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

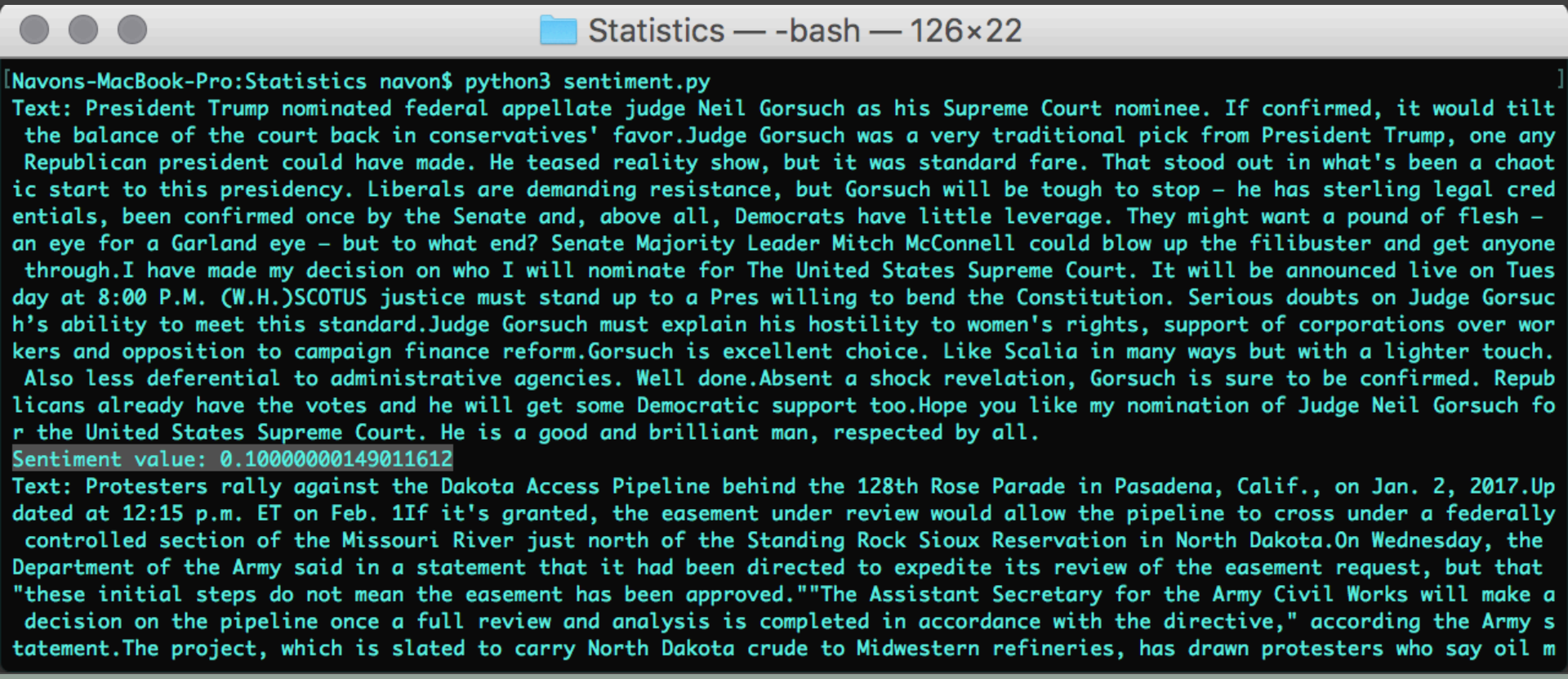
```
# For each article in the entire list of articles
for article in articles:
    # The text to analyze
    document = types.Document(content=article, type=enums.Document.Type.PLAIN_TEXT)

    # 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



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

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

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

Abbreviation:
1. SV: Sentiment value.

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 ---> $H_0 : \mu_{euro} = \mu_{usa}$
- “The mean emotion derived from EU news is not equal to US news”
- Alt. Hypothesis ---> $H_a : \mu_{euro} \neq \mu_{usa}$

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=0.03205256104364763 4)
```

