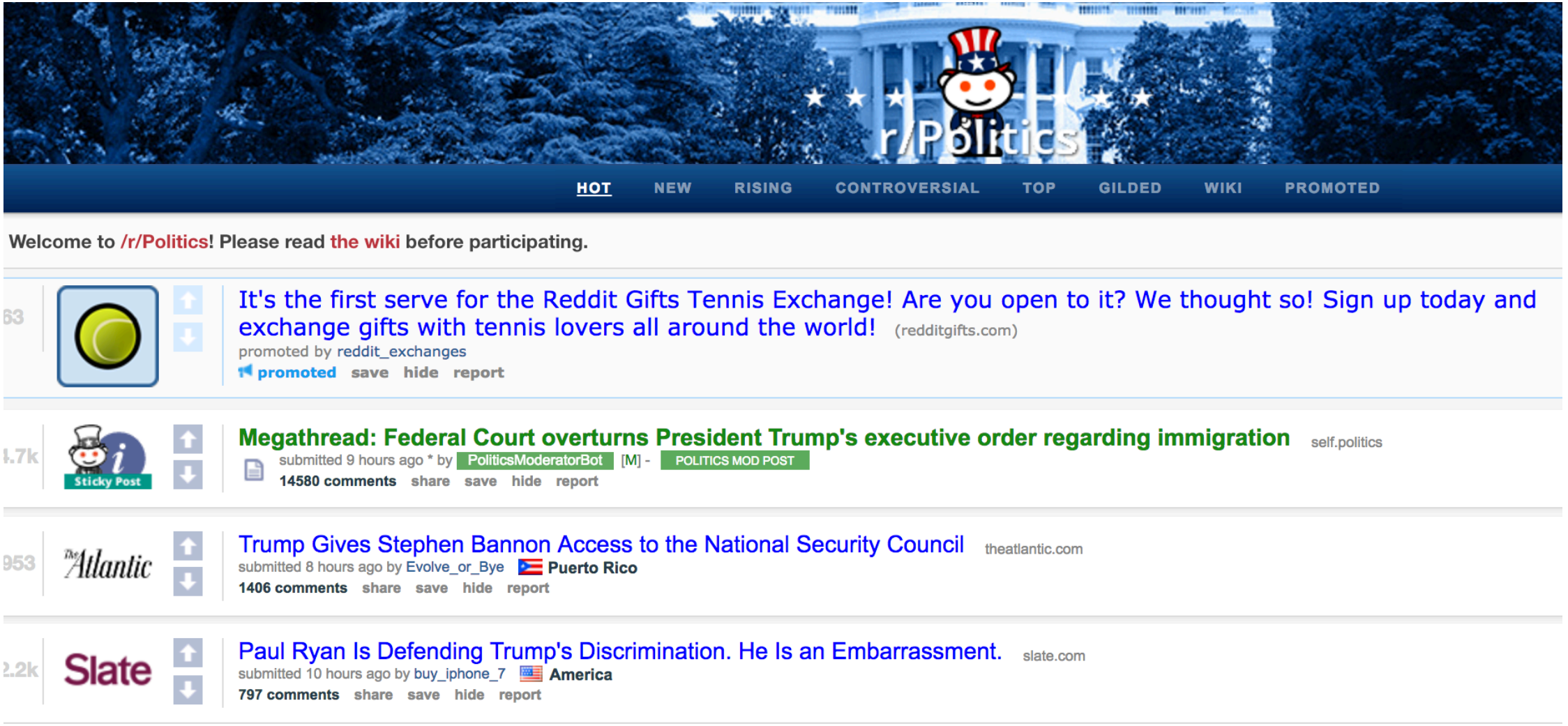
The Best Words

Data mining Reddit to
capture political moods

reddit

Reddit as the internet's id

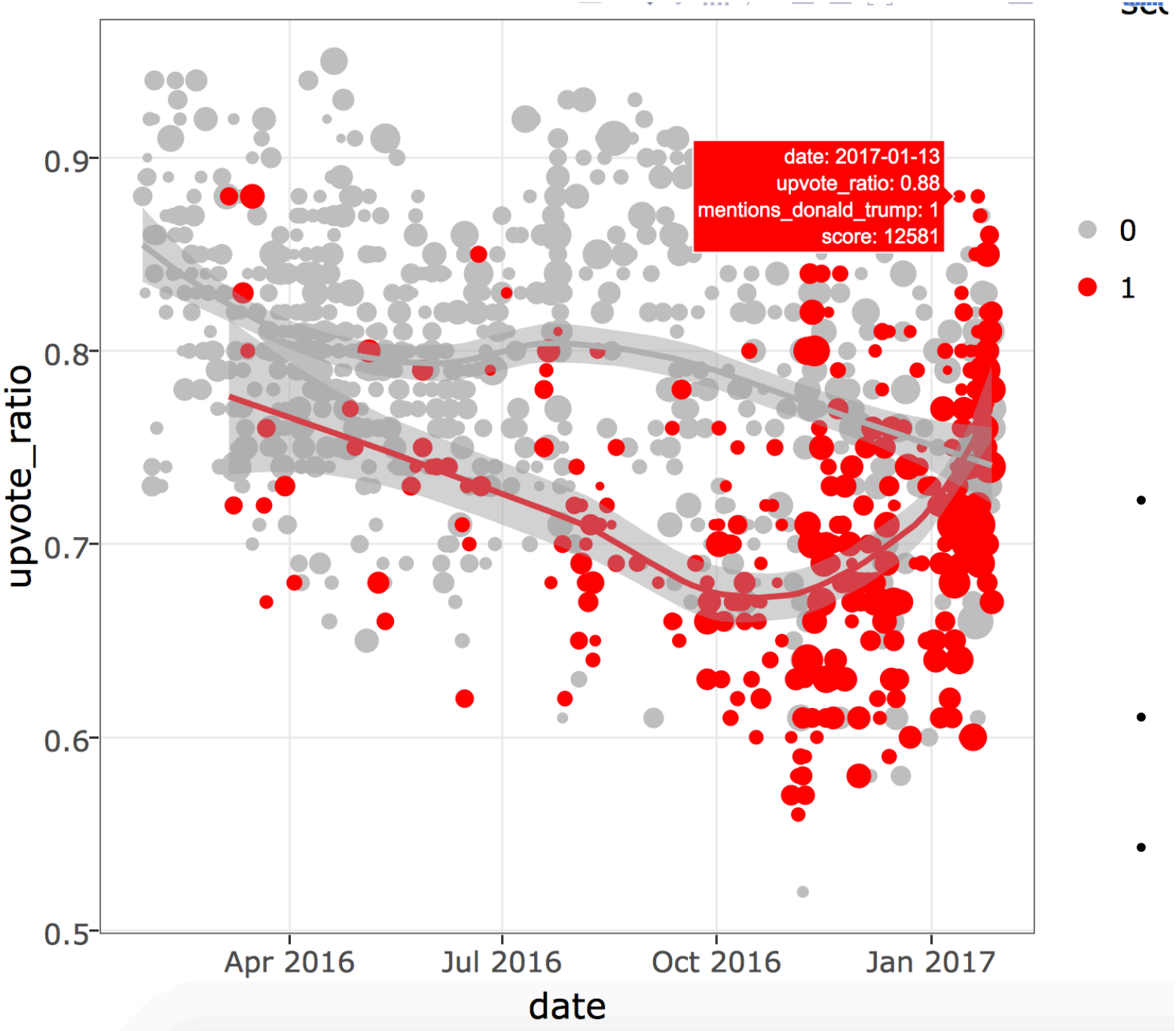
Reddit as a political looking glass



This project:

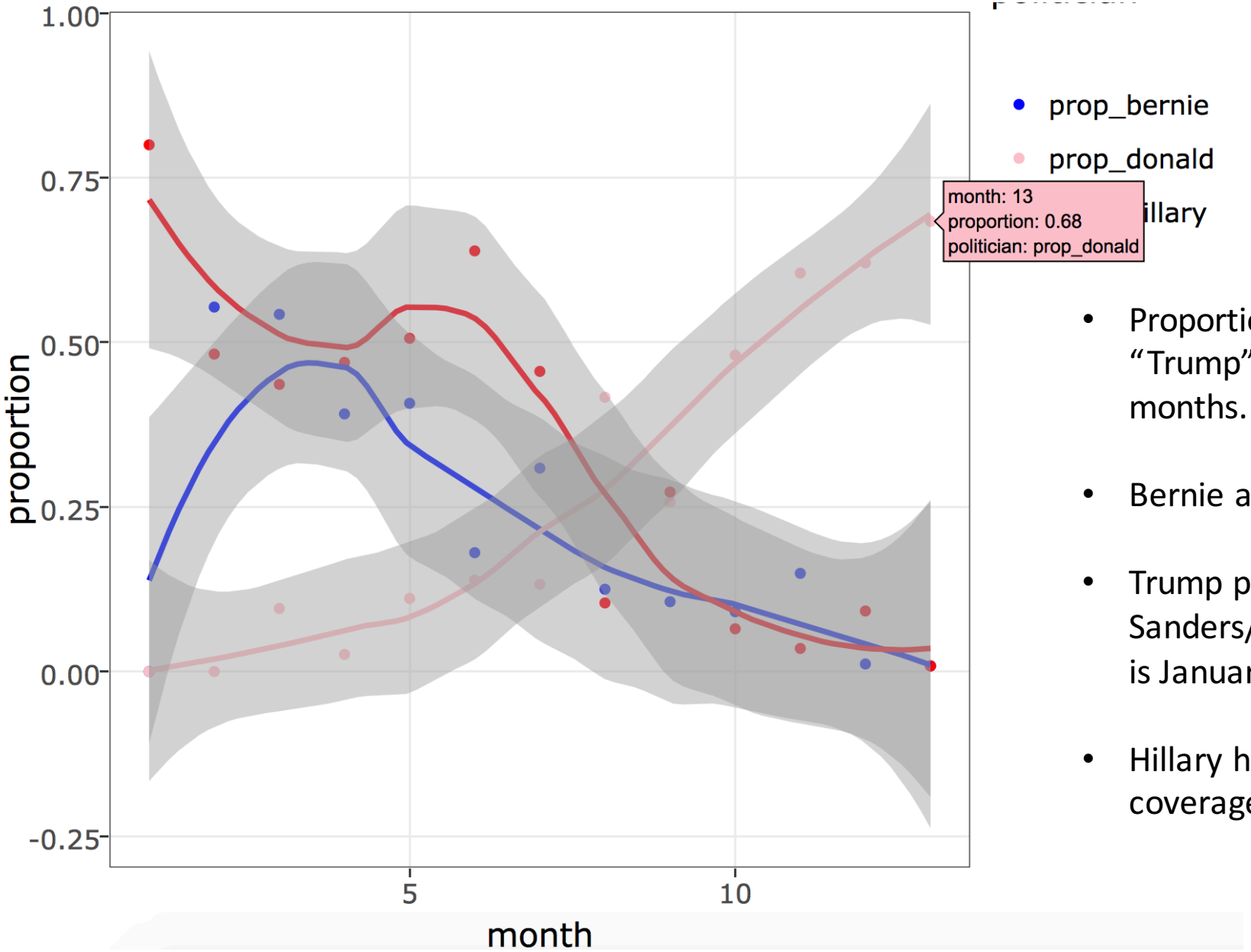
- Analyse the 1000 most “top” and “controversial” threads and comment trees on /r/ politics for previous 12 months

- Methods
  - **Python package PRAW**, a wrapper for Reddit’s API + Jupyter notebook
  - **Data cleaning, analysis and sentiment mapping** using R
    - tidyr, dpyr, stringr, purrr, tidytext
  - Microsoft Cognitive Services Azure **Text Analytics API**
  - **Data visualisation**
    - ggplot2, plotly



Upvote For Trump 2016

- The top 1000 posts in /r/politics in last year, have gradually gotten more controversial (lower upvote ratio) over the past 12 months
- Tons more post titles mention Trump (red dots) after the election (too late!)
- Trump (red line) started to be viewed much more positively starting in October! Bad omen.



