Github repo: Blog-Sentiment-Analysis

View these slides: bit.ly/2IxJ8i1

Blog sentiment analysis*

Eva Bacas

* and other stuff

Project overview

- Exploration of the Blog Authorship Corpus
  - 680,000 blogs by 20,000 bloggers
- Dataset ⇄ research questions
  - Machine learning techniques + sociolinguistic variation

bit.ly/2lxJ8i1

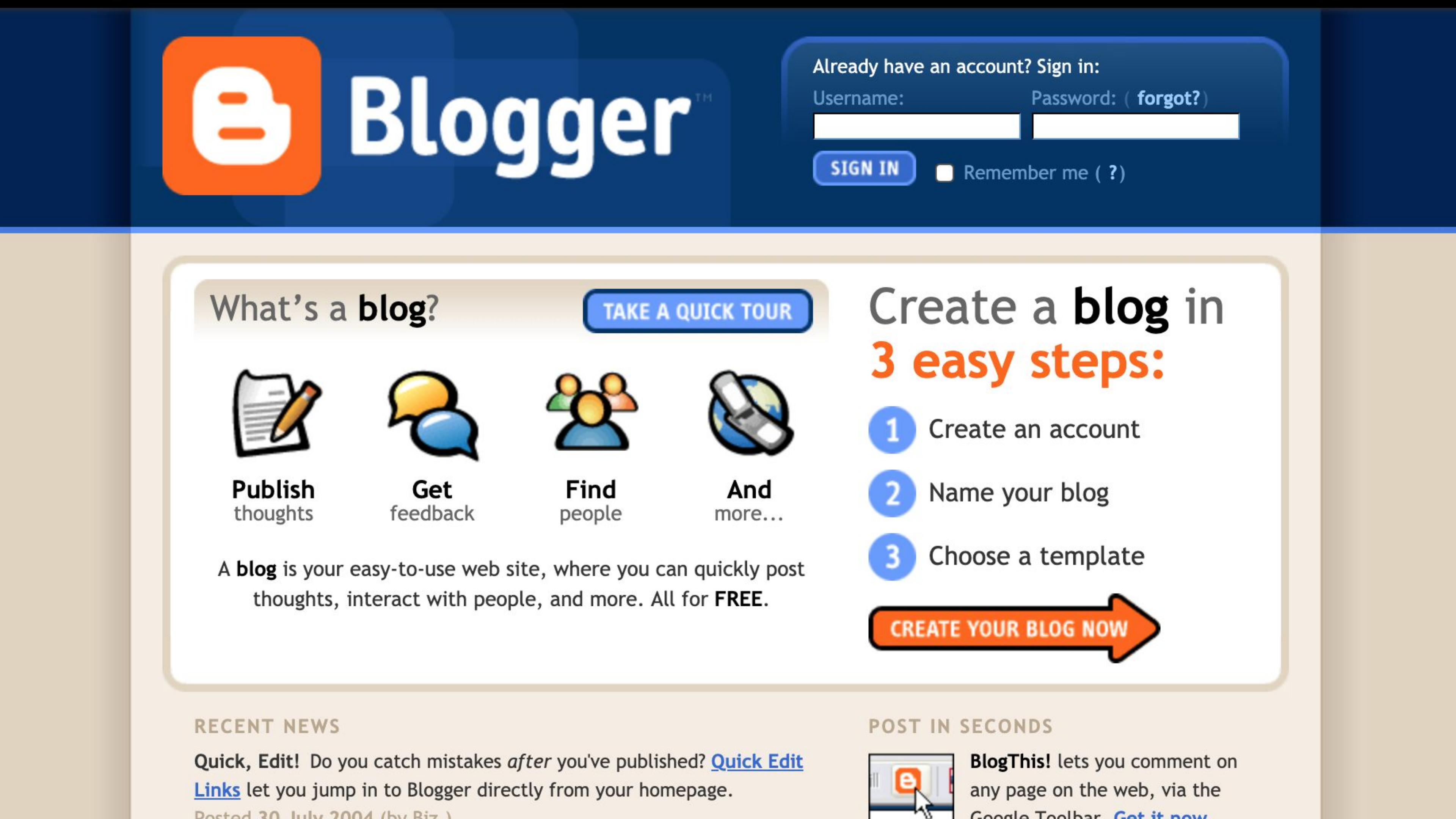
Big goals of this project

- What information can I get out of a text corpus?
- How can I use machine learning to extract that information?
- How can I show that information with data visualization?

bit.ly/2lxJ8i1

Specific goals of this project

- Word frequencies
- Blog topics
- Blog sentiment



What's a **blog**?

A **blog** is your easy-to-use web site, where you can quickly post thoughts, interact with people, and more. All for **FREE**.

Create a **blog** in 3 easy steps:

1. Create an account
2. Name your blog
3. Choose a template

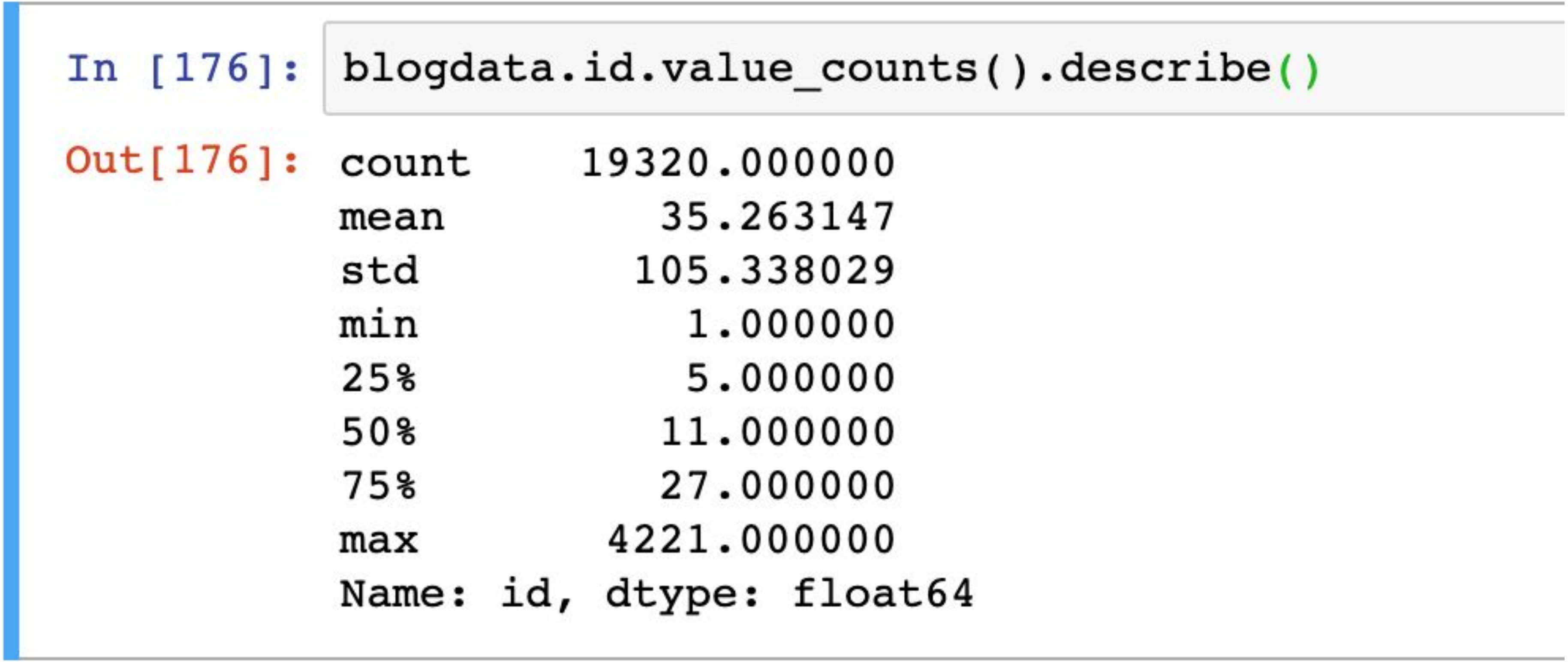
RECENT NEWS

Quick, Edit! Do you catch mistakes *after* you've published? Quick Edit Links let you jump in to Blogger directly from your homepage.

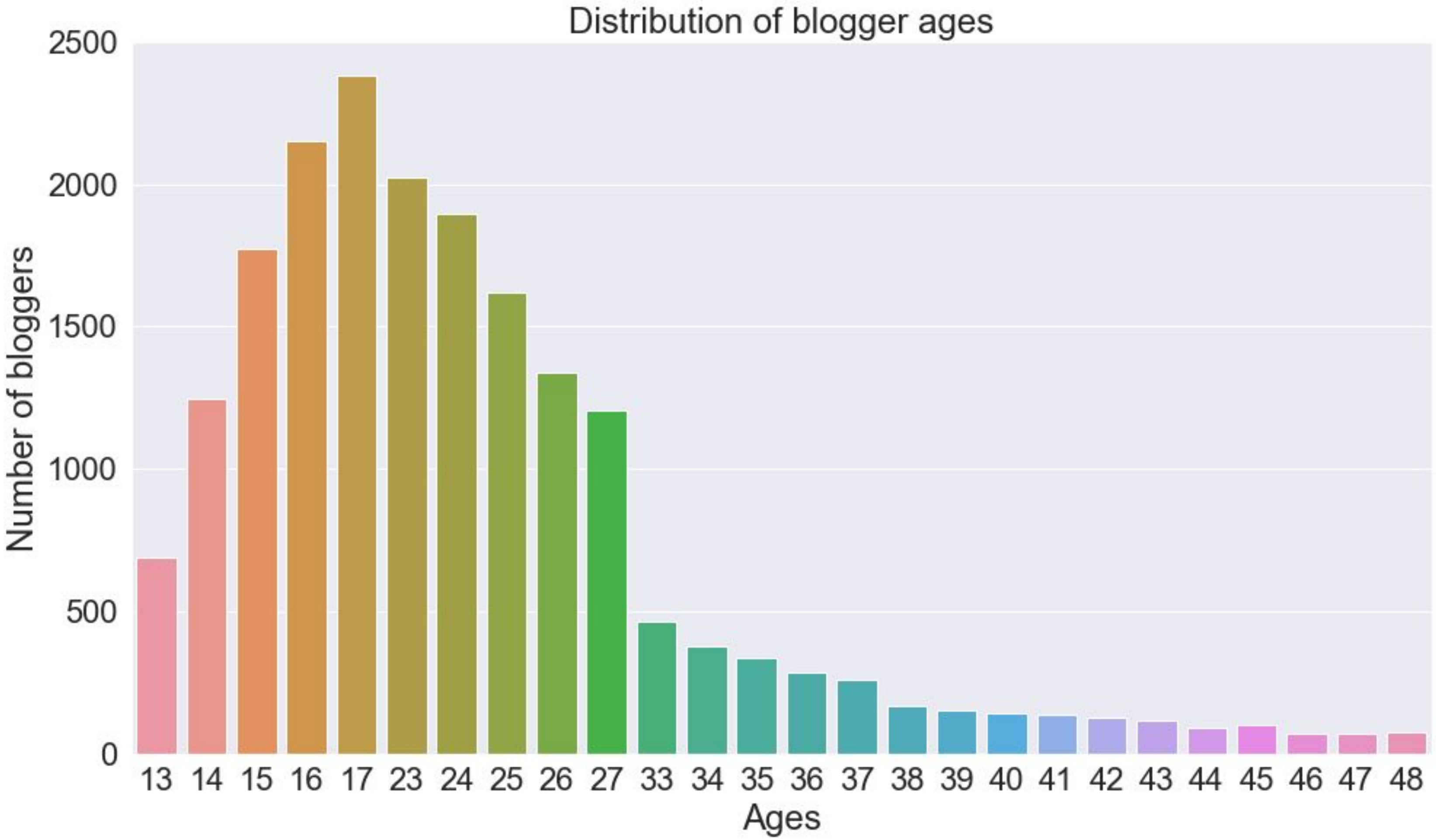
Posted 30 July 2004 (by Biz.)

