

News data

NAVON FRANCIS


The Question

How do we portray news?

The Data – Summary

- Newly created dataset, scraped from *euronews.com* and *npr.org*
- 110 articles chosen **i.i.d.** from both websites for 2017
- Articles then processed through Google's Natural Language API
- Performed sentiment analysis for each article
- Sentiment values range from -1 to 1, negative to positive emotion

sen·ti·ment

/ˈsen(t)əmənt/ 

noun

1. a view of or attitude toward a situation or event; an opinion.
"I agree with your sentiments regarding the road bridge"
synonyms: view, feeling, attitude, thought, opinion, belief
"the comments echo my own sentiments"
2. a feeling or emotion.
"an intense sentiment of horror"



Google Cloud Platform

The Data – API

- Google make's this really easy!
- This is the Natural Language API at work

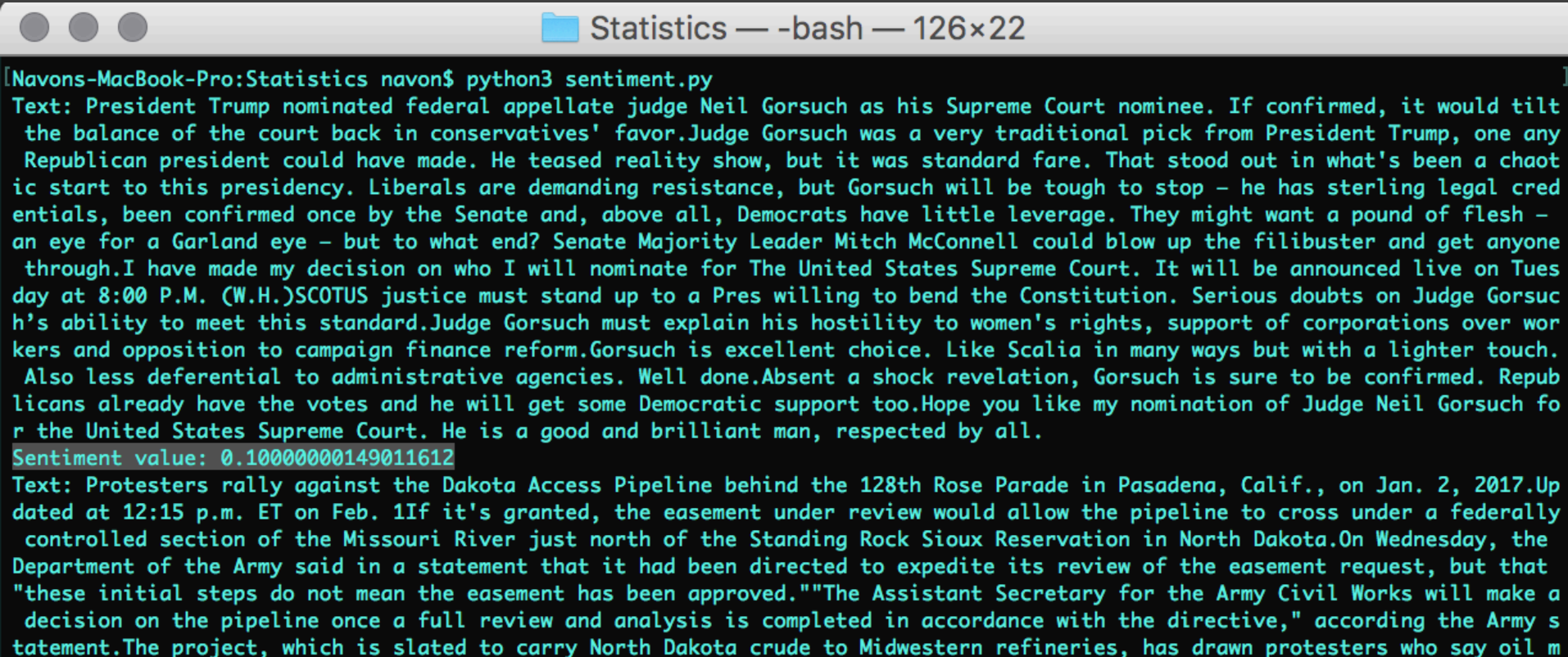
```
# For each article in the entire list of articles
for article in articles:
    # The text to analyze
    document = types.Document(content=article, type=enums.Document.Type.PLAIN_TEXT)

    # We send to the Google API to then analyze our text
    sentiment = client.analyze_sentiment(document=document).document_sentiment

    # Voilà!
    print('Text: {}'.format(article))
    print('Sentiment value: {}'.format(sentiment.score))
```