The eclipse

- Proportion of top submissions containing "Trump" or "Donald" over the last 12 months. It's now (month=13) up to 69%!
- Bernie and Hillary down to ~ 0%.
- Trump publicity really overtook Sanders/Clinton in late march (month of 0 is January 2016)
- Hillary has always enjoyed higher coverage than Bernie.

Trends in hate, disgust, fear, etc. towards Trump, Clinton, Sanders

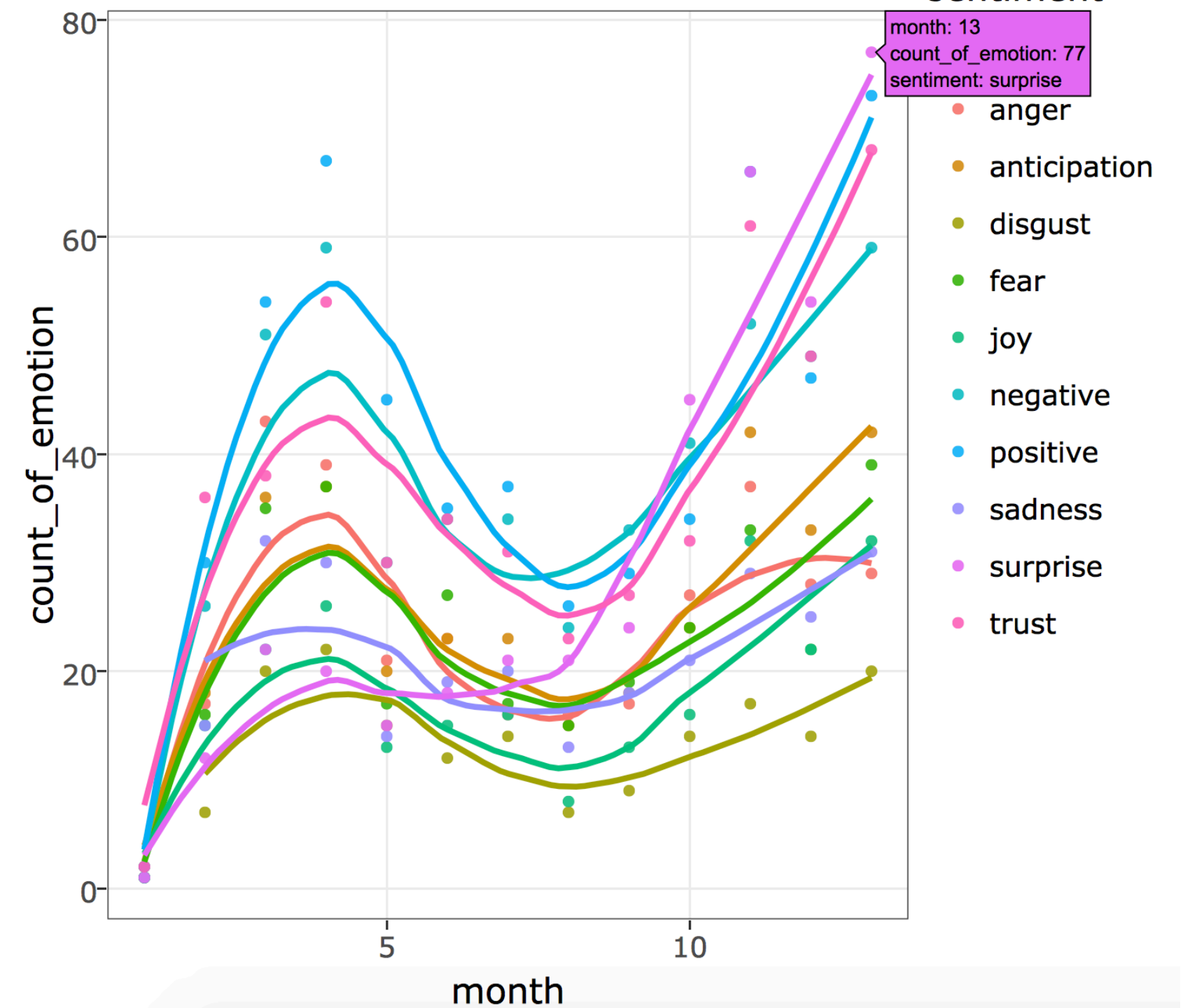
	word	sentiment
	<chr>	<chr>
1	abacus	trust
2	abandon	fear
3	abandon	negative
4	abandon	sadness
5	abandoned	anger
6	abandoned	fear