The Results – Summary

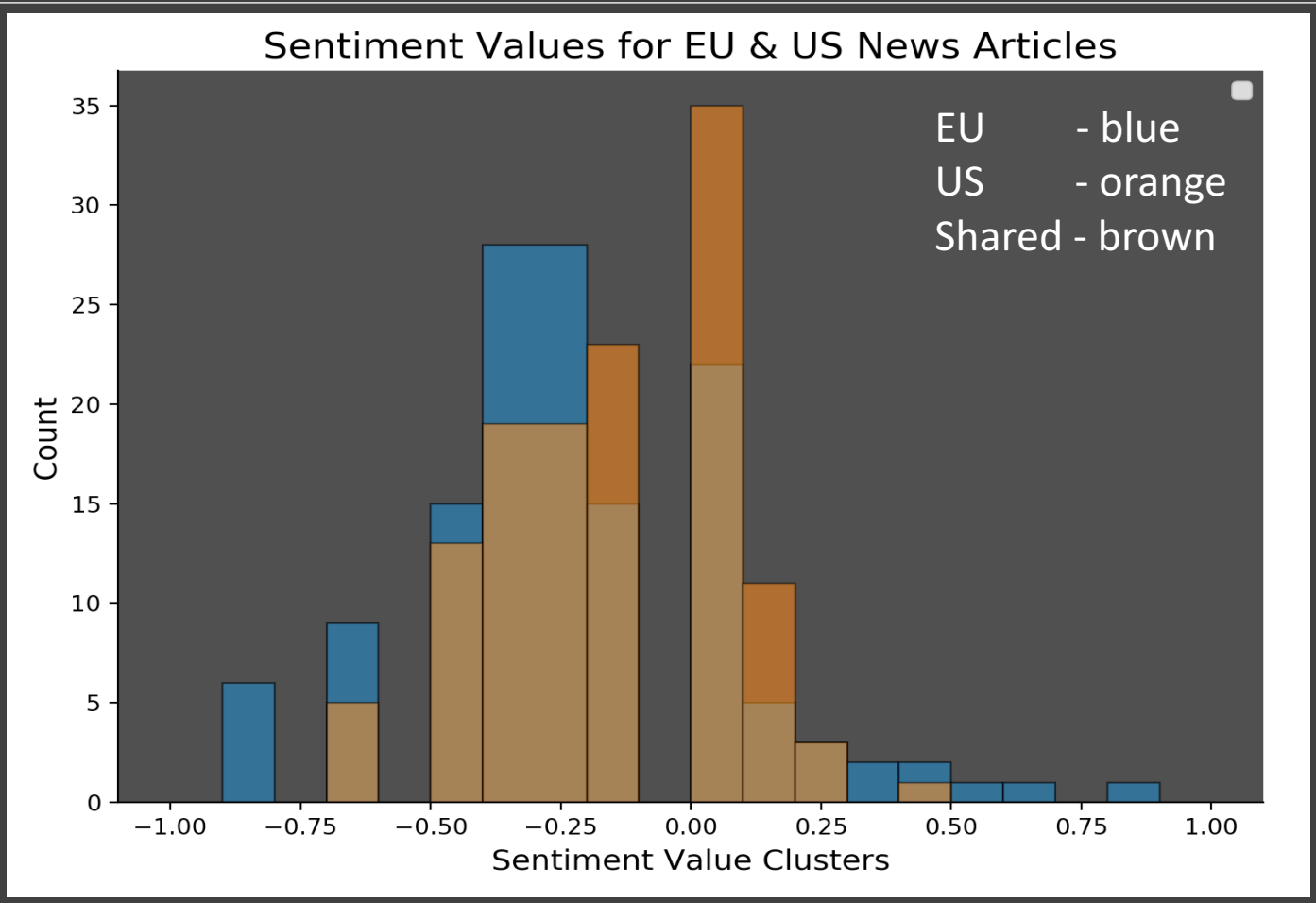
- t statistic = 2.158
- α = 0.05
- df = 218
- p val = 0.032

- Critical t val = 1.984

- Critical t (1.984) < calculated t (2.158)
- p val (0.032) < α (0.05)

- Recall $H_0 : \mu_{euro} = \mu_{usa}$
- 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