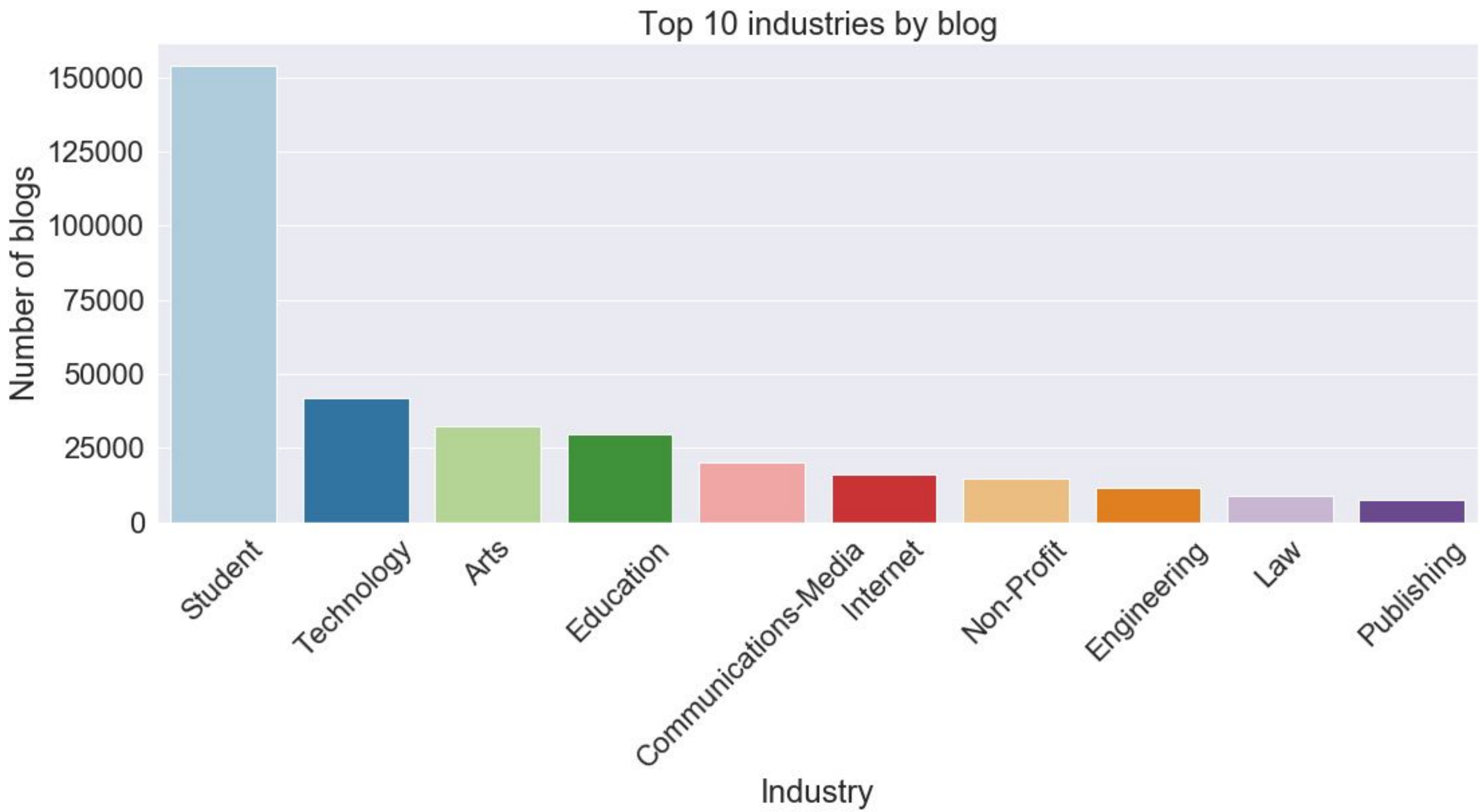
POST IN SECONDS

BlogThis! lets you comment on any page on the web, via the Google Toolbar. Get it now.