⋮

N=13901

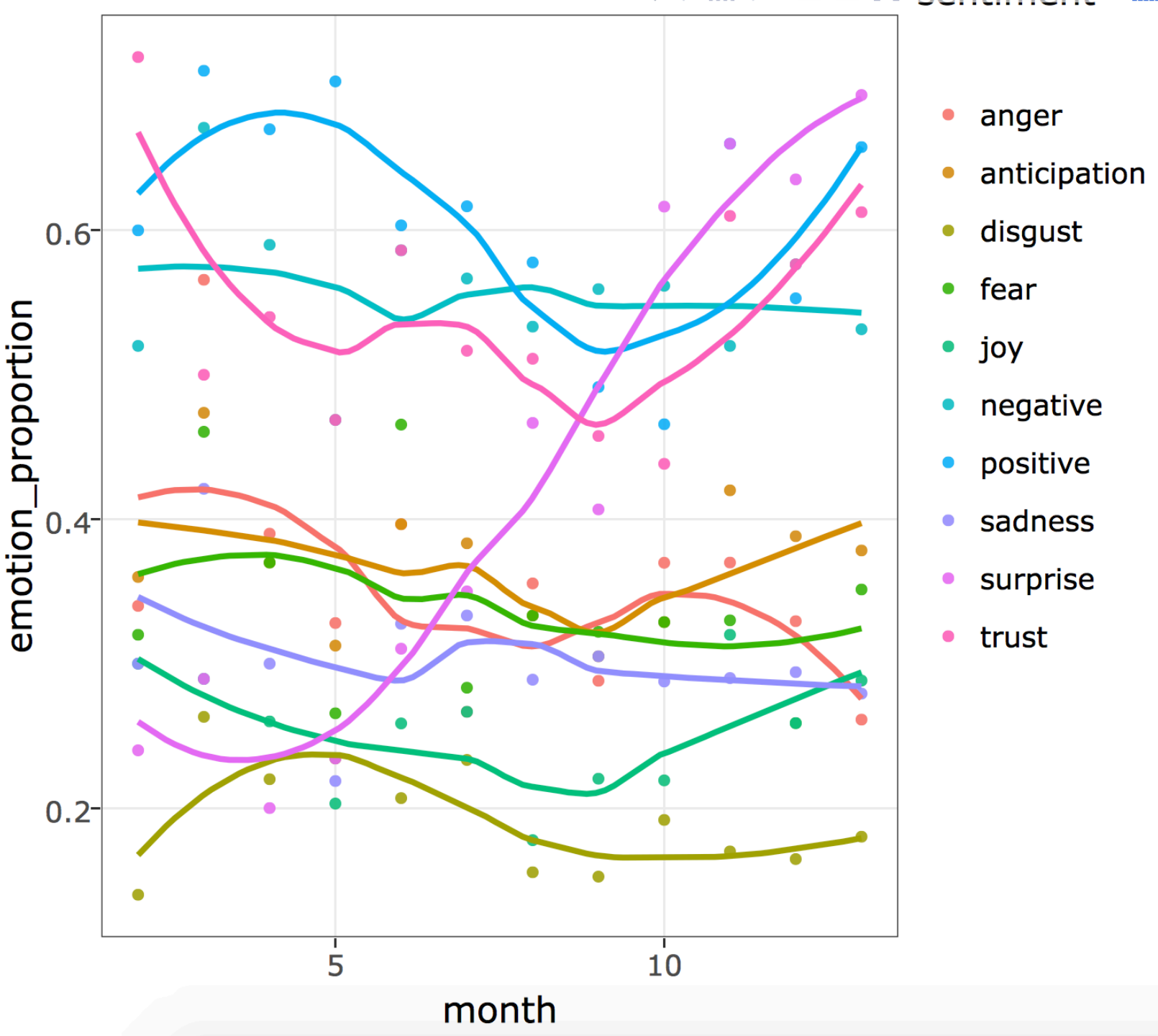
Titles of submissions

Text of the top comment per submission



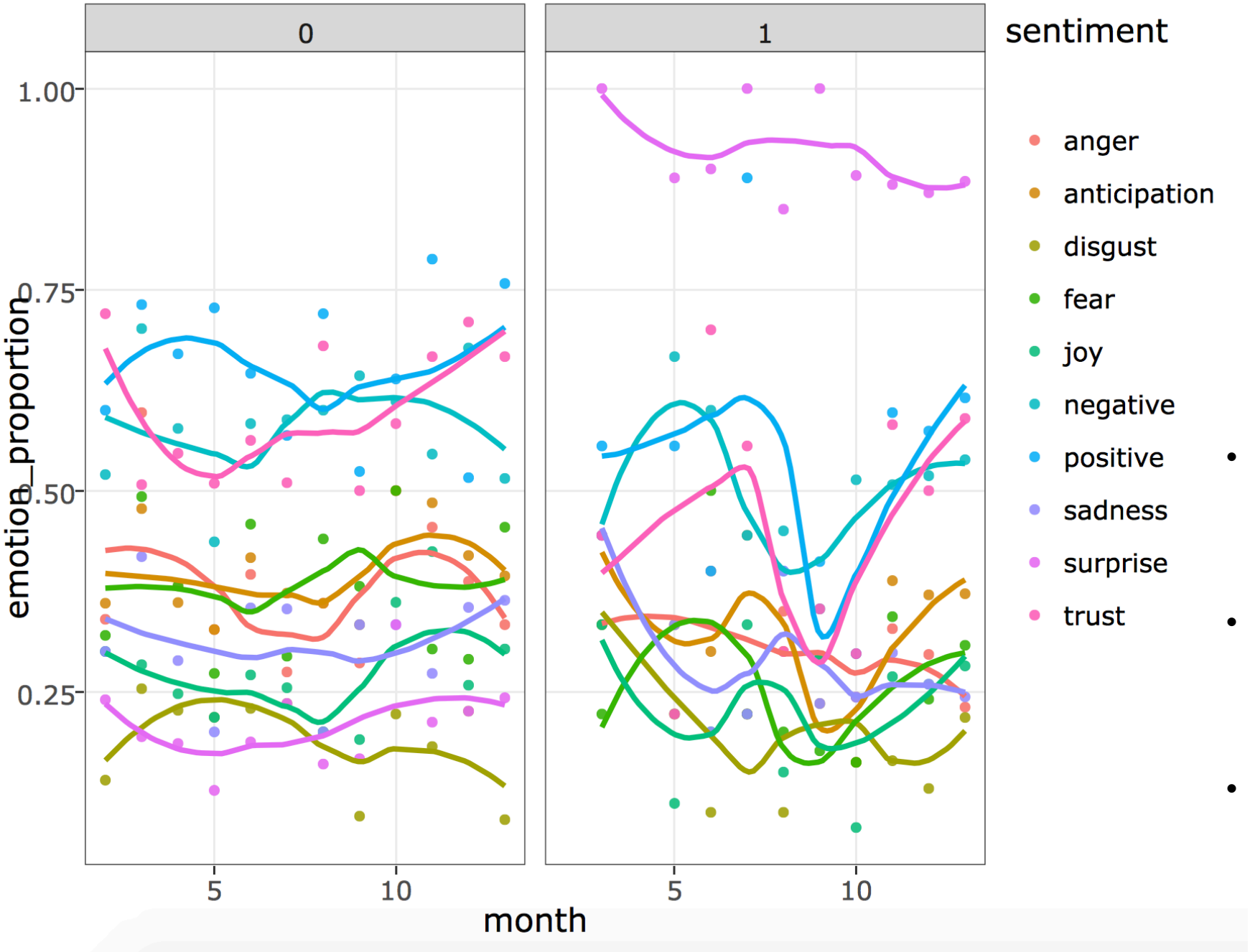
Emotions about politics in general

- Big correlations because raw counts, need to scale..



