Macron vs Le Pen Through Tweets

Spark EDA on Tweets
Group 2

Data & Pipeline

- Parsed json file with 216,912 lines, each line represented a tweet
  - Had try/except in case of non-utf8 chars
- Read data into Spark RDD with json class
- Filtered RDD to create Spark DataFrame with only our desired features
- Converted Spark DataFrame into Pandas DataFrame for plotting, analysis

Pipeline

RDD Pipeline to see what attributes there were

```
french_tweets_rdd = sc.textFile('data/french_tweets.json')
```

```
french_tweets_rdd.map(ss.parse_json).take(2)
```

Dataframe pipeline for further processing

```
french_tweets_df = (french_tweets.select('created_at', 'text', 'geo', 'retweet_count', 'favorite_count', 'entities')
                    .withColumn('hashtags', text_only_udf(french_tweets['entities']['hashtags']))
                    .withColumn('user_mentions', username_only_udf(french_tweets['entities']['user_mentions']))
                    .drop('entities')
                    .toPandas())
```

```
french_tweets_df['created_at'] = french_tweets_df['created_at'].astype('datetime64')
```

EDA

Features We Included

- Selected 7 features from 609 schema elements in json data
- “Created_at” - Date of tweet
- “Text” - Tweet content
- “Geo” - Tweet location in latitude/longitude
- “Retweet_count” - # users who retweeted tweet, *but all 0 or nan in json data*
- “Favorite_count” - # of users who favorited tweet, *but all 0 or nan in json data*
- “Entities”:
  - “Hashtags”
  - “User_mentions”

```
french_tweets_df.sample(3)
```

	created_at	text	geo	retweet_count	favorite_count	hashtags	user_mentions
140682	2017-04-28 09:32:29	La meilleure façon de suivre et d'organiser sa...	None	0.0	0.0	[memo, emploi, i4Emploi]	[]
204003	2017-04-28 21:44:45	@CindiMakowski 😂😂😂 \nlol actually I'm not arrog...	None	0.0	0.0	[]	[CindiMakowski]
185758	2017-04-28 18:51:25	Artfood ! 🤗 👌 Ça faisait un petit moment que j...	([48.85311872, 2.3691685], Point)	0.0	0.0	[]	[]

Overview of DataFrame

- Final dataframe had 214,936 rows and 7 columns
- 10 rows had N/A for “created_at” & “text” = 10 possibly invalid tweets/data
  - Other 214,926 tweets valid
- 18,234/214,926 (8.5%) of “geo” data from valid tweets was non-null

Types of Each Column

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 214936 entries, 0 to 214935
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   created_at      214926 non-null  datetime64[ns]
 1   text            214926 non-null  object
 2   geo             18234 non-null   object
 3   retweet_count   214926 non-null  float64
 4   favorite_count  214926 non-null  float64
 5   hashtags        214936 non-null  object
 6   user_mentions   214936 non-null  object
dtypes: datetime64[ns](1), float64(2), object(4)
memory usage: 11.5+ MB
```

Number of Null Values per Column

```
created_at          10
text                10
geo             196702
retweet_count       10
favorite_count      10
hashtags             0
user_mentions        0
dtype: int64
```

Summaries of Features

	created_at	text
count	214926	214926
unique	115448	213175
top	2017-04-28 08:17:24	#TPMP1000
freq	28	41
first	2017-04-26 13:30:45	NaN
last	2017-04-29 05:01:54	NaN

- Tweets were from April 26-29 in 2017 (top right table)
- All tweets from data showed 0 for retweet & favorite count (bottom right table)

Summary Statistics for # Hashtags