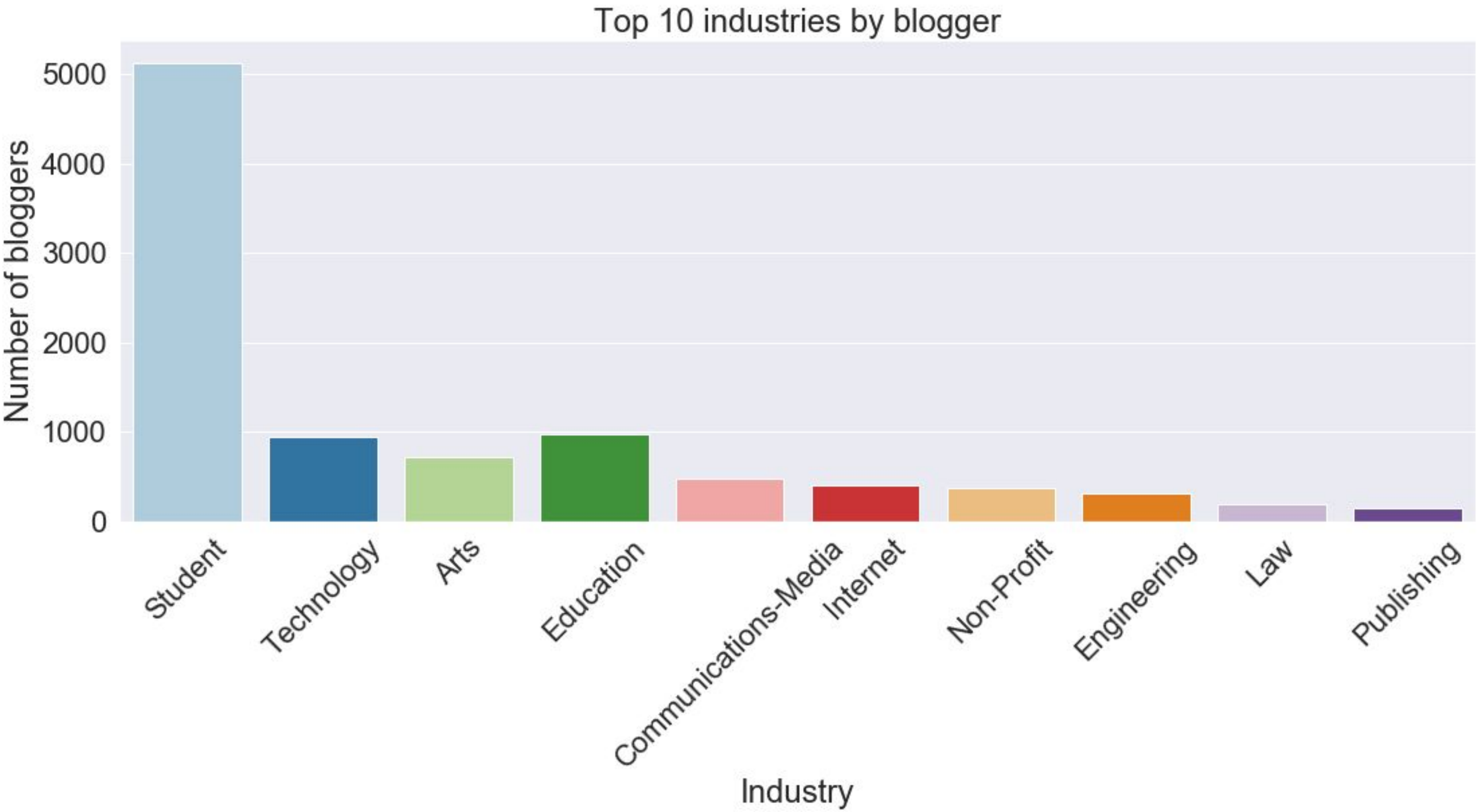


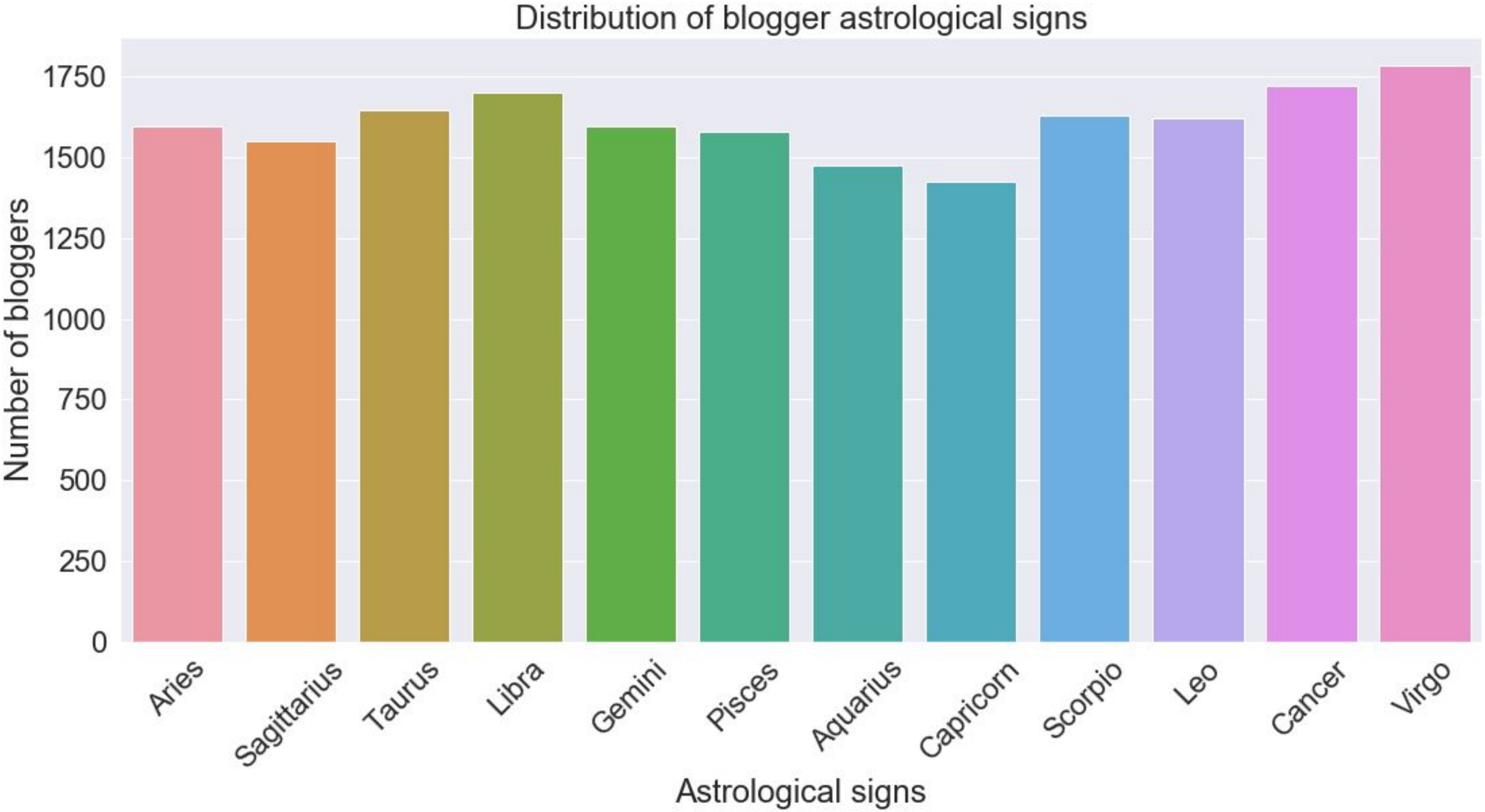
```
In [176]: blogdata.id.value_counts().describe()

Out[176]: count    19320.000000
          mean        35.263147
          std        105.338029
          min          1.000000
          25%          5.000000
          50%         11.000000
          75%         27.000000
          max       4221.000000
          Name: id, dtype: float64
```


Distribution of blogger ages
Number of bloggers
2500
2000
1500
1000
500
0
13
14
15
16
17
23
24
25
26
27
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
Ages



Top 10 industries by blog
Number of blogs
150000
125000
100000
75000
50000
25000
0
Student
Technology
Arts
Education
Communications-Media
Internet
Non-Profit
Engineering
Law
Publishing
Industry

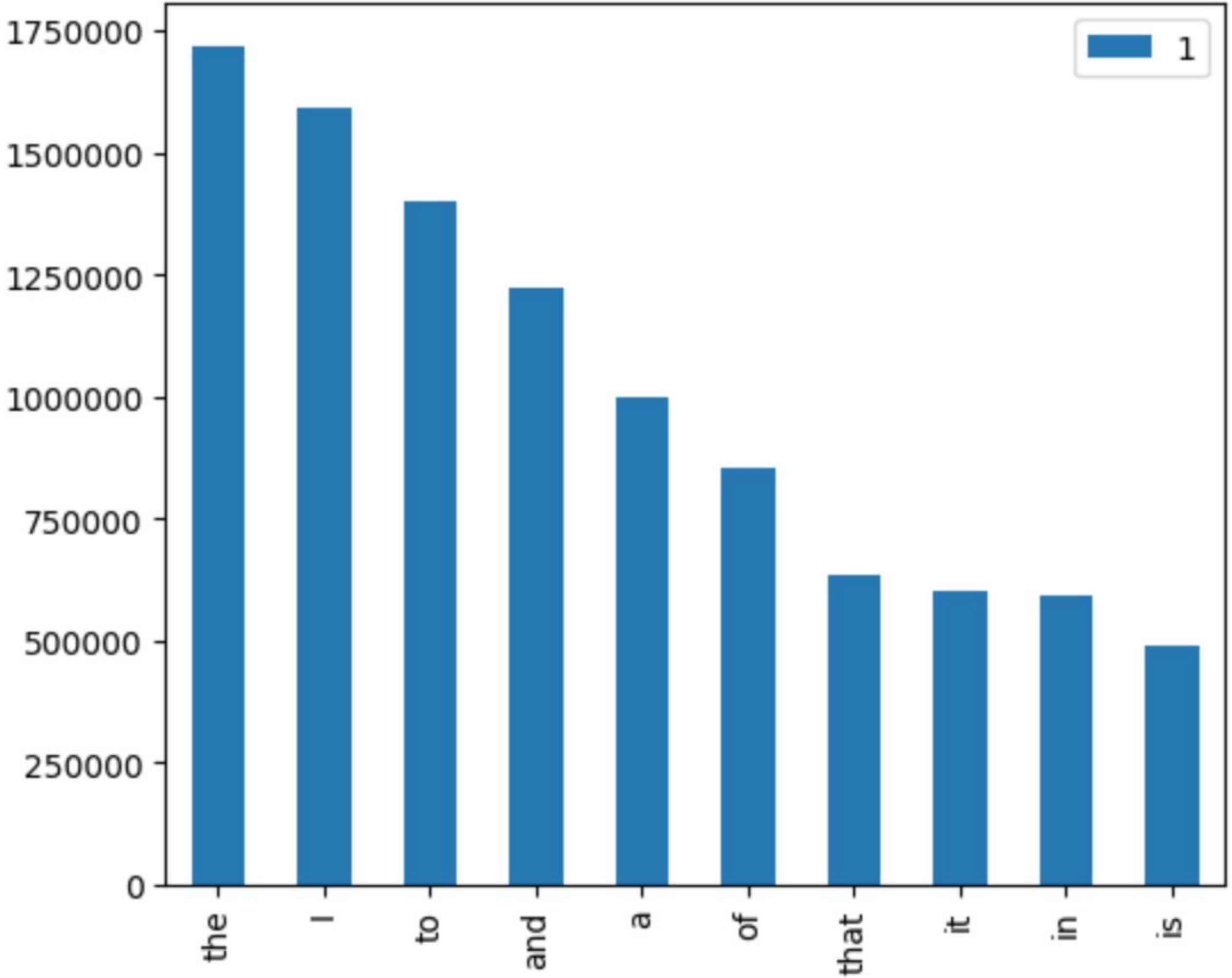

Top 10 industries by blogger
Number of bloggers
5000
4000
3000
2000
1000
0
Student
Technology
Arts
Education
Communications-Media
Internet
Non-Profit
Engineering
Law
Publishing
Industry

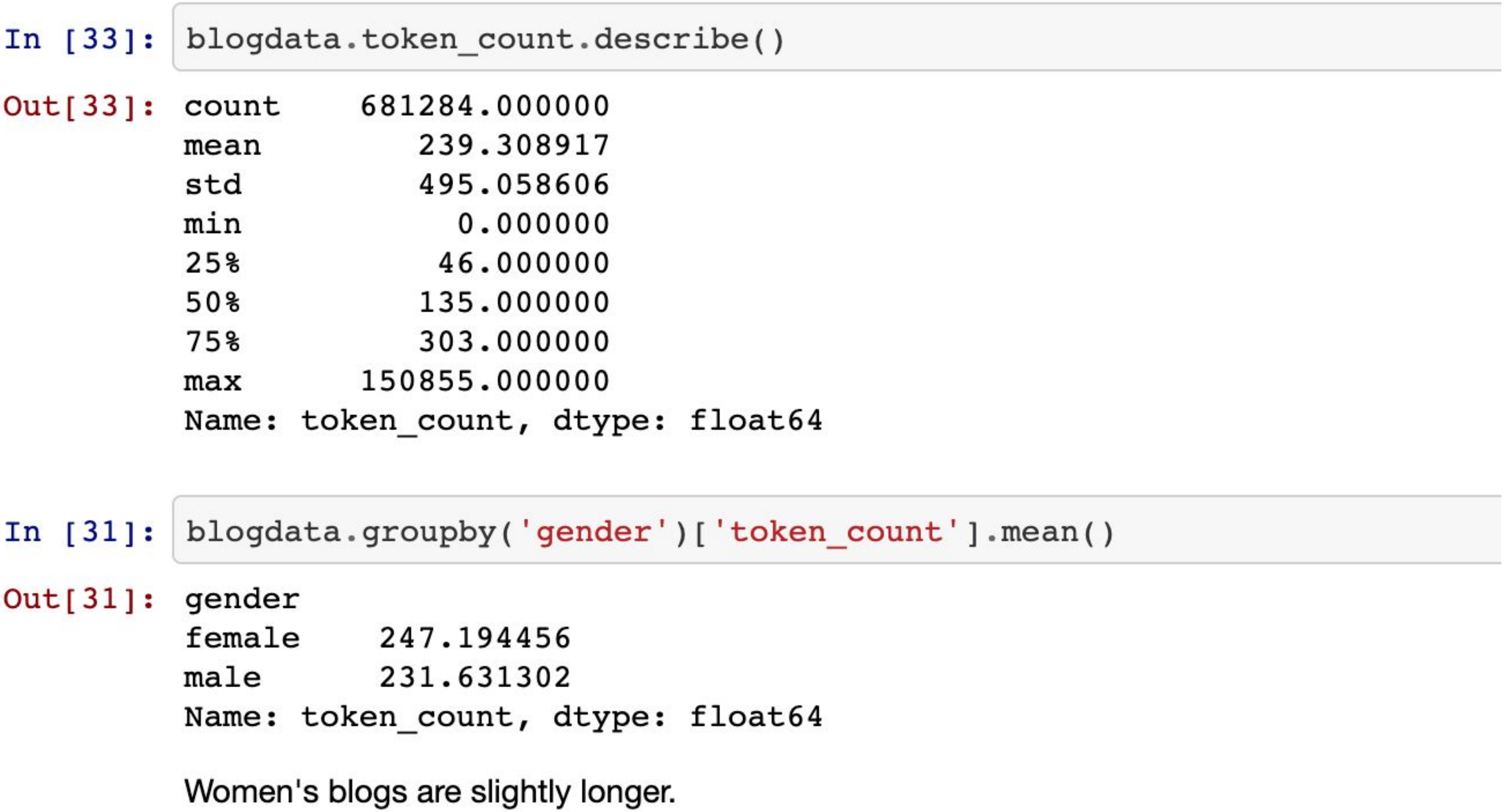

Distribution of blogger astrological signs
Number of bloggers
0
250
500
750
1000
1250
1500
1750
Aries
Sagittarius
Taurus
Libra
Gemini
Pisces
Aquarius
Capricorn
Scorpio
Leo
Cancer
Virgo
Astrological signs

