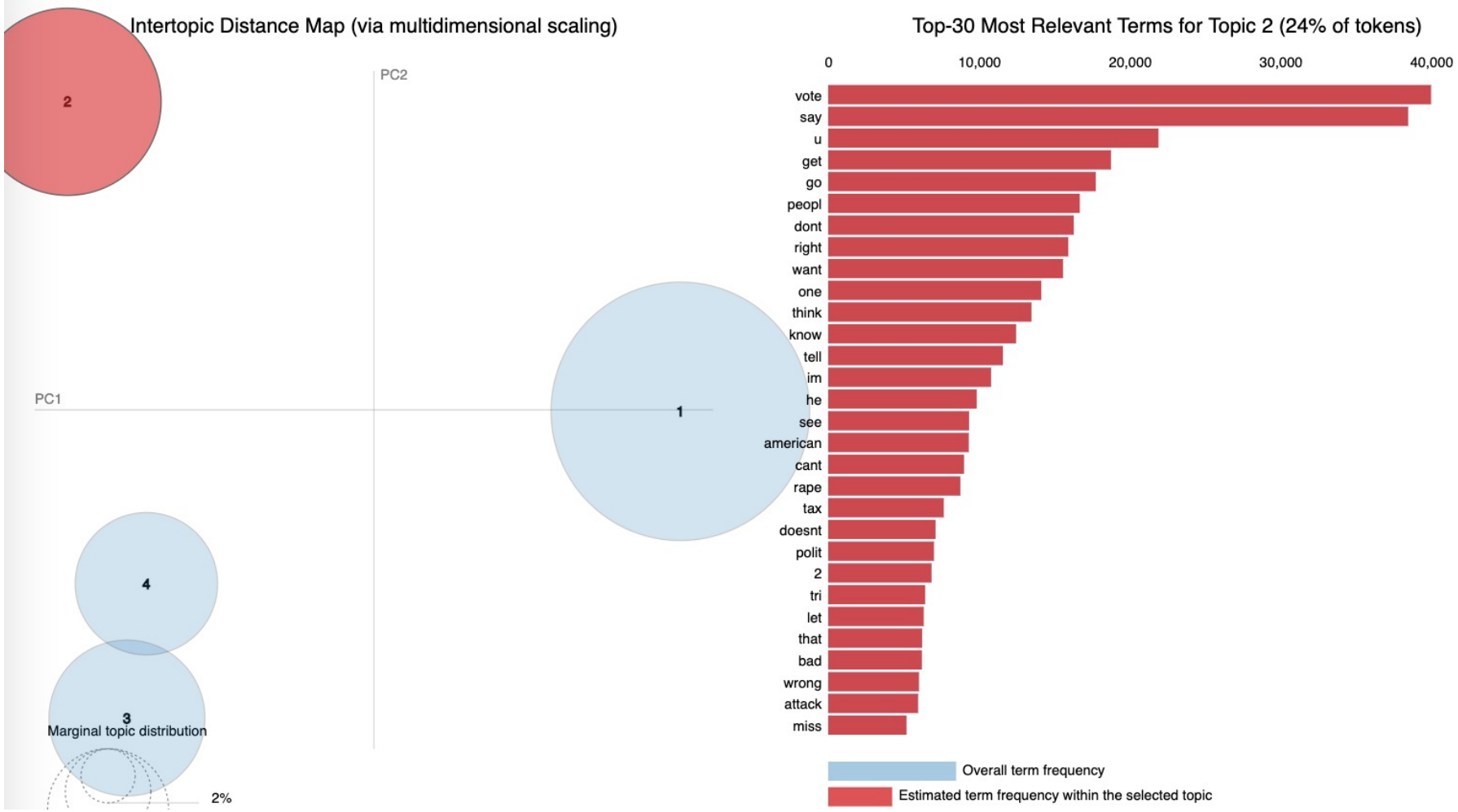
Mean of absolute sentiment value