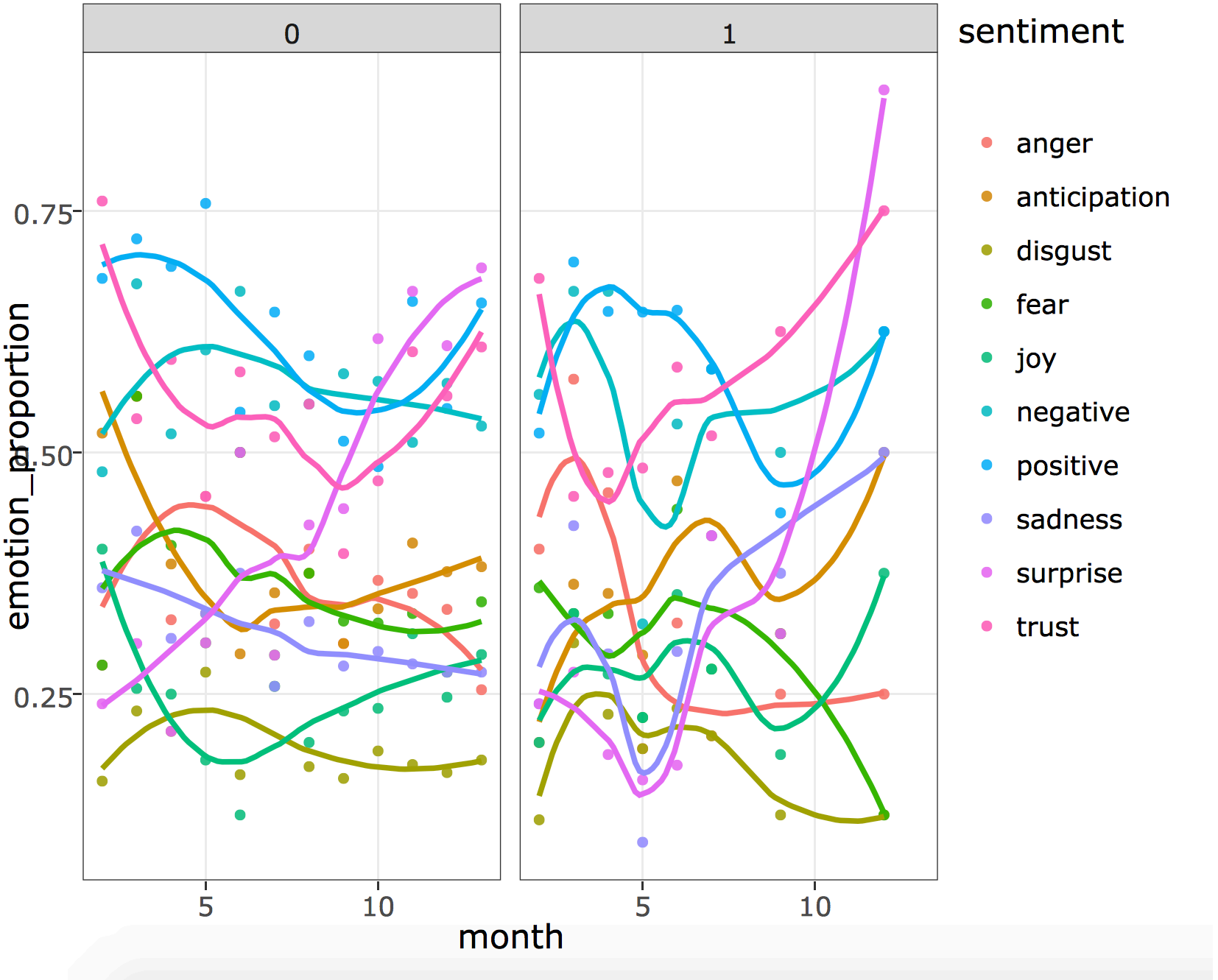
"Surprise" is going up and up and up!

- Trust hit the lowest point at the month before the election
- Post-election, we are seeing a lot of fear (green)!
- Weirdly, post-election, trust and positivity are also going up?
- The year started off pretty positive and then nosedived, recovering now



Emotional swings are due to The Donald

- Left panel (0) is posts not mentioning Donald, right panel mentions Donald
- The level of surprise is much much higher in posts related to Trump
- The huge dips in trust, joy, happiness pre-election we saw before are largely due to Trump!



Compared to Hillary..