Word frequencies

- Variation across industry
- Issues looking at word frequency
  - Varying length of text
  - Not all the texts are in English...
  - Tokenization issue
- Will revisit for sentiment

Top 10 most frequent words for indUnk





```
In [33]: blogdata.token_count.describe()

Out[33]: count    681284.000000
         mean        239.308917
         std         495.058606
         min           0.000000
         25%          46.000000
         50%         135.000000
         75%         303.000000
         max      150855.000000
         Name: token_count, dtype: float64
```

```
In [31]: blogdata.groupby('gender')['token_count'].mean()

Out[31]: gender
         female    247.194456
         male      231.631302
         Name: token_count, dtype: float64
```

Women's blogs are slightly longer.

Topic clustering

- Latent Dirichlet Allocation (LDA) with scikit-learn
  - Used with TF-IDF
- How it works: https://lettier.com/projects/lda-topic-modeling/

Topic #0: na la ko sa da

Topic #1: snow exam weather ben cold

Topic #2: youre dont like im just

Topic #3: urllink google search information art

Topic #4: dont im just know want

Topic #5: love life know heart just

Topic #6: school class im teacher just

Topic #7: book read books reading im

Topic #8: pm break spring im ah

Topic #9: site urllink page web website

Topic #10: wedding beach im apartment lake

Topic #11: baby women dave child woman

Topic #12: god jesus lord life christ

Topic #13: birthday happy party day im

Topic #14: dun den juz wat tt

Topic #15: movie movies film watch urllink

Topic #16: test don questions ve question

Topic #17: car road bike just driving

Topic #18: hair black color red like

Topic #19: hello dance kiss fly say

Topic #20: im boring nick going really

Topic #21: dad mom im paul just

Topic #22: bush urllink kerry war president

Topic #23: dog photos dogs concert tickets

Topic #24: lol im like ur gonna

Topic #25: bye welcome jumper991 im vacation

Topic #26: nbsp urllink pictures photo picture

Topic #27: fucking fuck shit im like

Topic #28: game team games win play

Topic #29: email hate im just dont

Topic #30: eyes like just sun sky

Topic #31: went got home like fun

Topic #32: said just like got went

Topic #33: blog post blogger im comments

Topic #34: weekend night saturday friday went

Topic #35: music song band songs urllink

Topic #36: yay im homework today math

Topic #37: im lunch eat today day

Topic #38: ha haha hi im oh

Topic #39: cheese 12 chocolate cream like

Topic #40: urllink people government states world

Topic #41: christmas phone hotel cell im

Topic #42: pants suck ball im clothes

Topic #43: urllink 2004 2003 july june

Topic #44: urllink brought link click check

Topic #45: job im doctor hospital work

Topic #46: im listening mood beer just

Topic #47: sleep im bed tired night

Topic #48: account card money internet computer

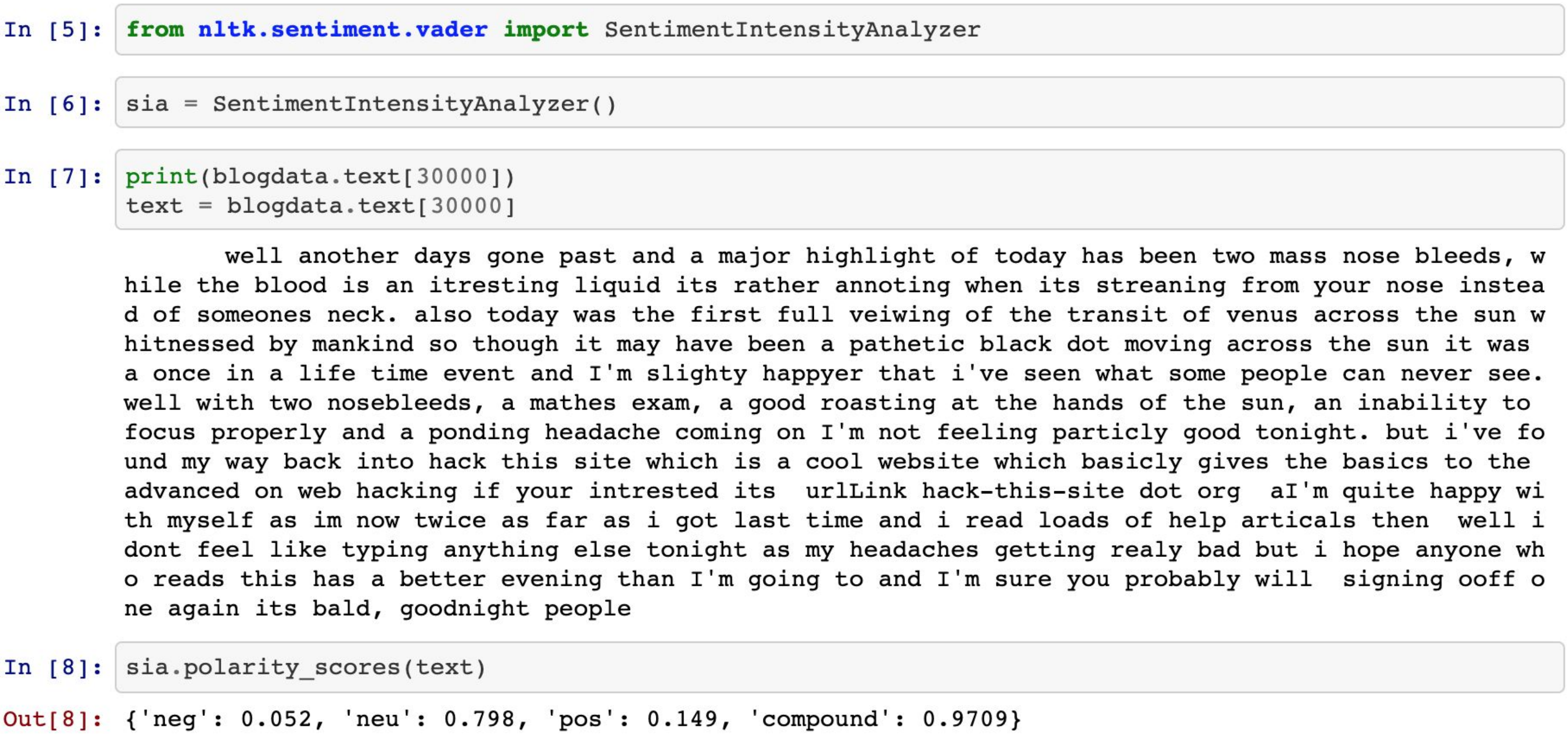
Topic #49: church group youth christian anybody

Better (?) topics

- Topic #0: blog post just
- Topic #1: car phone just
- Topic #2: game team birthday
- Topic #3: urllink nbsp brought
- Topic #4: went got night
- Topic #5: god church jesus
- Topic #6: like just hair
- Topic #7: movie night just
- Topic #8: nbsp urllink 2004
- Topic #9: urllink book people
- Topic #10: im dont lol
- Topic #11: class school test
- Topic #12: just don know
- Topic #13: na just like
- Topic #14: urllink com http
- Topic #15: urllink bush war
- Topic #16: love life like
- Topic #17: going just fun
- Topic #18: haha la den
- Topic #19: music song urllink