Preliminary Analysis



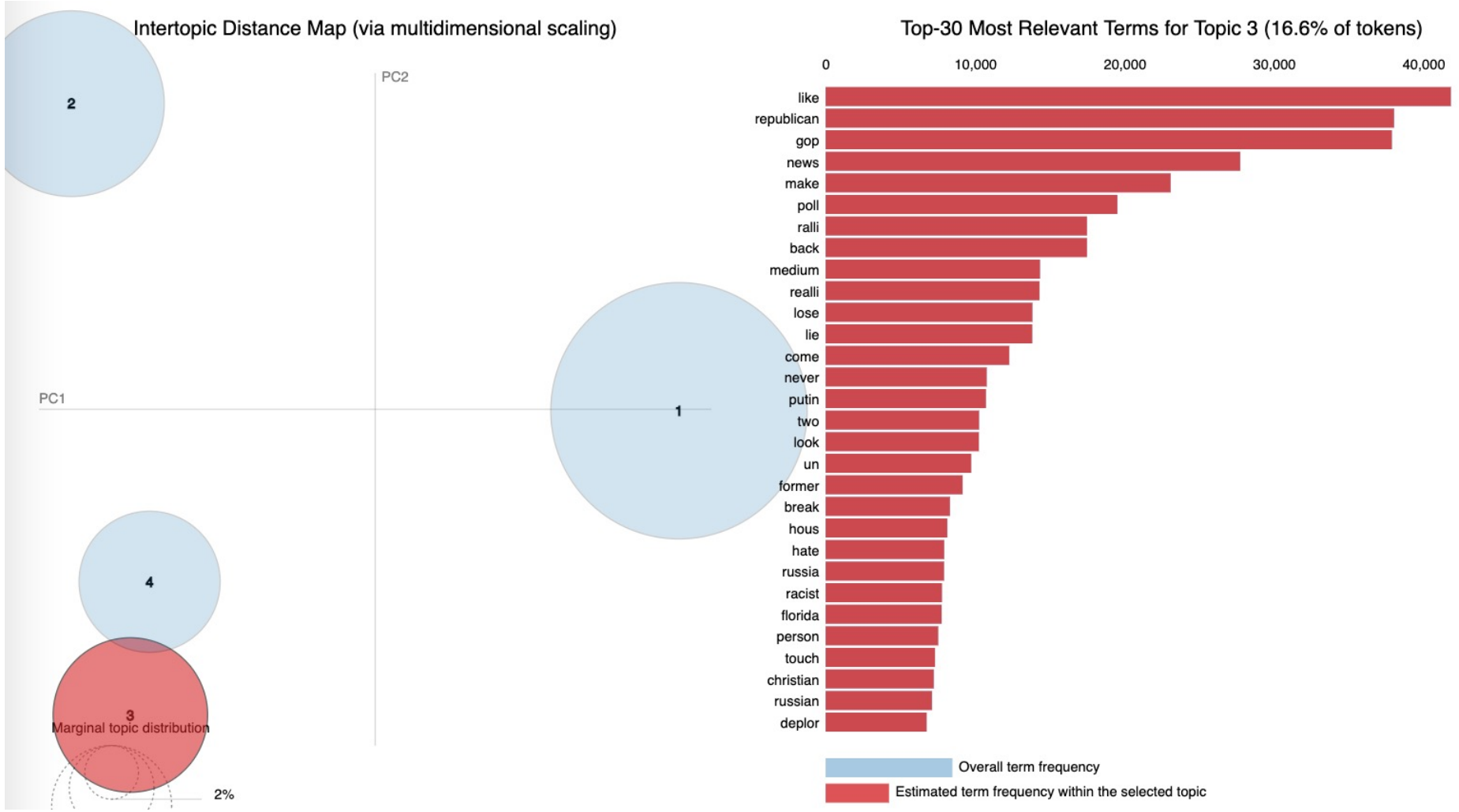
- Topic model (Trump)