The Data – Results

- A portion of the dataset, US news



```
Statistics — -bash — 126x22
[Navons-MacBook-Pro:Statistics navon$ python3 sentiment.py
Text: President Trump nominated federal appellate judge Neil Gorsuch as his Supreme Court nominee. If confirmed, it would tilt the balance of the court back in conservatives' favor. Judge Gorsuch was a very traditional pick from President Trump, one any Republican president could have made. He teased reality show, but it was standard fare. That stood out in what's been a chaotic start to this presidency. Liberals are demanding resistance, but Gorsuch will be tough to stop – he has sterling legal credentials, been confirmed once by the Senate and, above all, Democrats have little leverage. They might want a pound of flesh – an eye for a Garland eye – but to what end? Senate Majority Leader Mitch McConnell could blow up the filibuster and get anyone through. I have made my decision on who I will nominate for The United States Supreme Court. It will be announced live on Tuesday at 8:00 P.M. (W.H.) SCOTUS justice must stand up to a Pres willing to bend the Constitution. Serious doubts on Judge Gorsuch's ability to meet this standard. Judge Gorsuch must explain his hostility to women's rights, support of corporations over workers and opposition to campaign finance reform. Gorsuch is excellent choice. Like Scalia in many ways but with a lighter touch. Also less deferential to administrative agencies. Well done. Absent a shock revelation, Gorsuch is sure to be confirmed. Republicans already have the votes and he will get some Democratic support too. Hope you like my nomination of Judge Neil Gorsuch for the United States Supreme Court. He is a good and brilliant man, respected by all.
Sentiment value: 0.10000000149011612
Text: Protesters rally against the Dakota Access Pipeline behind the 128th Rose Parade in Pasadena, Calif., on Jan. 2, 2017. Updated at 12:15 p.m. ET on Feb. 1. If it's granted, the easement under review would allow the pipeline to cross under a federally controlled section of the Missouri River just north of the Standing Rock Sioux Reservation in North Dakota. On Wednesday, the Department of the Army said in a statement that it had been directed to expedite its review of the easement request, but that "these initial steps do not mean the easement has been approved." "The Assistant Secretary for the Army Civil Works will make a decision on the pipeline once a full review and analysis is completed in accordance with the directive," according to the Army's statement. The project, which is slated to carry North Dakota crude to Midwestern refineries, has drawn protesters who say oil m
```

The Model – Choices

- What is the best model to use for this data?
- We obviously can not look at every article ever written
- We have two sample sets of values
- Maybe a t test, but which one..?
- One sample t test?
- Independent samples t test?
- Paired samples t test?

The Model – Verdict

- The Independent Samples t test
- We will test the difference between the two sample means
- Our dependent variable is from -1 to 1, our interval
- Our dependent variable is from the sample population
- We need mean M , variance S^2 , and sample size n of ₍₁₎ SVs – for t

Abbreviation:

1. SV: Sentiment value.

$$S^2 = \frac{\sum (x - M)^2}{n - 1}$$

$$t = \frac{M_x - M_y}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

The Analysis – Summary

- Using the Independent samples t test
- Analyze both means of sample sets (US and Euro)
- Two independent groups to be compare
- Data parameters: 110 ₍₁₎ SVs US articles & 110 SVs of Euro articles
- We will compare calculated t value to the critical t value

Abbreviation:

1. SV: Sentiment value.

The Analysis – Computation

- At this point, we should think about the question again
- We will apply the independent samples t test to the dataset
- “The mean emotion derived from EU news is equal to US news”
- Null Hypothesis ---> $H_0 : \mu_{\text{euro}} = \mu_{\text{usa}}$
- “The mean emotion derived from EU news is not equal to US news”
- Alt. Hypothesis ---> $H_a : \mu_{\text{euro}} \neq \mu_{\text{usa}}$

The Analysis – Computation

- Find the variance:

```
N = 110
euro_mean = np.mean(euro)
euro_var = sum([pow(e - euro_mean, 2) for e in euro]) / (N - 1)
```

$$S^2 = \frac{\sum (x - M)^2}{n - 1}$$

```
[Navons-MacBook-Pro:Statistics navon$ python3 ttest.py
euro variance: 0.10210925825771637
[Navons-MacBook-Pro:Statistics navon$ python3 ttest.py
usa variance: 0.04232276980110364
```

The Analysis – Computation

- Calculate t statistic:

```
t = (euro_mean - usa_mean) / np.sqrt((euro_var/N) + (usa_var/N))
```


$$t = \frac{M_x - M_y}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

```
[Navons-MacBook-Pro:Statistics navon$ python3 ttest.py  
t statistic: -2.1575971306161366
```