Sentiment analysis

- Sentiments
  - Positive vs. negative
- VADER (Valence Aware Dictionary and sEntiment Reasoner)
  - “Specifically attuned to sentiments expressed in social media”
  - Polarity score (-1 to 1)



```
In [5]: from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

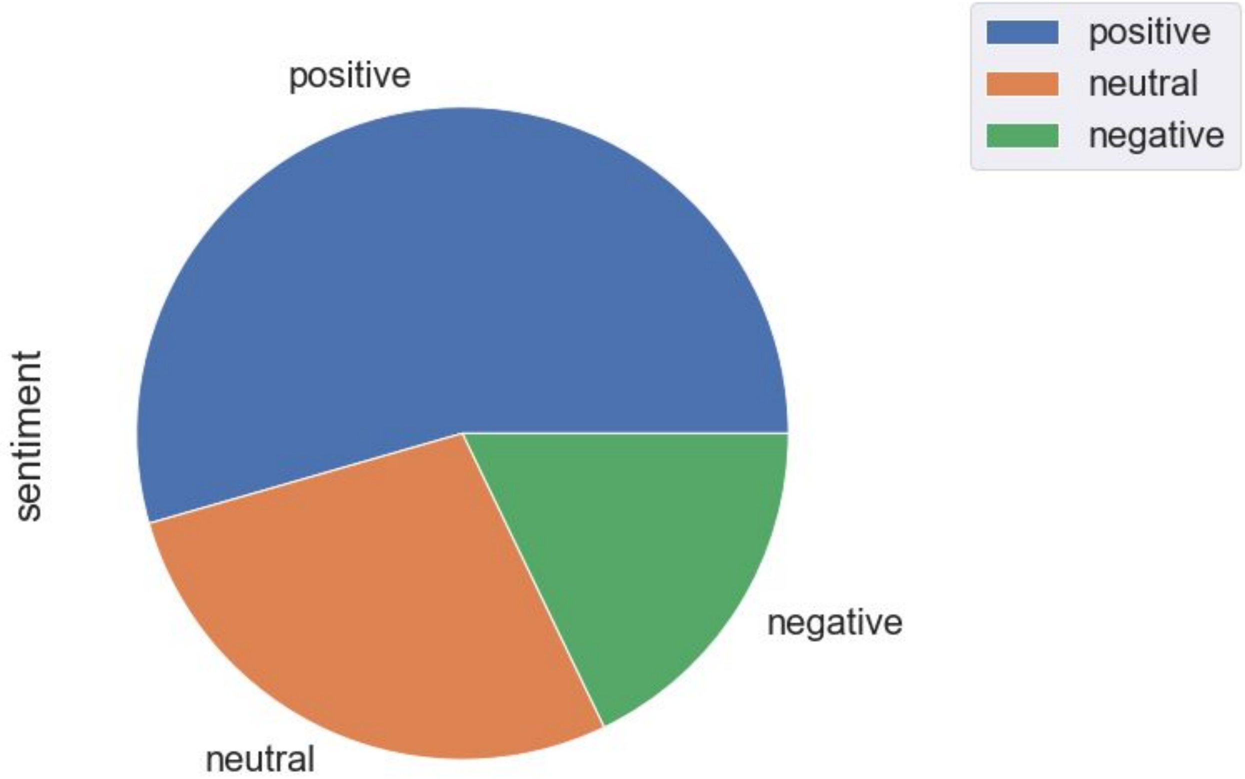
```
In [6]: sia = SentimentIntensityAnalyzer()
```

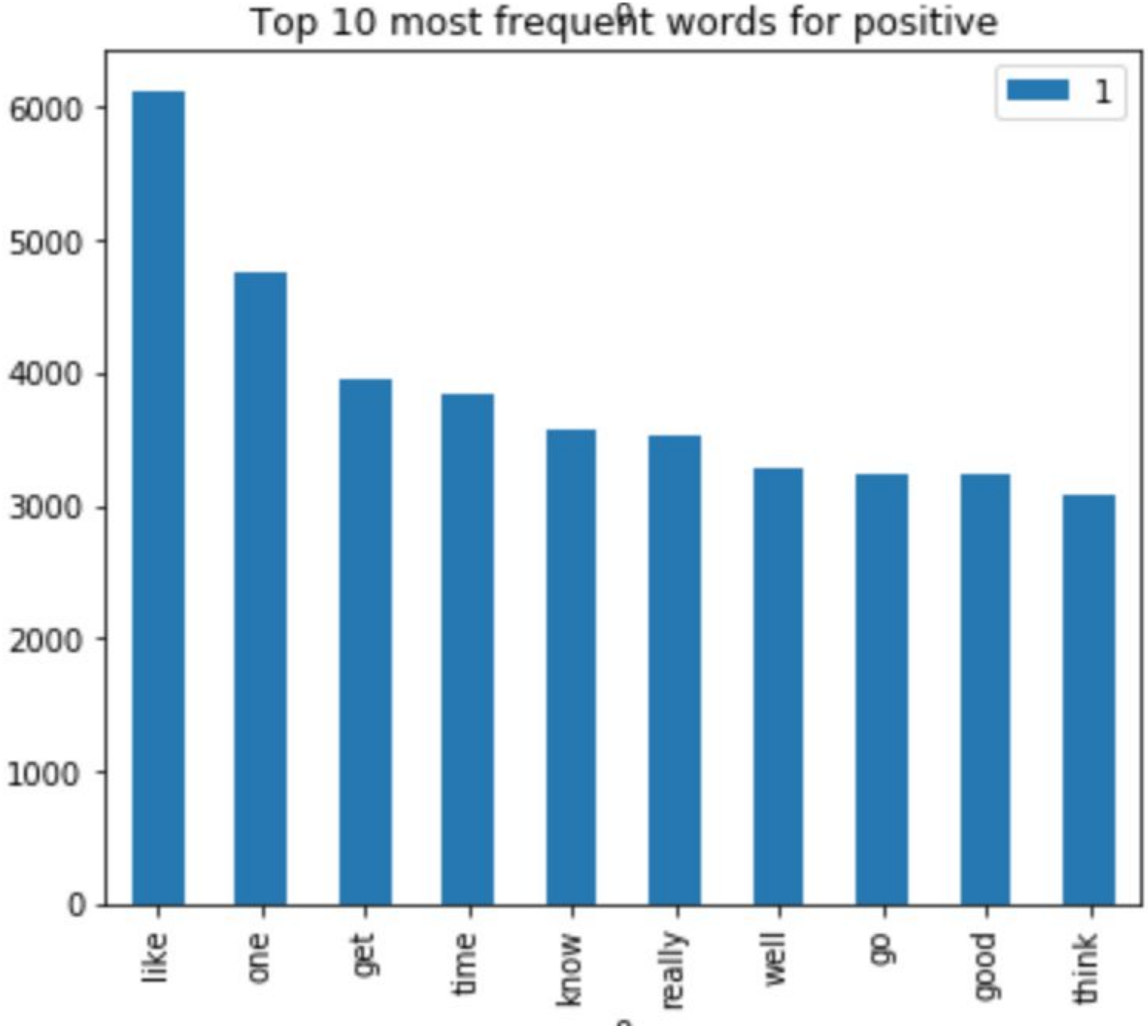
```
In [7]: print(blogdata.text[30000])
        text = blogdata.text[30000]
```

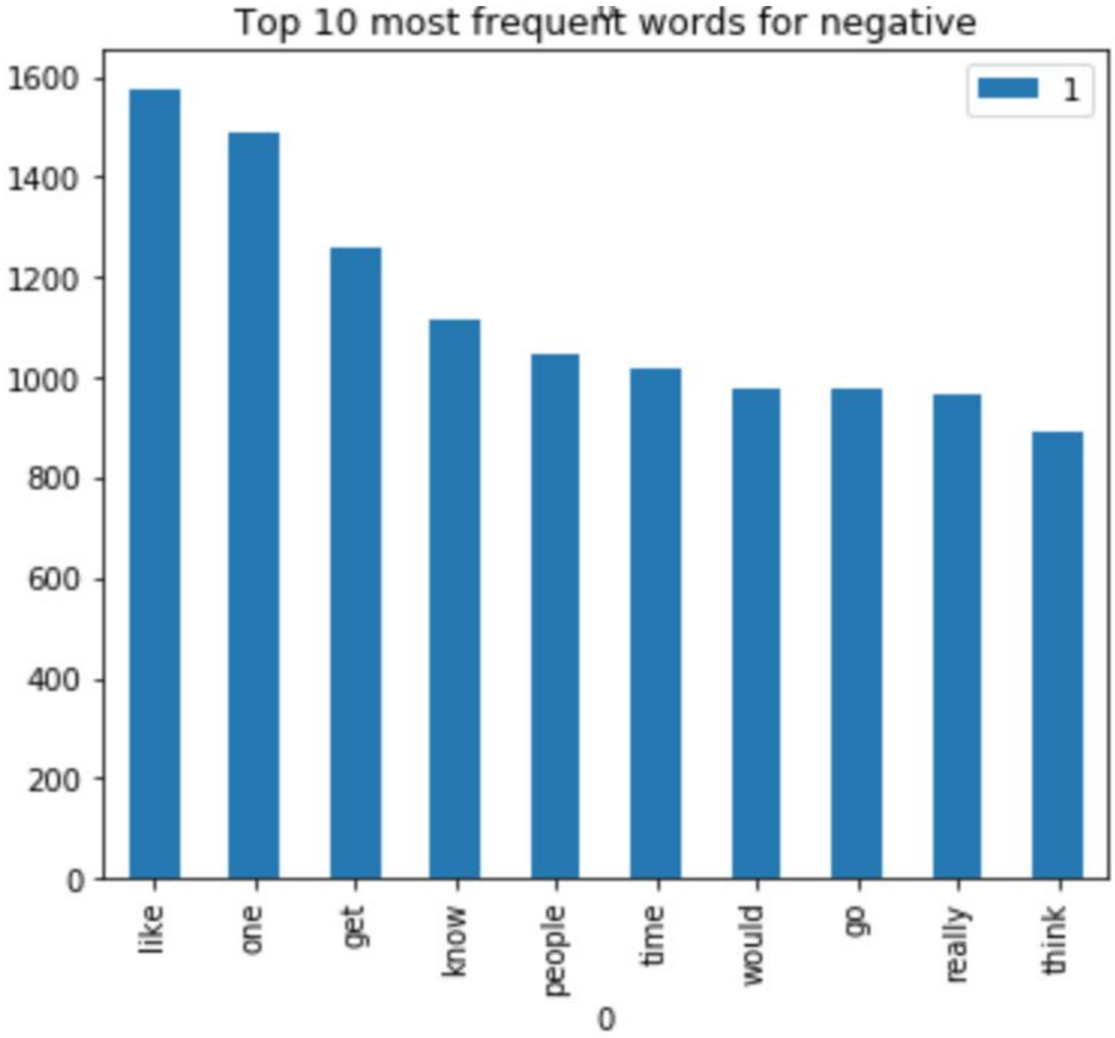
```
               well another days gone past and a major highlight of today has been two mass nose bleeds, w
        hile the blood is an itresting liquid its rather annoting when its streaning from your nose instea
        d of someones neck. also today was the first full veiwing of the transit of venus across the sun w
        hitnessed by mankind so though it may have been a pathetic black dot moving across the sun it was
        a once in a life time event and I'm slighty happyer that i've seen what some people can never see.
        well with two nosebleeds, a mathes exam, a good roasting at the hands of the sun, an inability to
        focus properly and a ponding headache coming on I'm not feeling particly good tonight. but i've fo
        und my way back into hack this site which is a cool website which basicly gives the basics to the
        advanced on web hacking if your intrested its  urlLink hack-this-site dot org  aI'm quite happy wi
        th myself as im now twice as far as i got last time and i read loads of help articals then  well i
        dont feel like typing anything else tonight as my headaches getting realy bad but i hope anyone wh
        o reads this has a better evening than I'm going to and I'm sure you probably will  signing ooff o
        ne again its bald, goodnight people
```

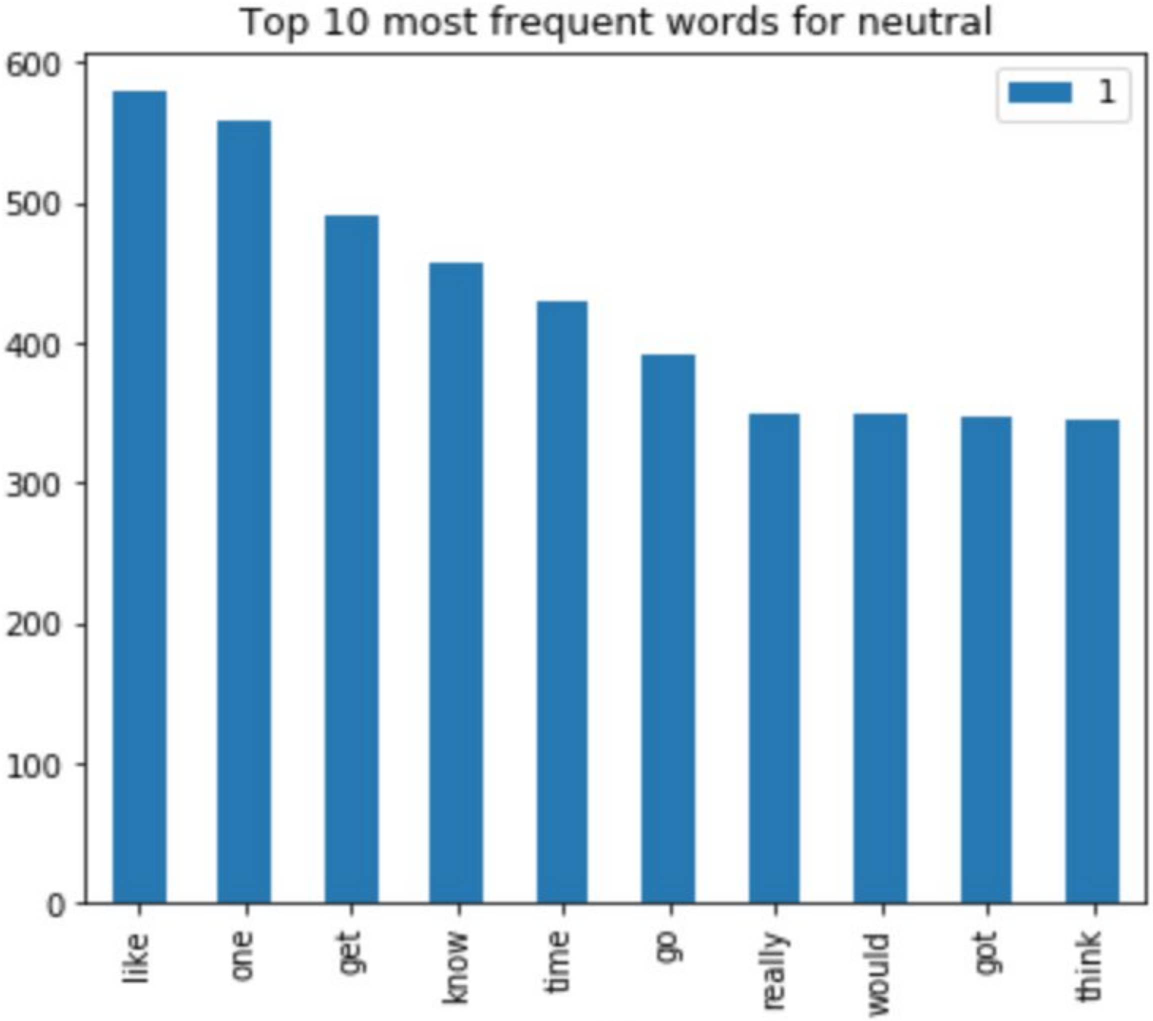
```
In [8]: sia.polarity_scores(text)
```

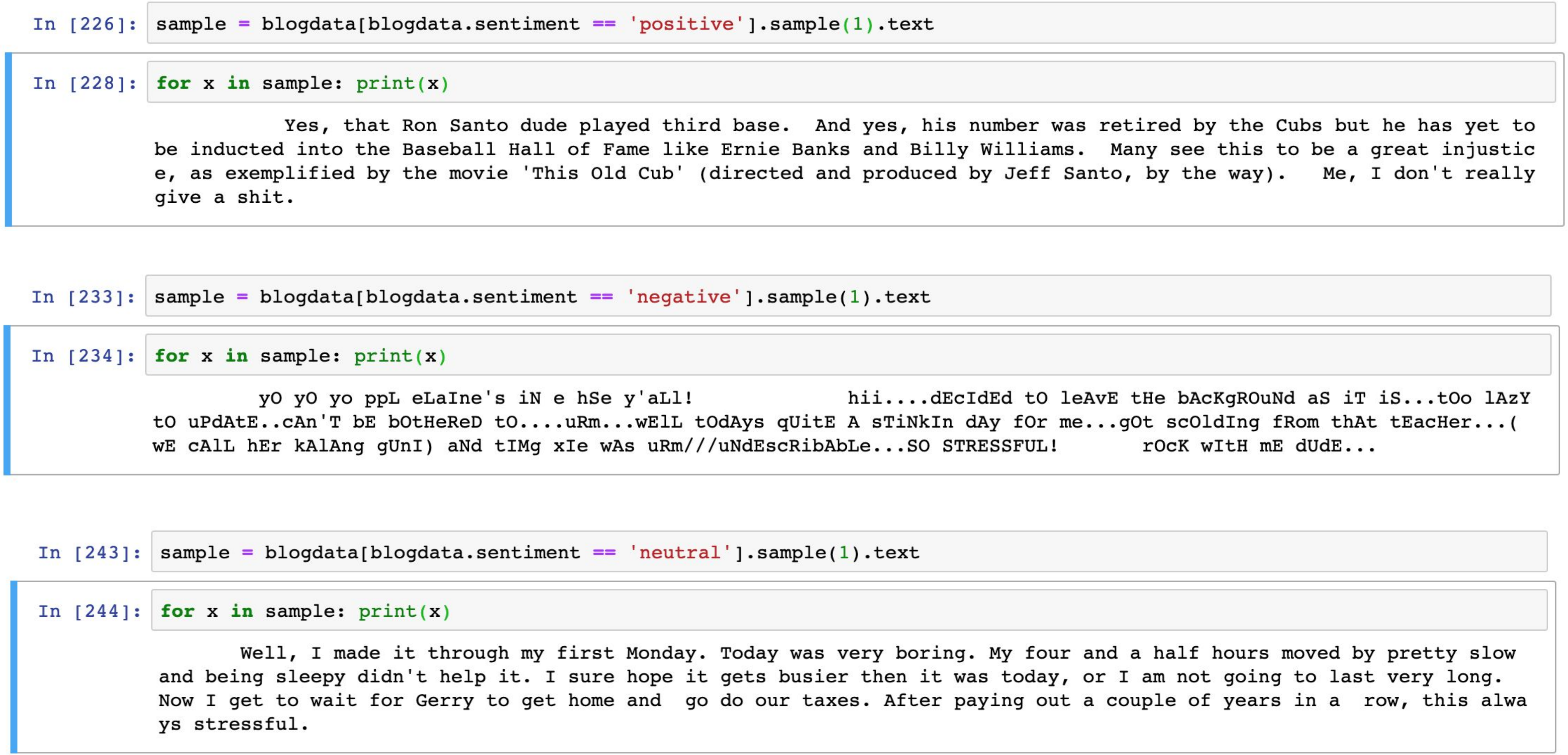
```
Out[8]: {'neg': 0.052, 'neu': 0.798, 'pos': 0.149, 'compound': 0.9709}
```


positive
neutral
negative
sentiment
positive
neutral
negative


Top 10 most frequent words for positive
0
1
6000
5000
4000
3000
2000
1000
0
like
one
get
time
know
really
well
go
good
think


Top 10 most frequent words for negative
1600
1400
1200
1000
800
600
400
200
0
1
like
one
get
know
people
time
would
go
really
think
0


Top 10 most frequent words for neutral
1
600
500
400
300
200
100
0
like
one
get
know
time
go
really
would
got
think