Preliminary Analysis



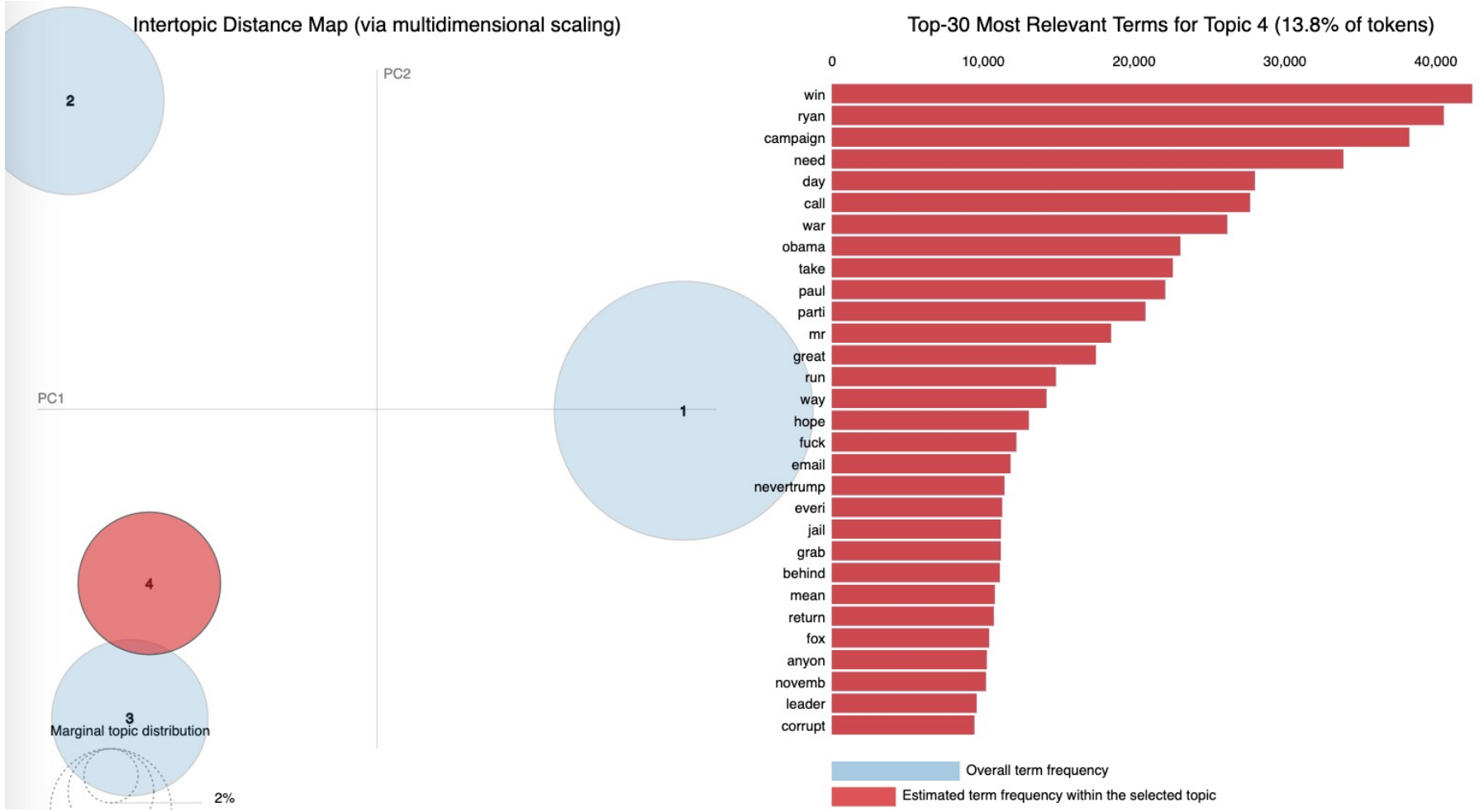
- Topic model (Trump)