```
count    214936.000000
mean          0.444509
std           1.134995
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max          14.000000
Name: hashtag_len, dtype: float64
```

Summary Statistics for # User Mentions

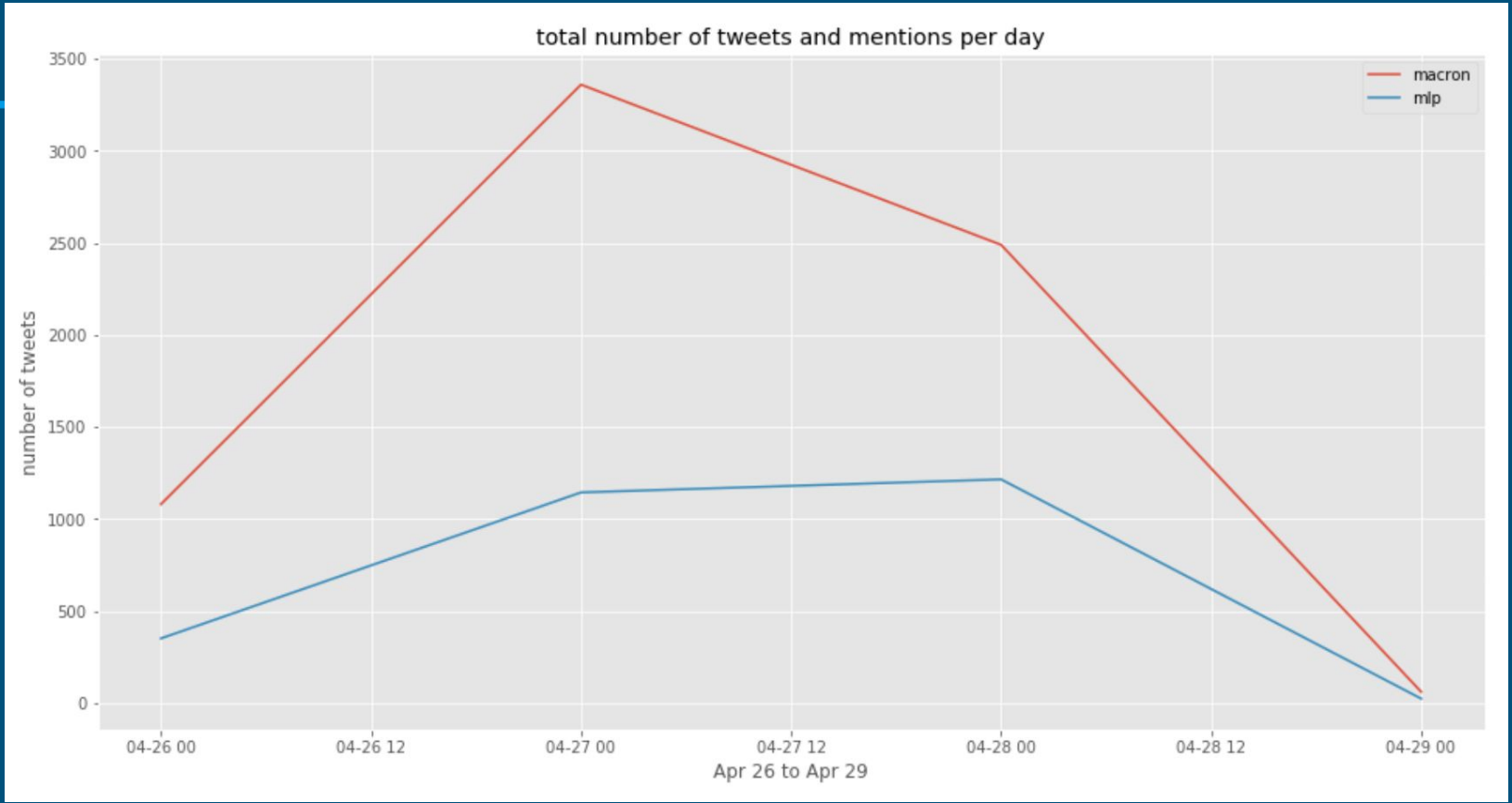
```
count    214936.000000
mean          0.677913
std           1.247606
min           0.000000
25%           0.000000
50%           0.000000
75%           1.000000