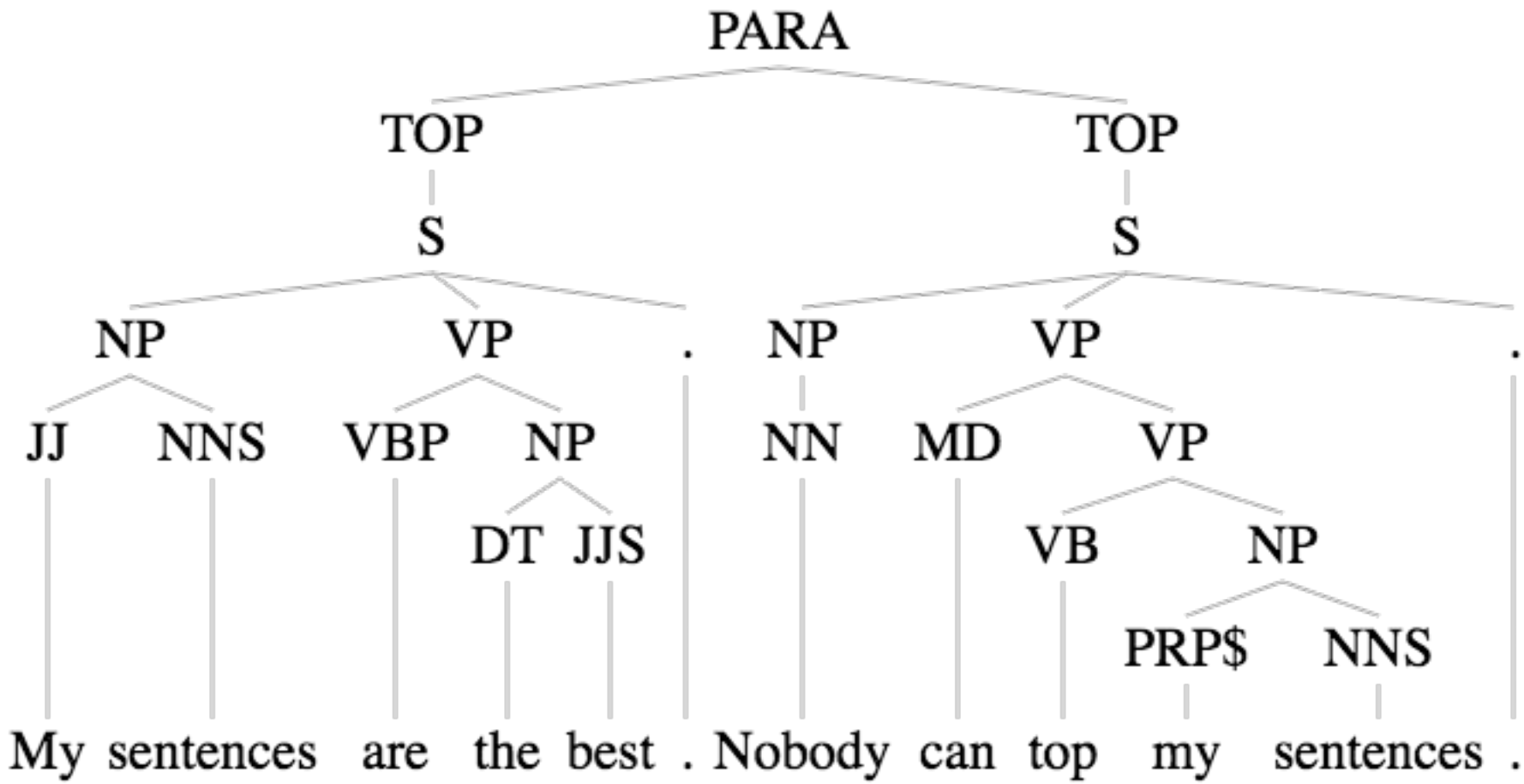
- Posts about Hillary Clinton (facet with '1') don't have such strong trends, mostly just follow the background levels of emotions about politics
- Main trend we do see here – anger/sadness/fear peaked right around the time that Donald was gaining lots of popularity (see plots 2)... related to the scandals, probably.

Topic and sentiment detection with Azure

In [8]:
```
# Azure portal URL.
base_url = 'https://westus.api.cognitive.microsoft.com/'

# Your account key goes here.
account_key = 'ac455400c61d4c5792c1b4bc57b55715'

headers = {'Content-Type':'application/json', 'Ocp-Apim-Subscription-Key':account_key}

input_texts = '{"documents":['+df[['top_comment_body','title']].to_json()+']}'

num_detect_langs = 1;

# Detect key phrases.
batch_keyphrase_url = base_url + 'text/analytics/v2.0/keyPhrases'
req = urllib2.Request(batch_keyphrase_url, input_texts, headers)
response = urllib2.urlopen(req)
result = response.read()
obj = json.loads(result)
for keyphrase_analysis in obj['documents']:
    kp=keyphrase_analysis['keyPhrases'][0]


# Detect sentiment.
batch_sentiment_url = base_url + 'text/analytics/v2.0/sentiment'
req = urllib2.Request(batch_sentiment_url, input_texts, headers)
response = urllib2.urlopen(req)
result = response.read()
obj = json.loads(result)
for sentiment_analysis in obj['documents']:
    print('Sentiment ' + ' score: ' + str(sentiment_analysis['score']))
```

- Convert pandas dataframe containing text of interest into a .json dictionary and pass as result to sentiment mapper as a url using urllib2
- Unfortunately I couldn't get this to work at scale. HTTP timeouts, too many requests sent?
- Still interesting for a few cases.
- Would be better if it could map multiple emotions instead of positive/negative.