Preliminary Analysis



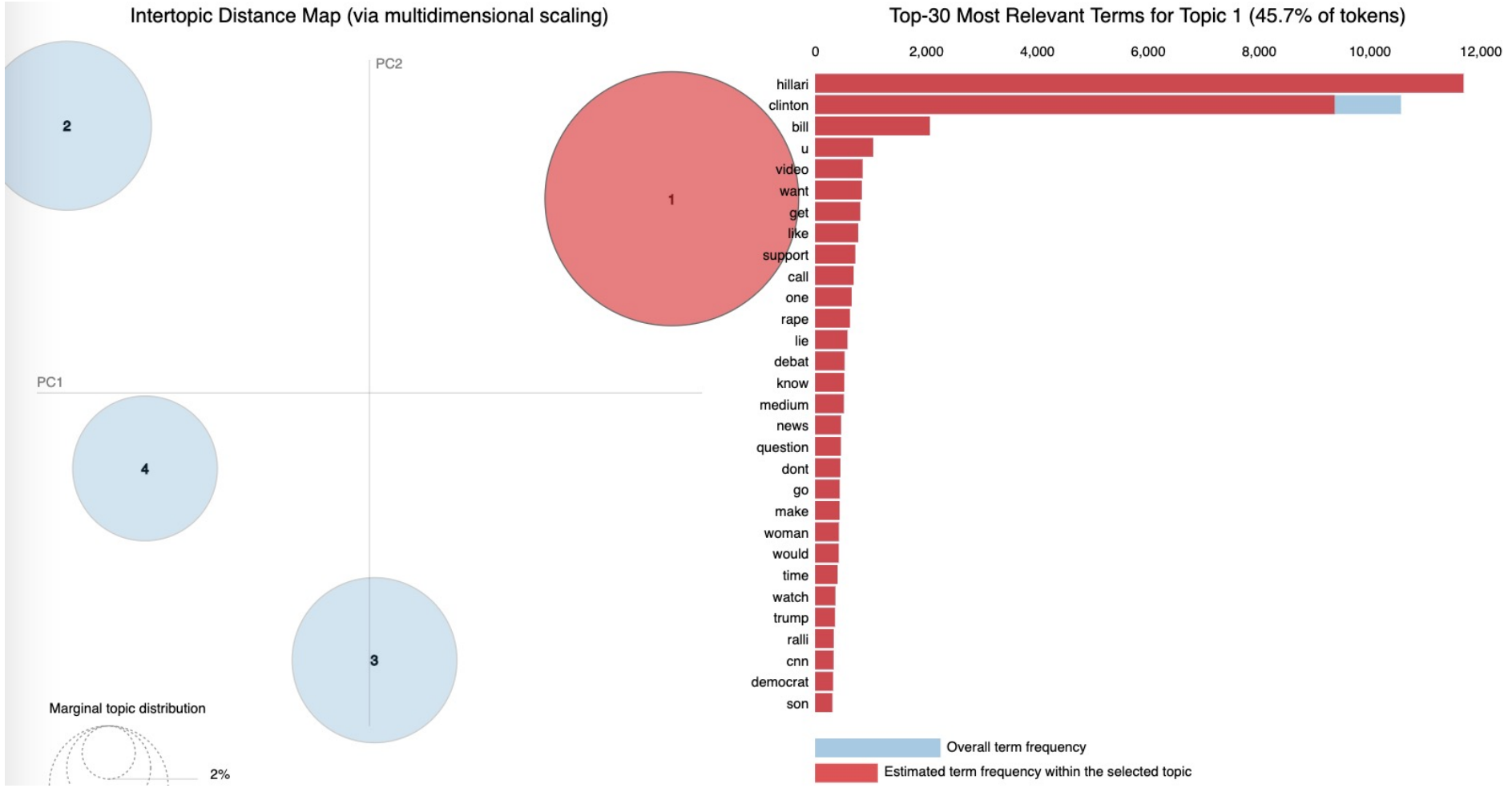
- Topic model (Trump)