```
In [226]: sample = blogdata[blogdata.sentiment == 'positive'].sample(1).text
```

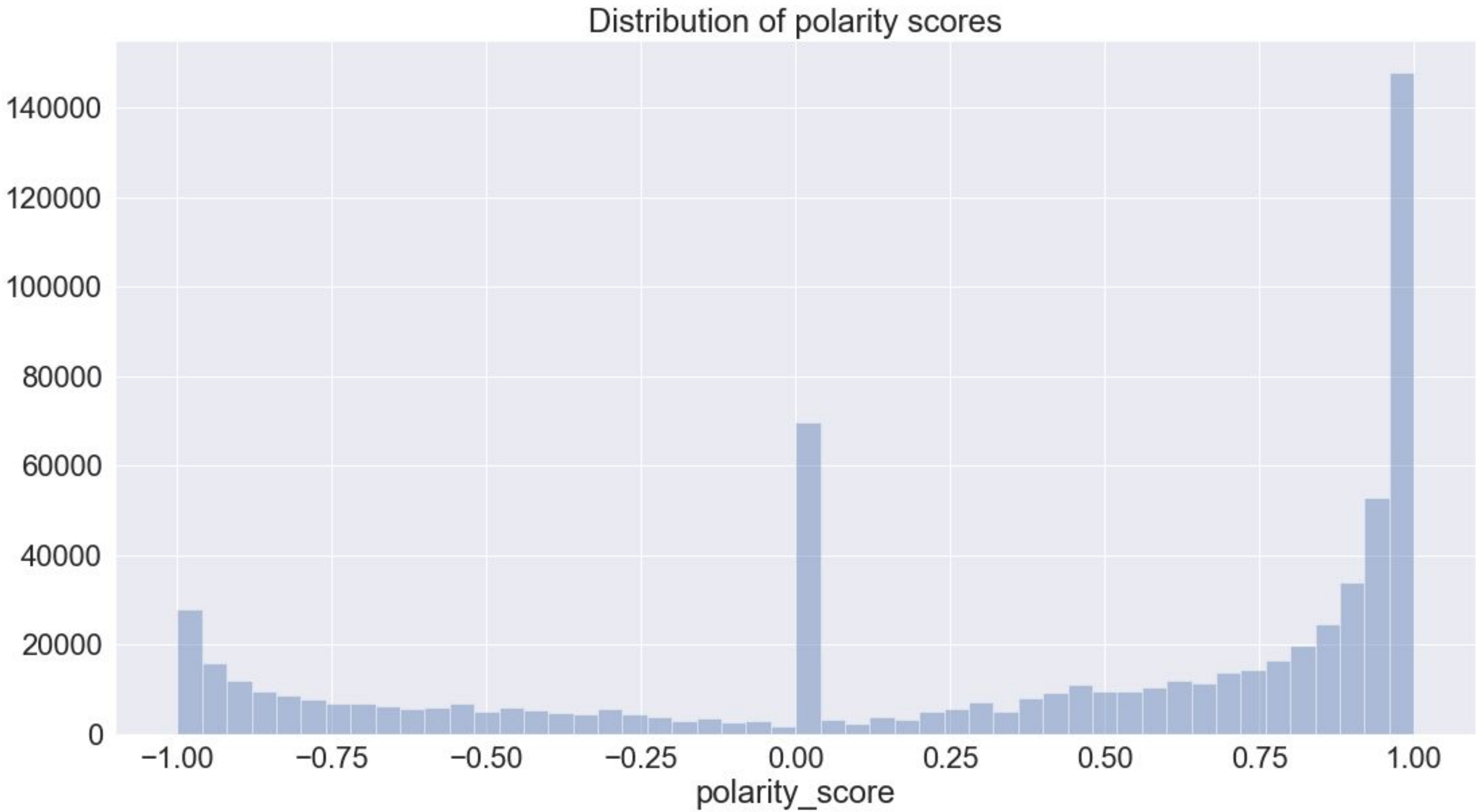
```
In [228]: for x in sample: print(x)