Future work



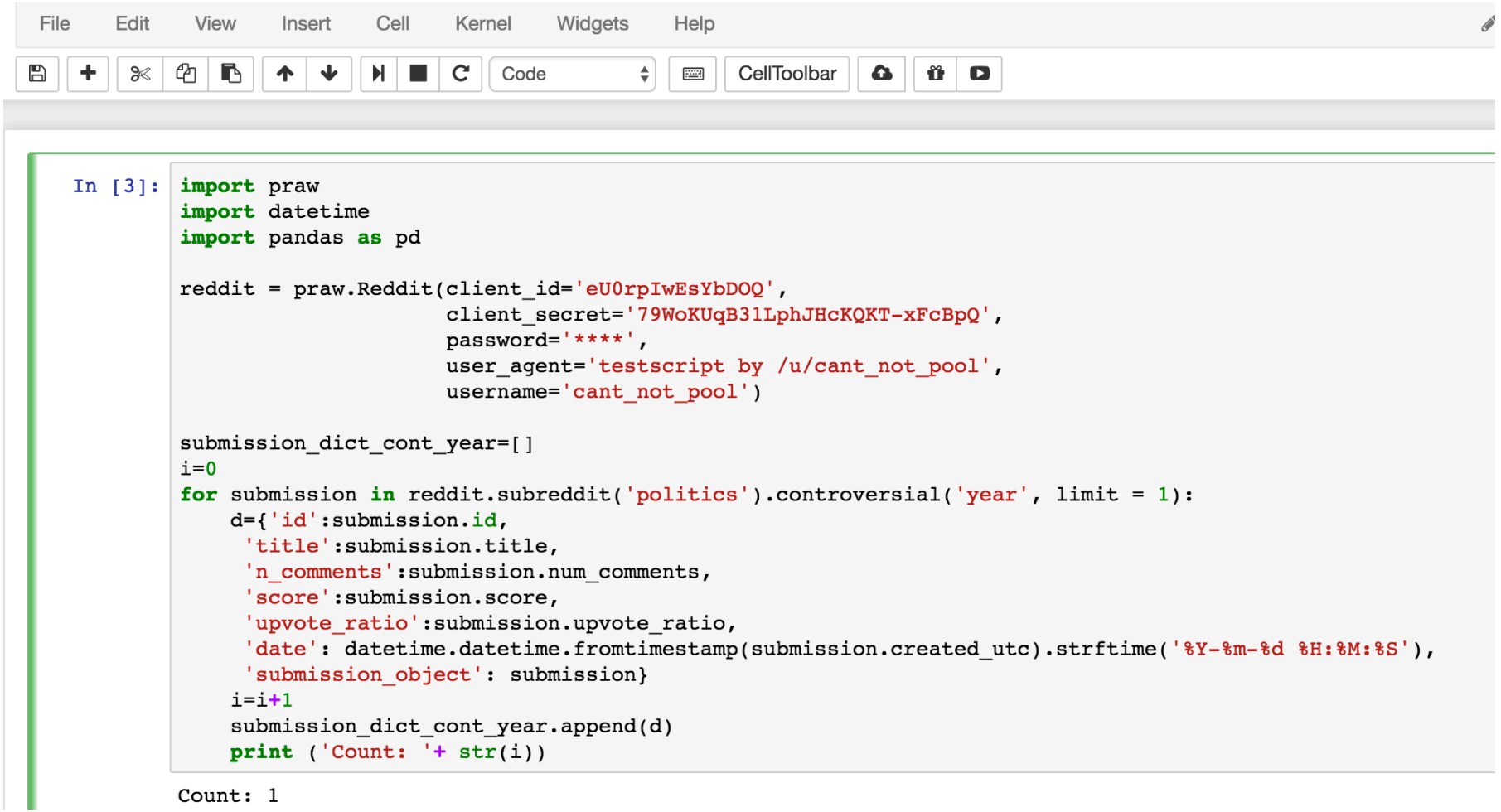
Microsoft syntactic structure detection… soon to be foiled by Trump's sentences?


Anger 0.00000
Contempt 0.94772
Disgust 0.00002
Fear 0.00000
Happiness 0.04198
Neutral 0.01028
Sadness 0.00000
Surprise 0.00000

Appendix

(Methods + Code)

Mining reddit data



- Used jupyter notebook + python to mine reddit data for /r/politics
- Can also collect the data live
- Can specify if you want “top”, or “controversial”
- Known issue – can’t get more than 1000 posts
- Rough/quick code at https://github.com/nslatysheva/hack_cambridge_2017

Output is something like this

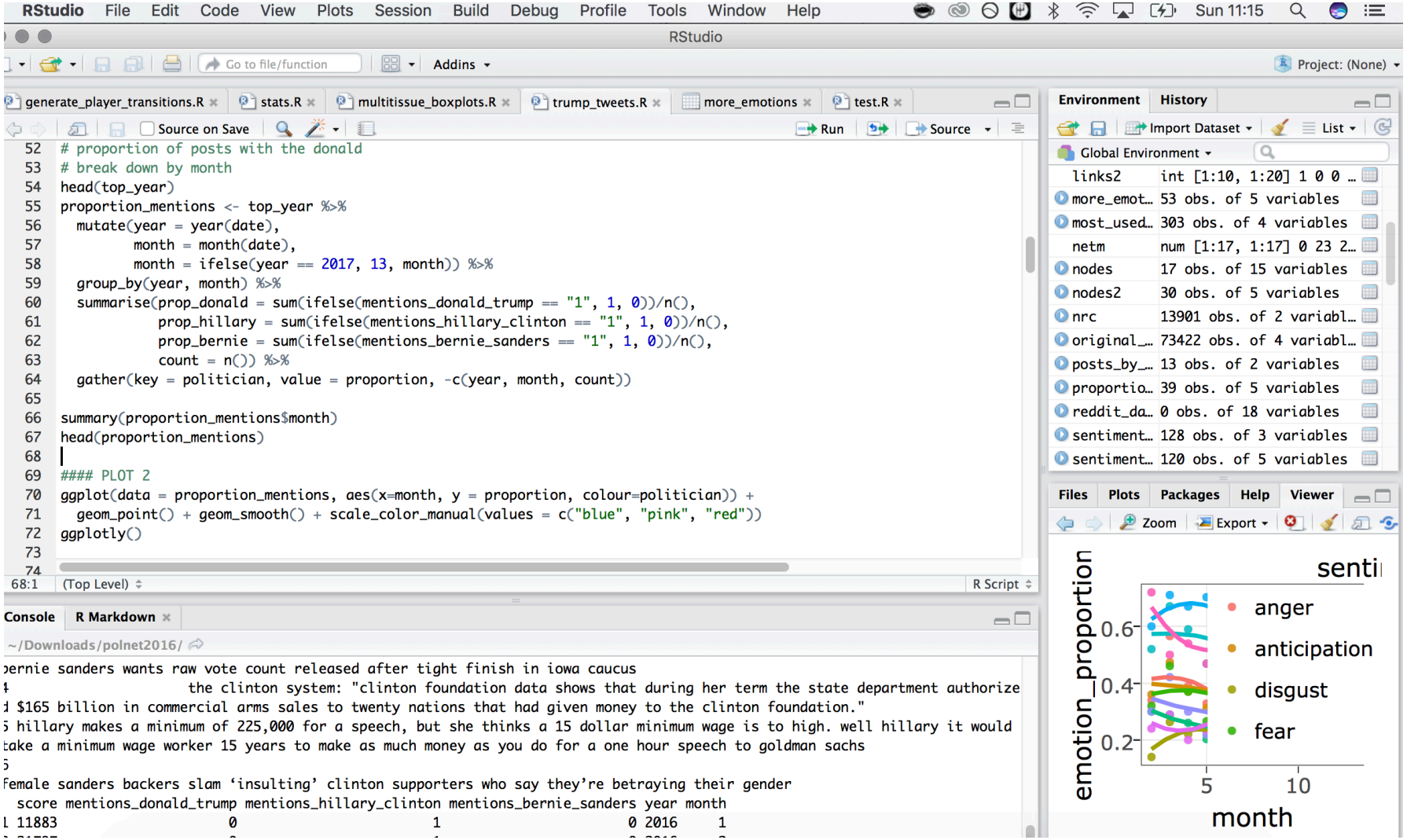
```
#df['top_comment_body'].apply(lambda x: x.encode('ascii', 'ignore'))
df.to_csv('politics_controversial_1000.csv',encoding='utf-8')
```

```
df.head()
```