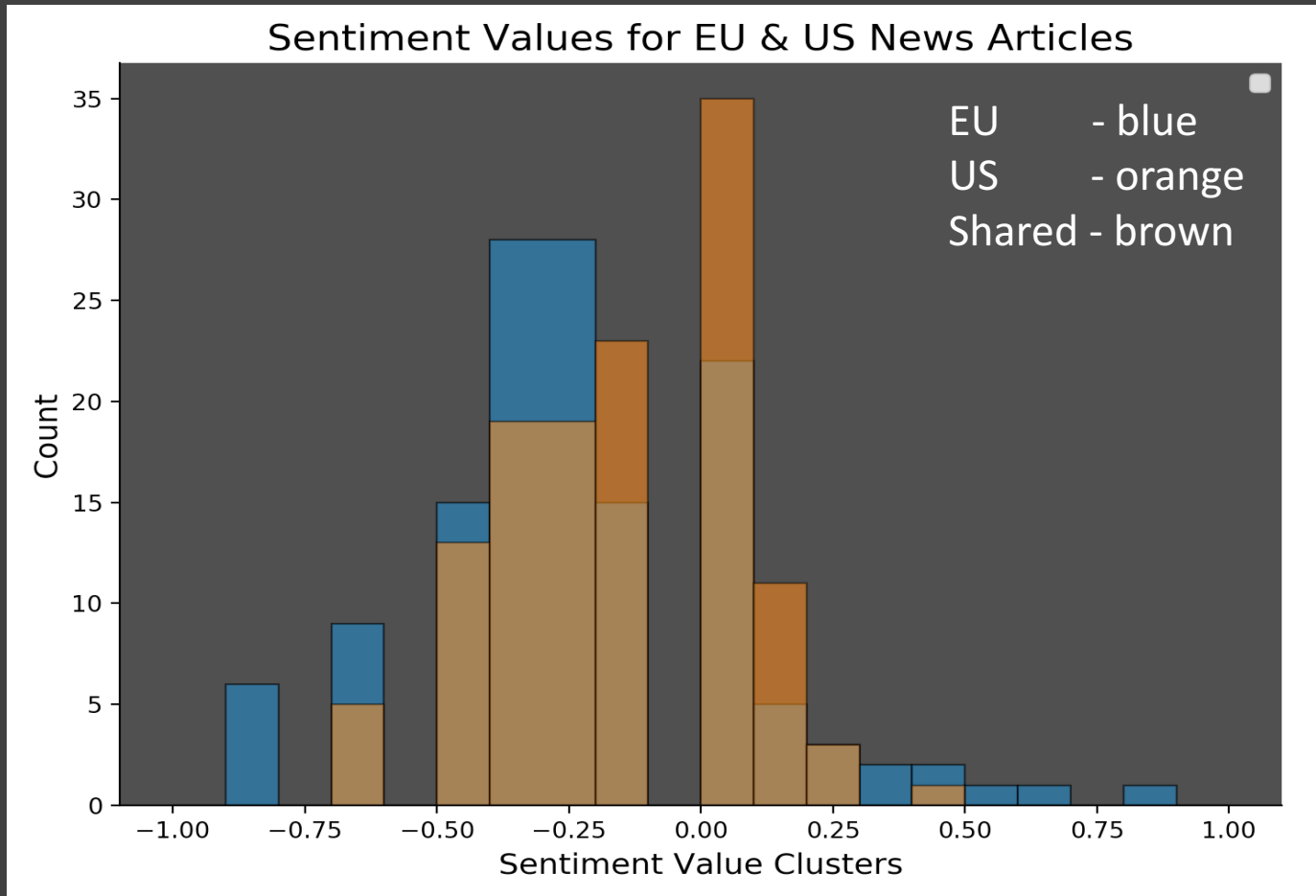
- Double check with scipy's independent t test function

```
[Navons-MacBook-Pro:Statistics navon$ python3 ttest.py  
t: Ttest_indResult(statistic=-2.1575971306161366, pvalue=0.032052561043647634)
```

The Results – Summary

- t statistic = 2.158
 - α = 0.05
 - df = 218
 - p val = 0.032
- 
- Critical t val = 1.984
 - Critical t (1.984) < calculated t (2.158)
 - p val (0.032) < α (0.05)
- Recall $H_0 : \mu_{\text{euro}} = \mu_{\text{usa}}$
 - The critical t value is less than the calculated t statistic, reject H_0 !
 - There is a statistical significance between the two population means

The Results – Histogram



The Conclusion

How do we portray news?

- In summation, there is a difference between the two samples
- Important because it provides some insight on how we portray things
- US articles seem to have more 0's, which means more mixed emotion
- EU articles have more negative sentiment, can be more serious
- This could lead to more discoveries in biased news
- Further analysis to identify factual news based on this can be made

Thank you