                     Yes, that Ron Santo dude played third base.  And yes, his number was retired by the Cubs but he has yet to
          be inducted into the Baseball Hall of Fame like Ernie Banks and Billy Williams.  Many see this to be a great injustic
          e, as exemplified by the movie 'This Old Cub' (directed and produced by Jeff Santo, by the way).   Me, I don't really
          give a shit.
```

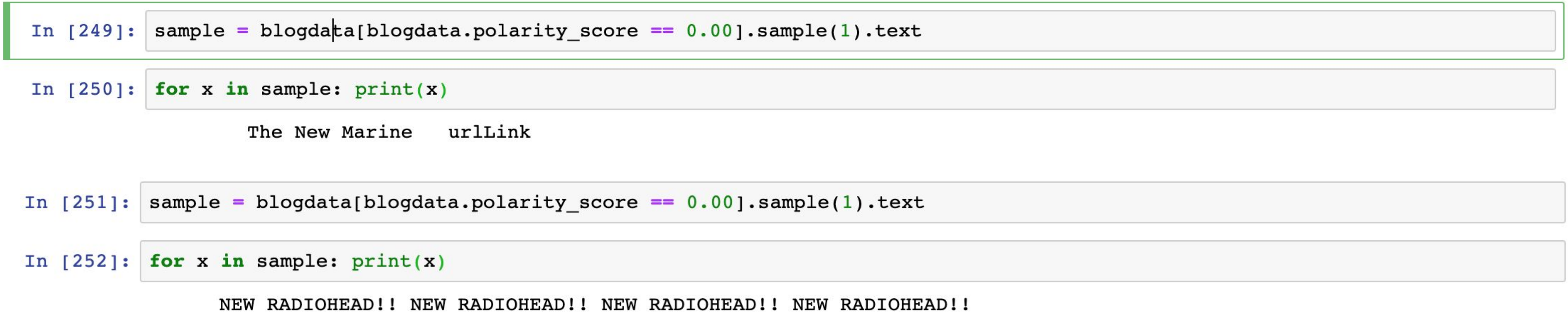
```
In [233]: sample = blogdata[blogdata.sentiment == 'negative'].sample(1).text
```

```
In [234]: for x in sample: print(x)

                 yO yO yo ppL eLaIne's iN e hSe y'aLl!             hii....dEcIdEd tO leAvE tHe bAcKgROuNd aS iT iS...tOo lAzY
          tO uPdAtE..cAn'T bE bOtHeReD tO....uRm...wElL tOdAys qUitE A sTiNkIn dAy fOr me...gOt scOldIng fRom thAt tEacHer...(
          wE cAlL hEr kAlAng gUnI) aNd tIMg xIe wAs uRm///uNdEscRibAbLe...SO STRESSFUL!       rOcK wItH mE dUdE...
```

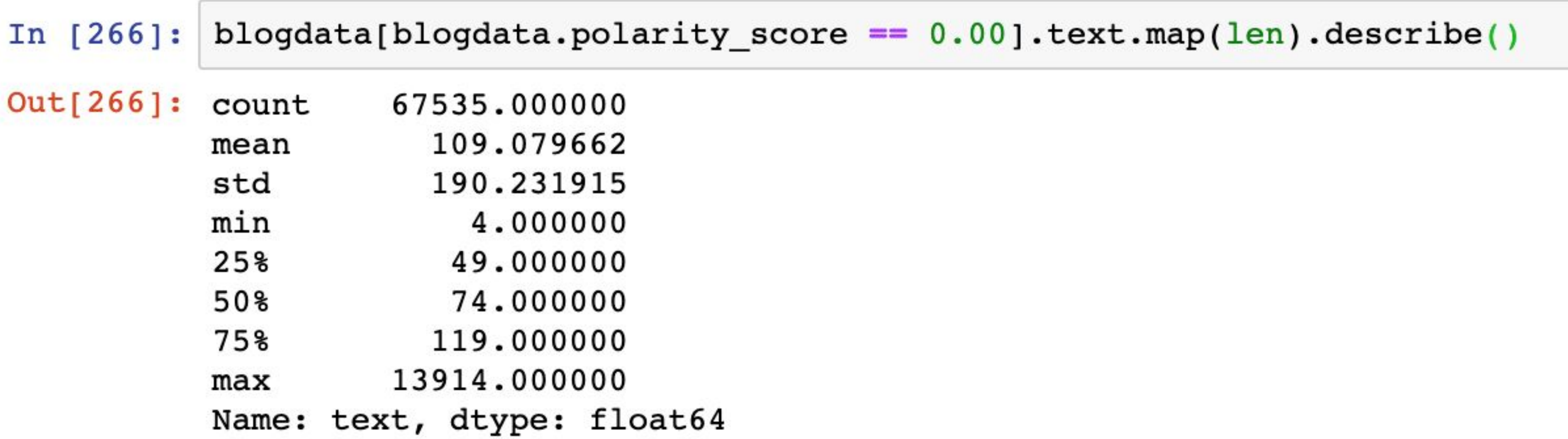
```
In [243]: sample = blogdata[blogdata.sentiment == 'neutral'].sample(1).text
```

```
In [244]: for x in sample: print(x)

                  Well, I made it through my first Monday. Today was very boring. My four and a half hours moved by pretty slow
          and being sleepy didn't help it. I sure hope it gets busier then it was today, or I am not going to last very long.
          Now I get to wait for Gerry to get home and  go do our taxes. After paying out a couple of years in a  row, this alwa
          ys stressful.
```

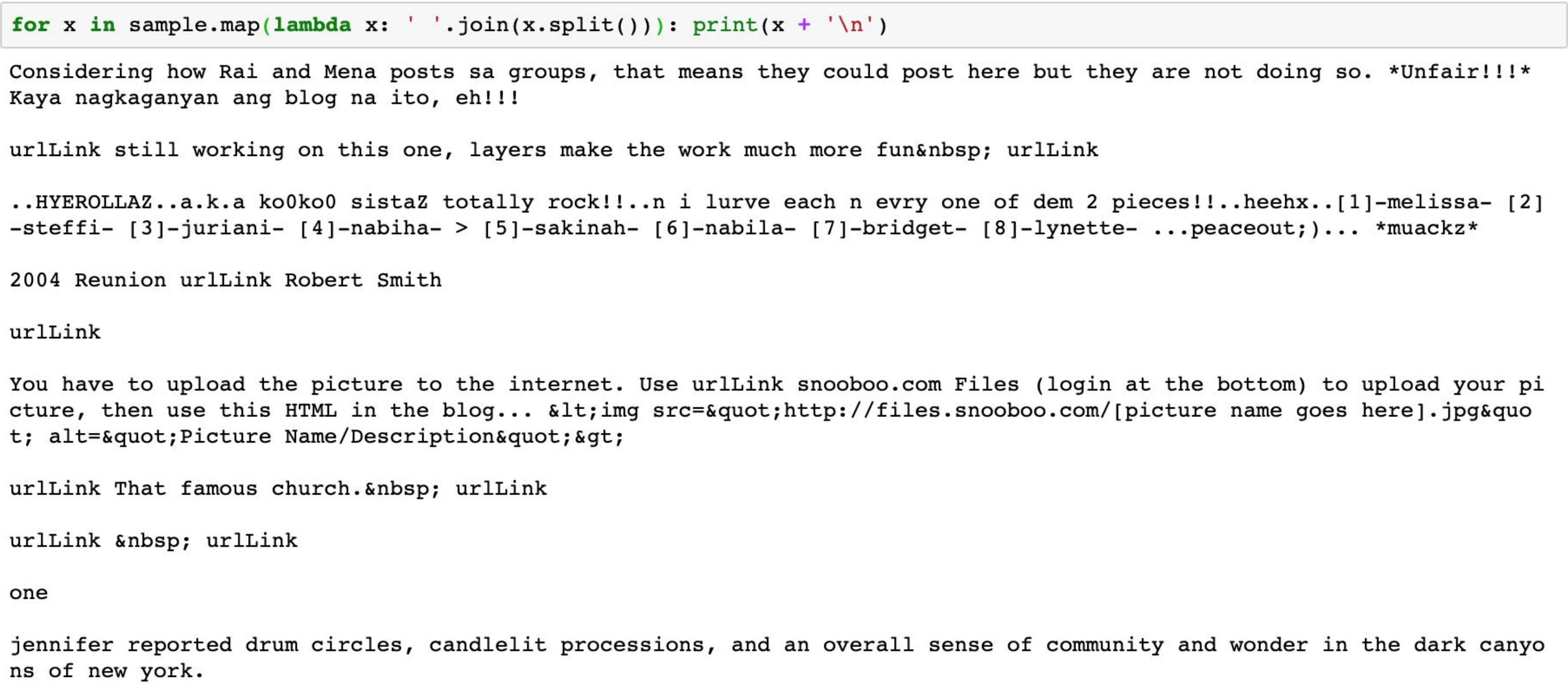

Distribution of polarity scores
140000
120000
100000
80000
60000
40000
20000
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
polarity_score



```
In [249]: sample = blogdata[blogdata.polarity_score == 0.00].sample(1).text

In [250]: for x in sample: print(x)