max          12.000000
Name: user_mentions_len, dtype: float64
```

Breakdown of Numerical Values

	retweet_count	favorite_count
count	214926.0	214926.0
mean	0.0	0.0
std	0.0	0.0
min	0.0	0.0
25%	0.0	0.0
50%	0.0	0.0
75%	0.0	0.0
max	0.0	0.0

Analysis

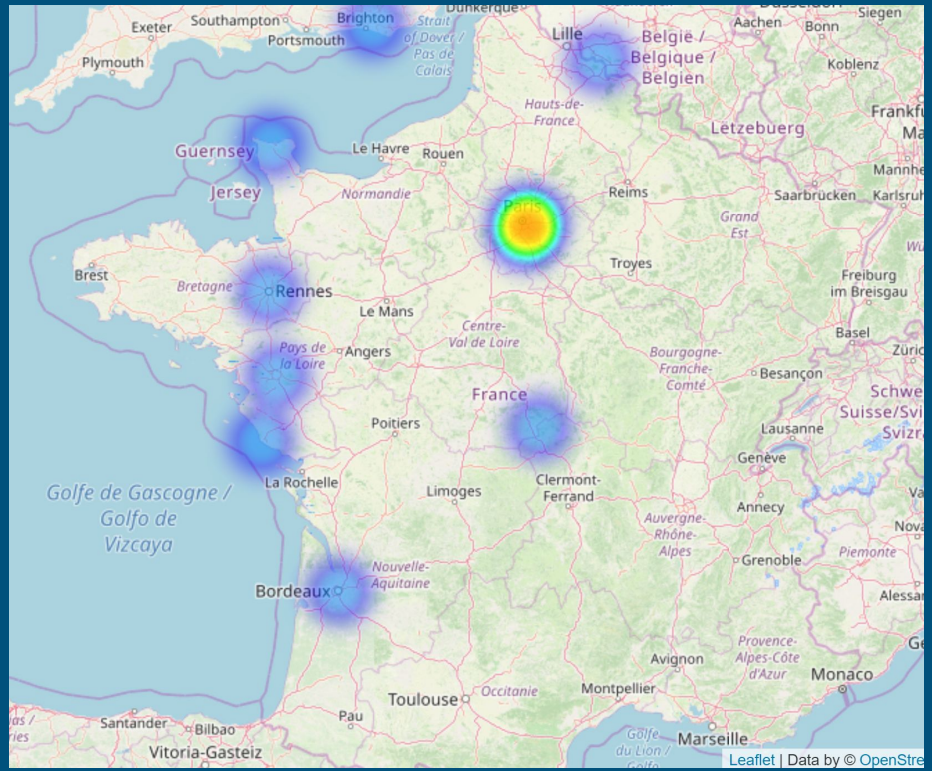
Tweets and Mentions Per Day



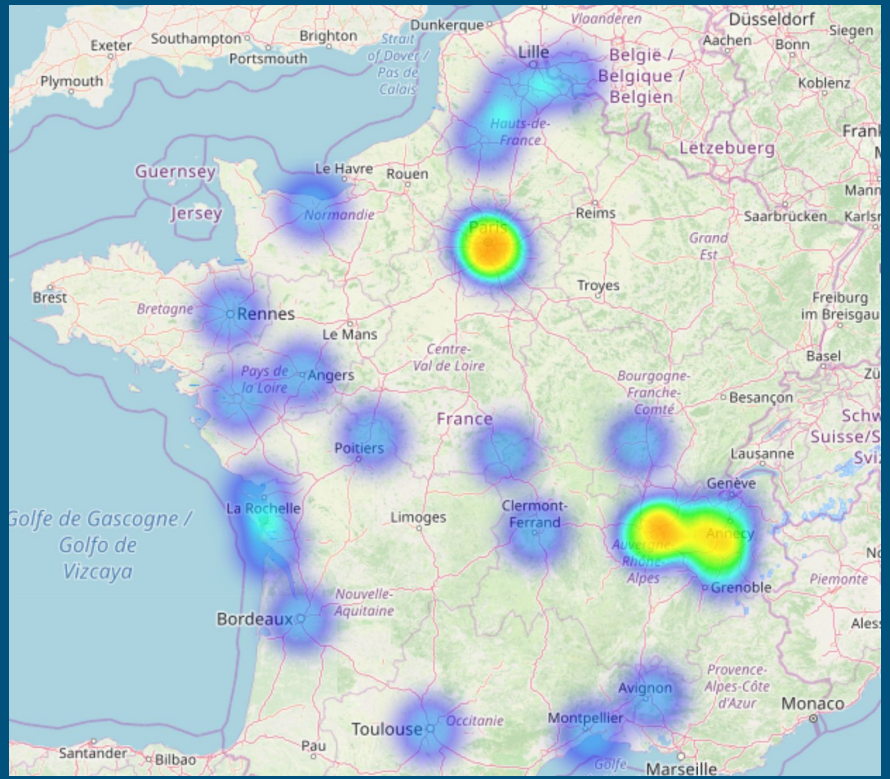
Cumulative Tweets and Mentions Per Day

Geolocation of Twitter Mentions