	date	id	n_comments	score	submission_object	title	upvote_ratio	top_comment_score	top_comment_gilded	top_comment_b
0	2016-11-07 11:19:09	5blmi4	12059	1554	5blmi4	I Was With Bernie Till the End; Now We All Mus...	0.50	10006.0	2.0	I'll vote for whoev the fuck I want\n\n**EDI...
1	2016-08-24 16:02:42	4zd07t	2081	0	4zd07t	My Writing in The Huffington Post, Salon, and ...	0.48	765.0	0.0	Thank you for doi this AMA! \n\nI'd to ...
2	2016-10-13 19:36:08	57buyp	537	52	57buyp	Hi Reddit, I am Maria Teresa Kumar, Emmy-nomin...	0.50	47.0	0.0	Are you surprisec the record low pu tru...
						Hillary				

- Data dictionary then converted to dataframe and exported into R
- Just because I am most comfortable in doing analysis with R

R analysis took a while



- Using Rstudio for R scripting
- Tons of data manipulation to do for every plot basically
- Visualisation with ggplot2, interactivity with ggplotly

Topic and sentiment detection with Azure

In [8]:
```
# Azure portal URL.
base_url = 'https://westus.api.cognitive.microsoft.com/'

# Your account key goes here.
account_key = 'ac455400c61d4c5792c1b4bc57b55715'

headers = {'Content-Type':'application/json', 'Ocp-Apim-Subscription-Key':account_key}

input_texts = '{"documents":['+df[['top_comment_body','title']].to_json()+']}'

num_detect_langs = 1;

# Detect key phrases.
batch_keyphrase_url = base_url + 'text/analytics/v2.0/keyPhrases'
req = urllib2.Request(batch_keyphrase_url, input_texts, headers)
response = urllib2.urlopen(req)
result = response.read()
obj = json.loads(result)
for keyphrase_analysis in obj['documents']:
    kp=keyphrase_analysis['keyPhrases'][0]


# Detect sentiment.
batch_sentiment_url = base_url + 'text/analytics/v2.0/sentiment'
req = urllib2.Request(batch_sentiment_url, input_texts, headers)
response = urllib2.urlopen(req)
result = response.read()
obj = json.loads(result)
for sentiment_analysis in obj['documents']:
    print('Sentiment ' + ' score: ' + str(sentiment_analysis['score']))
```

- Convert pandas dataframe containing text of interest into a .json dictionary and pass as result to sentiment mapper as a url using urllib2
- Unfortunately I couldn't get this to work at scale. HTTP timeouts, too many requests sent?
- Still interesting for a few cases.
- Would be better if it could map multiple emotions instead of positive/negative.

Failed attempt: Pusher with Reddit API stream

```
Natashas-MBP:hack_cambridge_2017 nat$ node pusher
/Users/nat/Projects/random_things/hack_cambridge_
var app = new PusherPlatform.App({
          ^

ReferenceError: PusherPlatform is not defined
    at Object.<anonymous> (/Users/nat/Projects/ra
    at Module._compile (module.js:570:32)
    at Object.Module._extensions..js (module.js:5
    at Module.load (module.js:487:32)
    at tryModuleLoad (module.js:446:12)
    at Function.Module._load (module.js:438:3)
    at Module.runMain (module.js:604:10)
    at run (bootstrap_node.js:394:7)
    at startup (bootstrap_node.js:149:9)
    at bootstrap_node.js:509:3
```

```
1
2  var app = new PusherPlatform.App({
3    appId: 'cbf29c93-840e-421c-af69-856cfcba01b5',
4  });
5
6  var myFeed = app.feed('playground');
7
8  myFeed.subscribe({
9    onOpen: () => console.log('Connection established'),
10   onItem: item => console.log('Item:', item),
11   onError: error => console.error('Error:', error),
12 });
13
14 myFeed.append('Hello, world!')
15   .then(response => console.log('Success:', response))
16   .catch(err => console.error('Error:', err));
17
18 // You're not limited to appending string values;
19 // you can also append objects, arrays and numbers.
20 myFeed.append({ yourKey: 'your value' })
21   .then(response => console.log('Success:', response))
22   .catch(err => console.error('Error:', err));
23
```

- Main problem: I don't know javascript (???)
- Installed and tried to use node.js
- Error codes a mystery
- Not enough time
- Decided to just stick to Python and R and data ;P