Preliminary Analysis



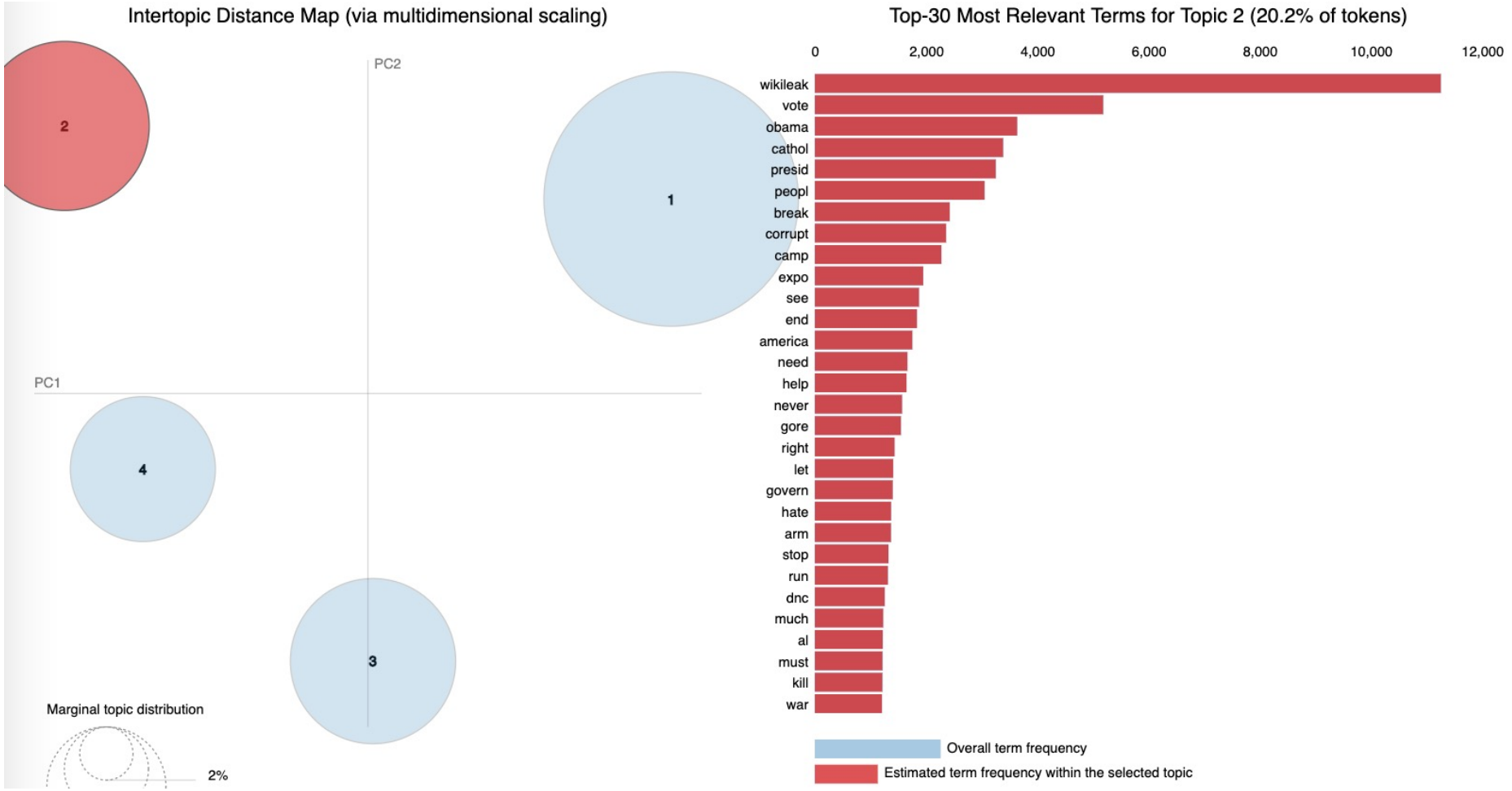
- Topic model (Trump)