              The New Marine   urlLink

In [251]: sample = blogdata[blogdata.polarity_score == 0.00].sample(1).text

In [252]: for x in sample: print(x)

          NEW RADIOHEAD!! NEW RADIOHEAD!! NEW RADIOHEAD!! NEW RADIOHEAD!!
```

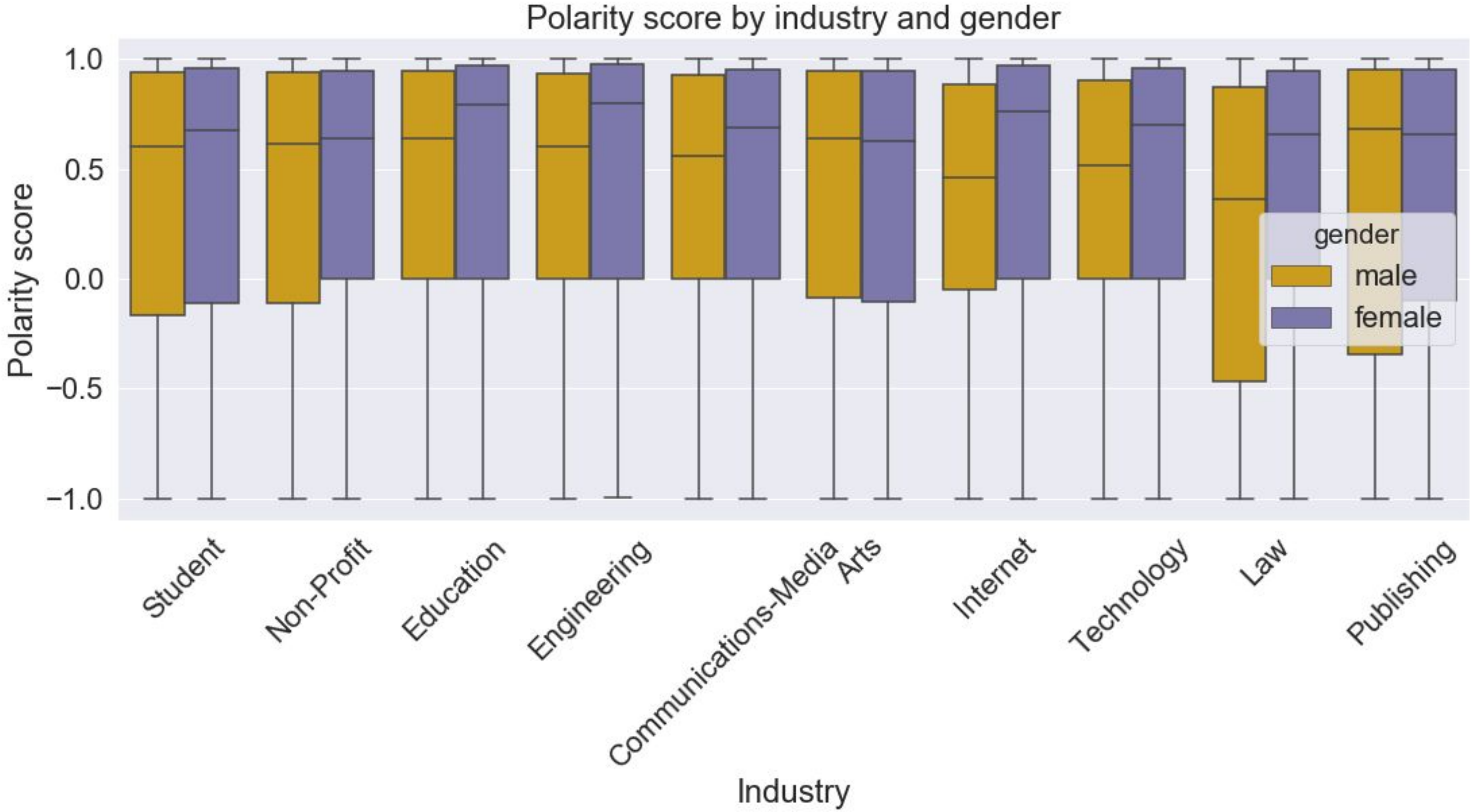


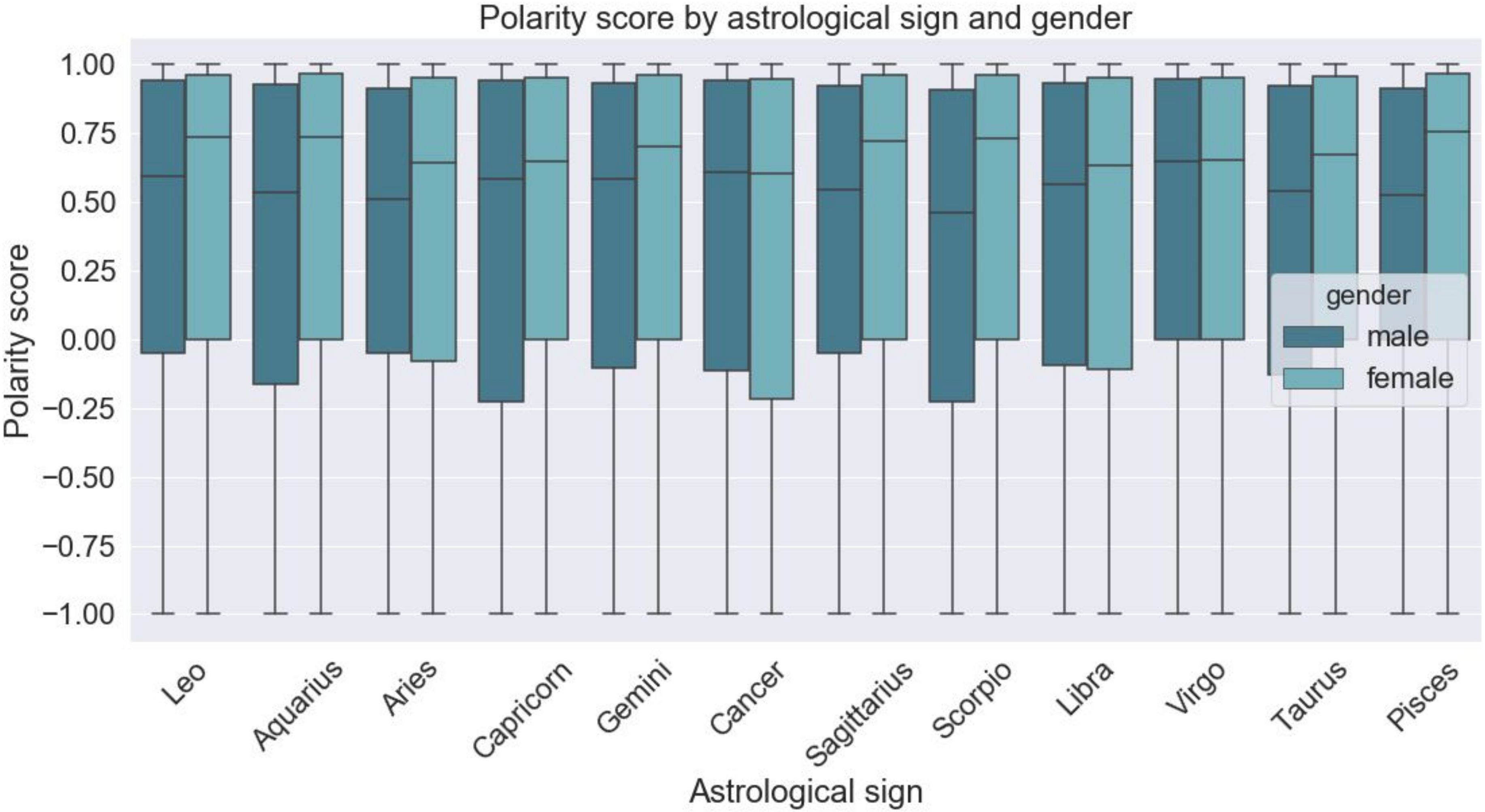
```
In [266]: blogdata[blogdata.polarity_score == 0.00].text.map(len).describe()

Out[266]: count    67535.000000
          mean       109.079662
          std        190.231915
          min          4.000000
          25%         49.000000
          50%         74.000000
          75%        119.000000
          max      13914.000000
          Name: text, dtype: float64
```



```python
for x in sample.map(lambda x: ' '.join(x.split())): print(x + '\n')
```

```
Considering how Rai and Mena posts sa groups, that means they could post here but they are not doing so. *Unfair!!!*
Kaya nagkaganyan ang blog na ito, eh!!!

urlLink still working on this one, layers make the work much more fun&nbsp; urlLink

..HYEROLLAZ..a.k.a ko0ko0 sistaZ totally rock!!..n i lurve each n evry one of dem 2 pieces!!..heehx..[1]-melissa- [2]
-steffi- [3]-juriani- [4]-nabiha- > [5]-sakinah- [6]-nabila- [7]-bridget- [8]-lynette- ...peaceout;)... *muackz*

2004 Reunion urlLink Robert Smith

urlLink

You have to upload the picture to the internet. Use urlLink snooboo.com Files (login at the bottom) to upload your pi
cture, then use this HTML in the blog... &lt;img src=&quot;http://files.snooboo.com/[picture name goes here].jpg&quo
t; alt=&quot;Picture Name/Description&quot;&gt;

urlLink That famous church.&nbsp; urlLink

urlLink &nbsp; urlLink

one

jennifer reported drum circles, candlelit processions, and an overall sense of community and wonder in the dark canyo
ns of new york.
```


Polarity score by industry and gender
Polarity score
1.0
0.5
0.0
−0.5
−1.0
Industry
Student
Non-Profit
Education
Engineering
Communications-Media
Arts
Internet
Technology
Law
Publishing
gender
male
female


Polarity score by astrological sign and gender
Polarity score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Astrological sign
Leo
Aquarius
Aries
Capricorn
Gemini
Cancer
Sagittarius
Scorpio
Libra
Virgo
Taurus
Pisces
gender
male
female

Brief aside: Mixed effects regression and R

- Looked for correlations between polarity score and demographic information
    - There weren’t any
- Fit a mixed effects model anyway
    - Results were meaningless and nothing was significant

Conclusions

- Difficult to get meaningful results from big datasets
    - Very basic self-reported information might not always be the best way to explore sociolinguistic variation
- Hope to further explore topic clustering and sentiment analysis