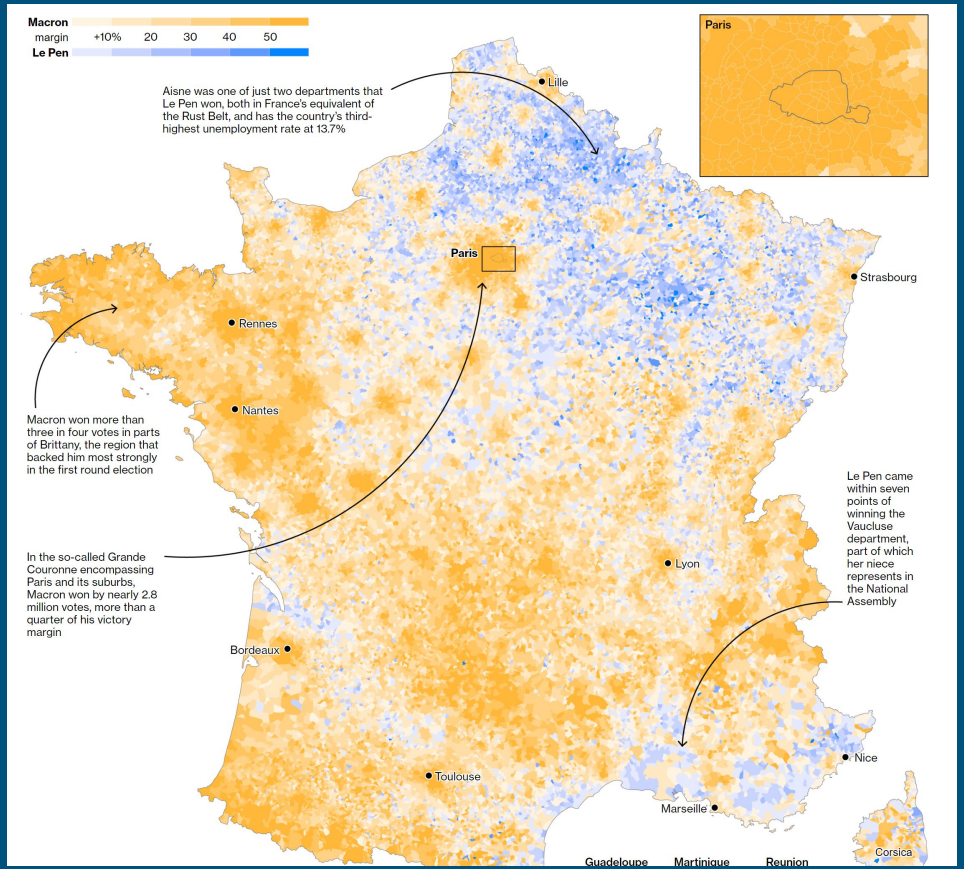
Le Pen



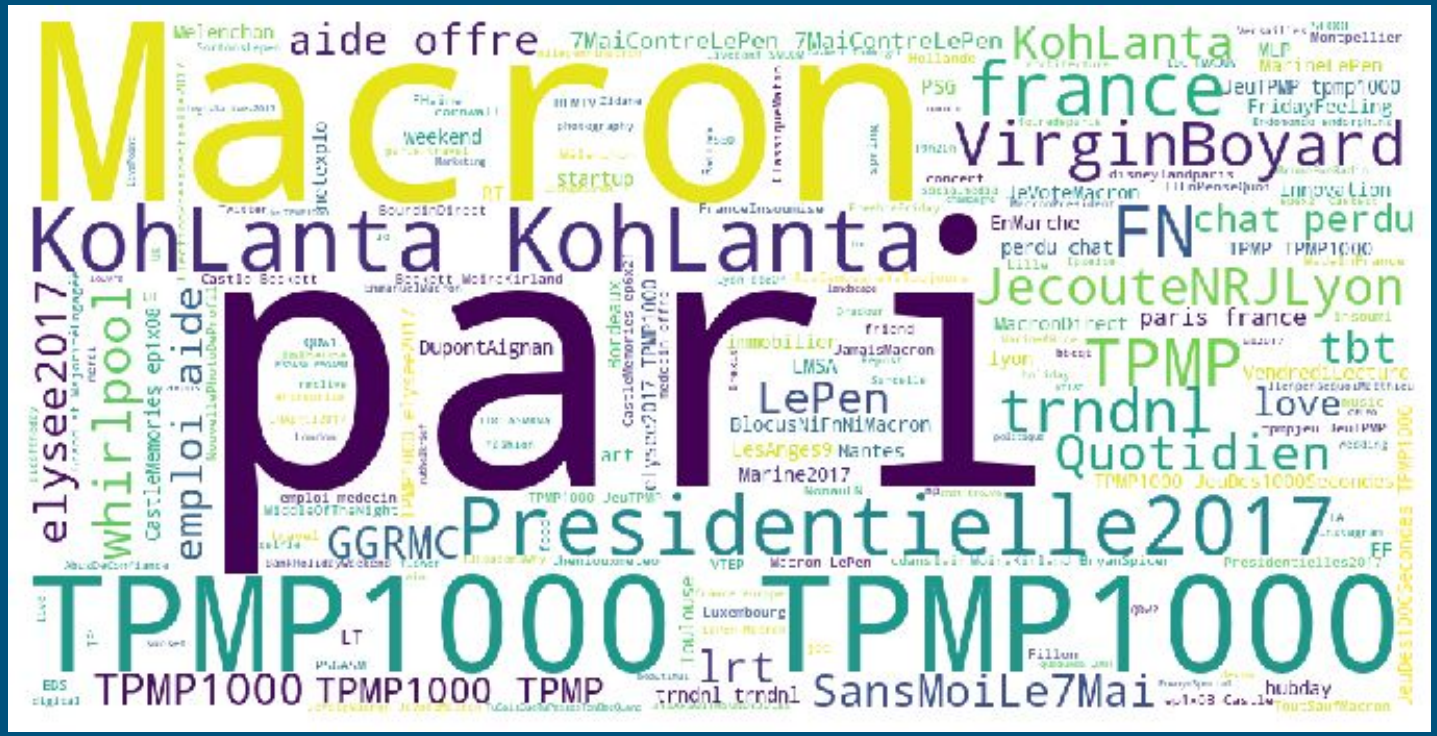
Macron



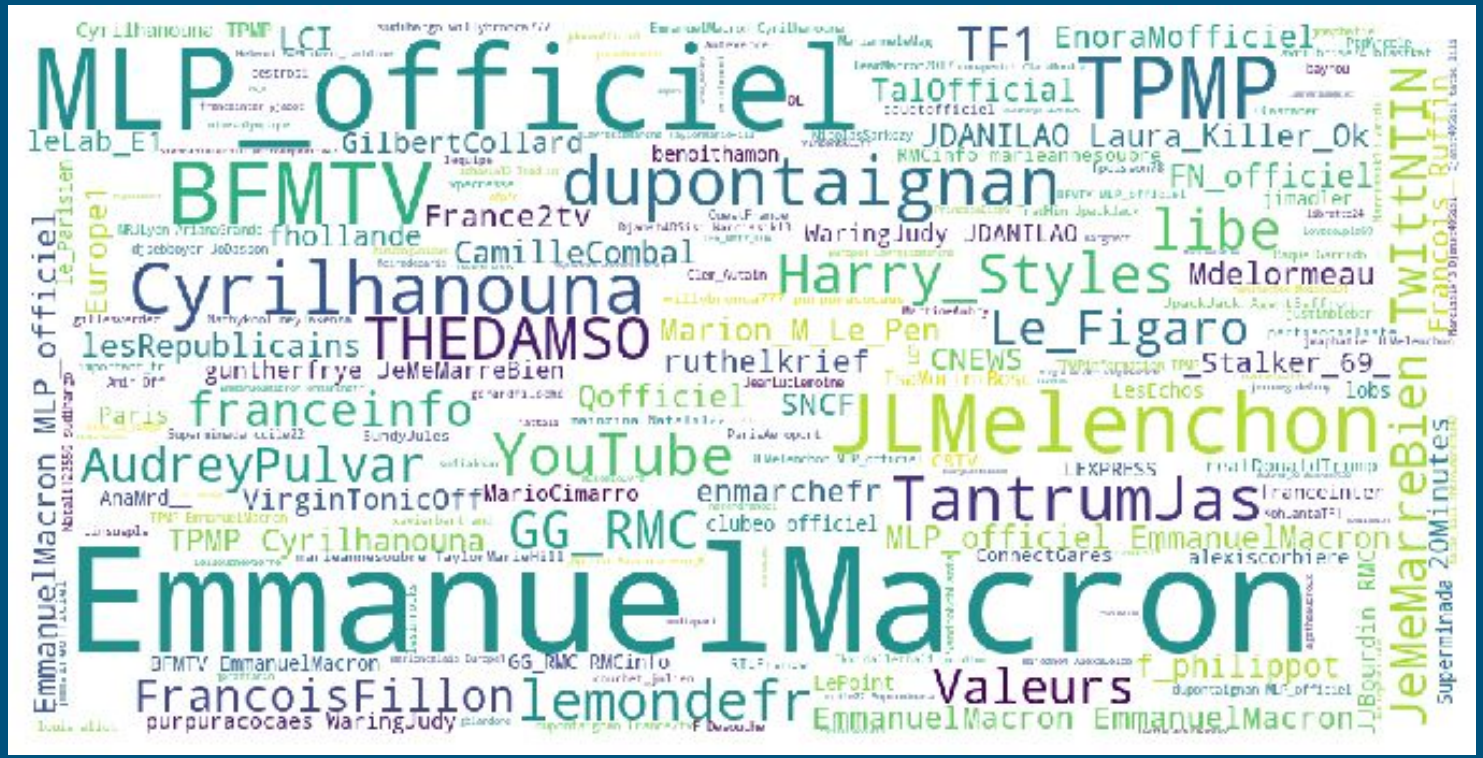
Actual Election Results



#Hashtag Word Cloud



Mentions Cloud



Conclusion