Preliminary Analysis



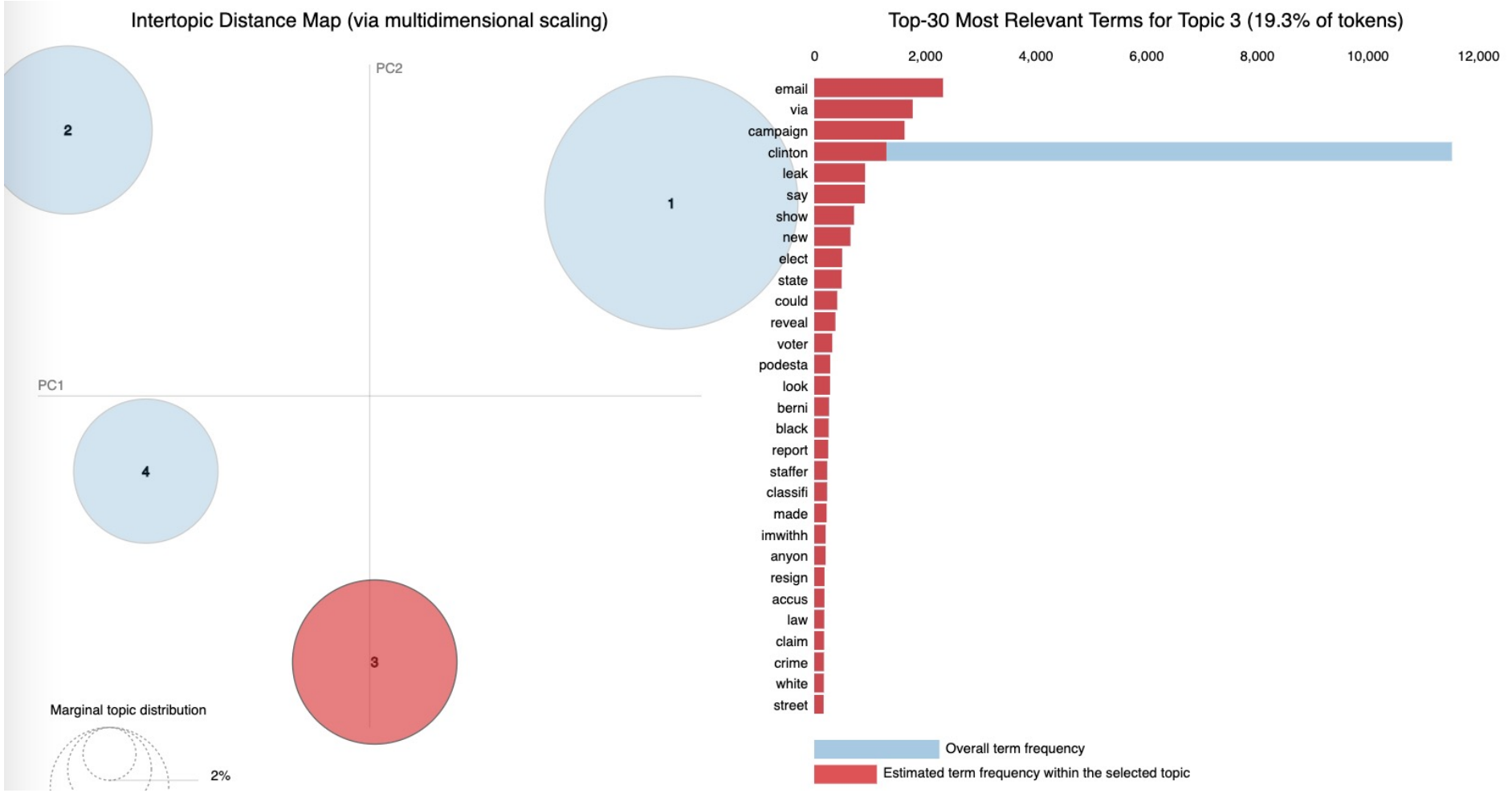
- Topic model (Clinton)

Preliminary Analysis



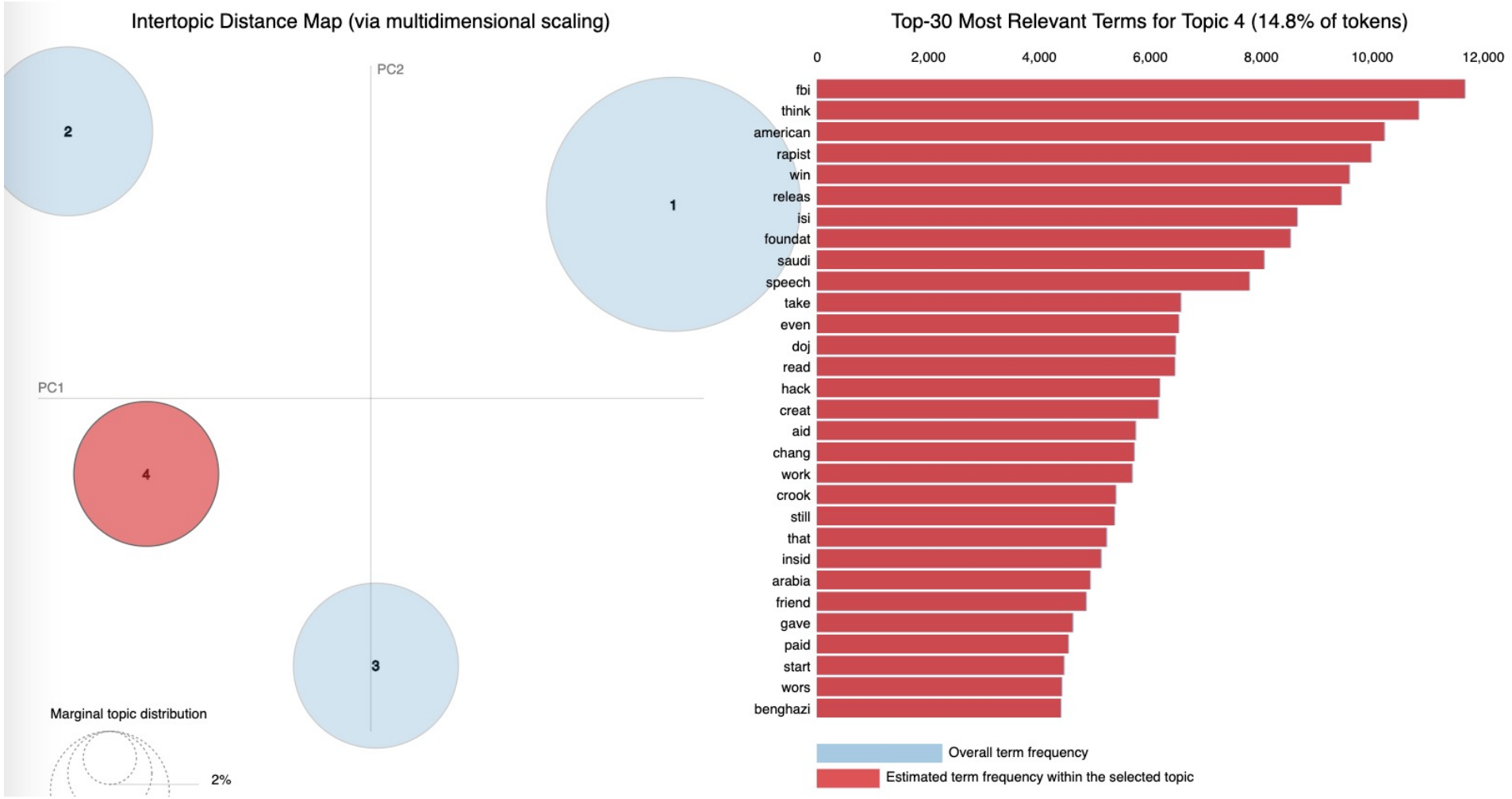
- Topic model (Clinton)

Preliminary Analysis



- Topic model (Clinton)

Preliminary Analysis



- Topic model (Clinton)

Preliminary Analysis

- Topic models
  - Answered Question (3):
    - From the most frequent words (especially the top 5 from each topic). For trump the majority are the supporting words, however for Clinton, 'wikileak', 'email', 'leak' appears at the very top of the topics. Also there're some negative words resulted from Bill Clinton shows up at the top 5 most frequent words of the topic.
    - This is consistent with our previous sentiment analysis that Trump have higher positive sentiment than Clinton overall.

Importance

- To political field specifically, this study is important, as it could reveal the potential problem of political polarization and quantify how serious the problem is and which way it is going.
- To general fields (including political), it can find out the topics from the text, so that we can get better understanding on what people are paying attention to.
- In terms of business use case, tech company (such as Twitter) could use real time tweets data to build the hot topics. It could either be used on ensuring integrity or improving users engagement (thus creating value for the company).

Thank you!
Any questions?

- Twitter data credit: Prof. Kayla Jordan