A Primer on Textual Analysis

MAS – 2022 – Doctoral Consortium

Ties de Kok
University of Washington

UNIVERSITY *of* WASHINGTON

FOSTER SCHOOL OF BUSINESS

Session objectives

Primary objective:

Enable you to incorporate textual data into your projects

What I will not do:

- Focus on technical and mathematical details
- Throw buzz-words at you for 1.5 hours
- Provide you with a comprehensive literature review

My background

I specialize in combining computer science techniques with empirical accounting research

Disclaimer:

> My primary field of research is financial accounting, but I tried to make this session broadly applicable!

Session framework

1. **What** is textual analysis?

2. **Why/When** use textual analysis?

3. **How** to perform textual analysis

   → **Hands-on portion with Python**

Sneak peak – hands on portion

We will analyze about 2,300 employee reviews from Meta (i.e., Facebook) using a Python notebook.

We'll do some pretty cool things!

> Keyword analysis
> Sentiment analysis
> Topic modeling using LDA

Let's get some details out of the way first!

What is textual analysis?

Simple → the analysis of textual data

Similar inter-related names:

> Computational Linguistics

> Text Mining

> Natural Language Processing

Examples of textual data

- Contracts
- Financial disclosures
- Compensation
- Regulations
- ESG disclosures
- Company websites
- Social media posts
- Audio transcripts
- Reviews
- Job postings
- News articles
- Survey responses
- Inspection reports
- Policy documents
- Interviews

Why? Textual data is everywhere!

When does it make sense to use textual data?

Problem: textual data is often difficult to work with…

Why difficult?

- A lot of variation
- Difficult to objectively evaluate the “message”
  - Nuance is important
- Large file sizes
  - Requires sufficient storage & compute power

Generally → avoid textual data if possible

When does it make sense to use textual data?

Do the benefits outweigh the costs?

Textual data is worth analyzing if:

1. The text is reasonably available and processable
2. The text contains information that:
   - You care about
   - Isn’t available otherwise

Primary approaches

You have found some textual data that you want to analyze, **now what?**

Options:

1. Perform manual analysis
2. Use a plug-and-play textual analysis tool
3. Set up an “end-to-end” approach

   → i.e. programming

Examples of plug-and-play tools

Microsoft Word

→ Readability

Grammarly

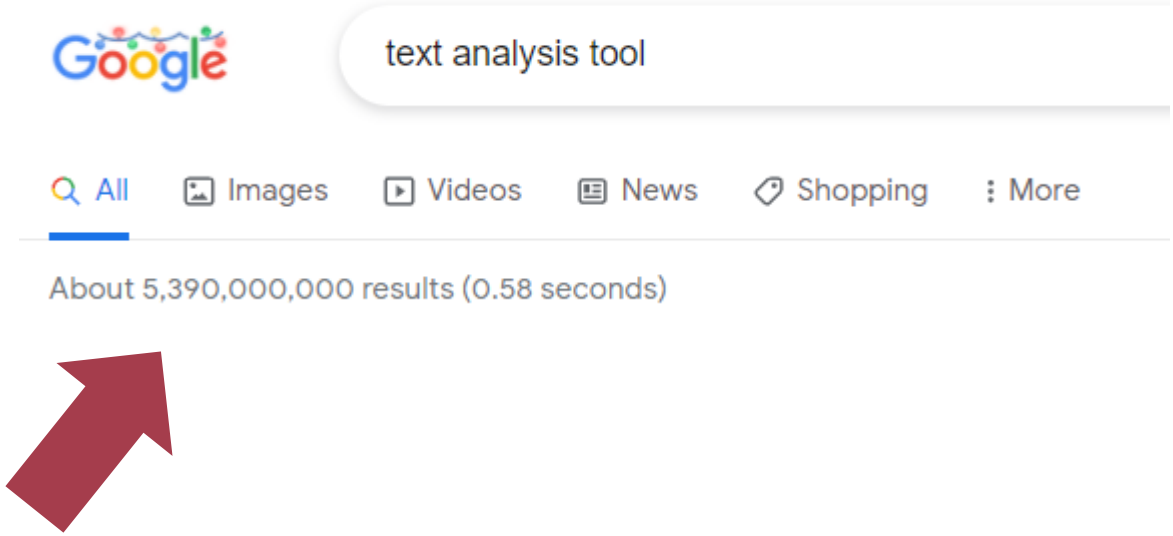
→ Readability

→ Emotion

Cloud services

e.g., Azure

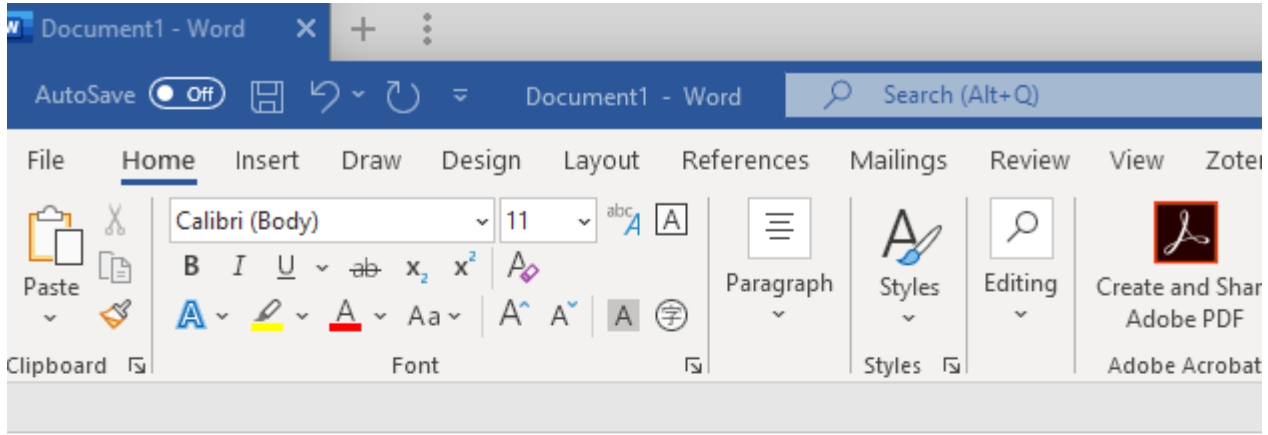
There is no lack of existing tools:



Examples of plug-and-play tools

Microsoft Word

→ Readability



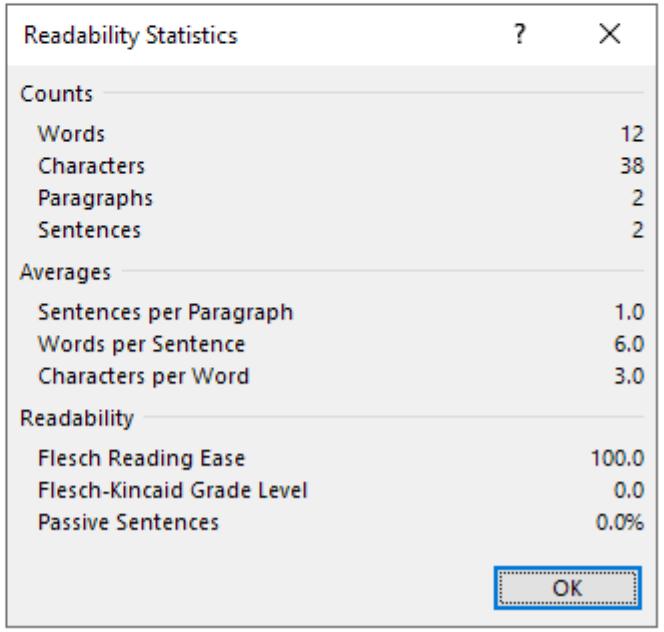
Grammarly

→ Readability

→ Emotion

This is a piece of text.

This is a piece of text.



Cloud services

e.g., Azure

Examples of plug-and-play tools

Grammarly

→ Readability

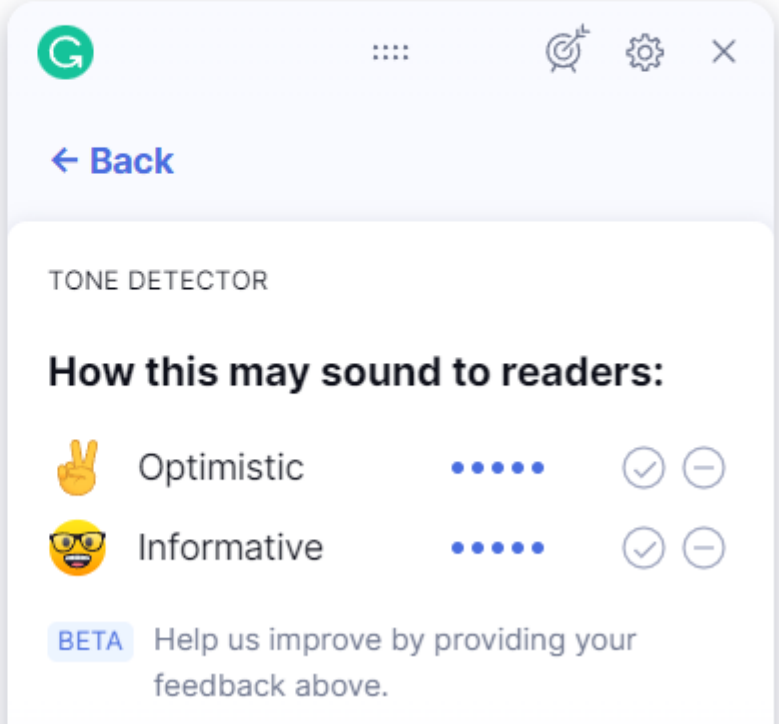
→ Emotion

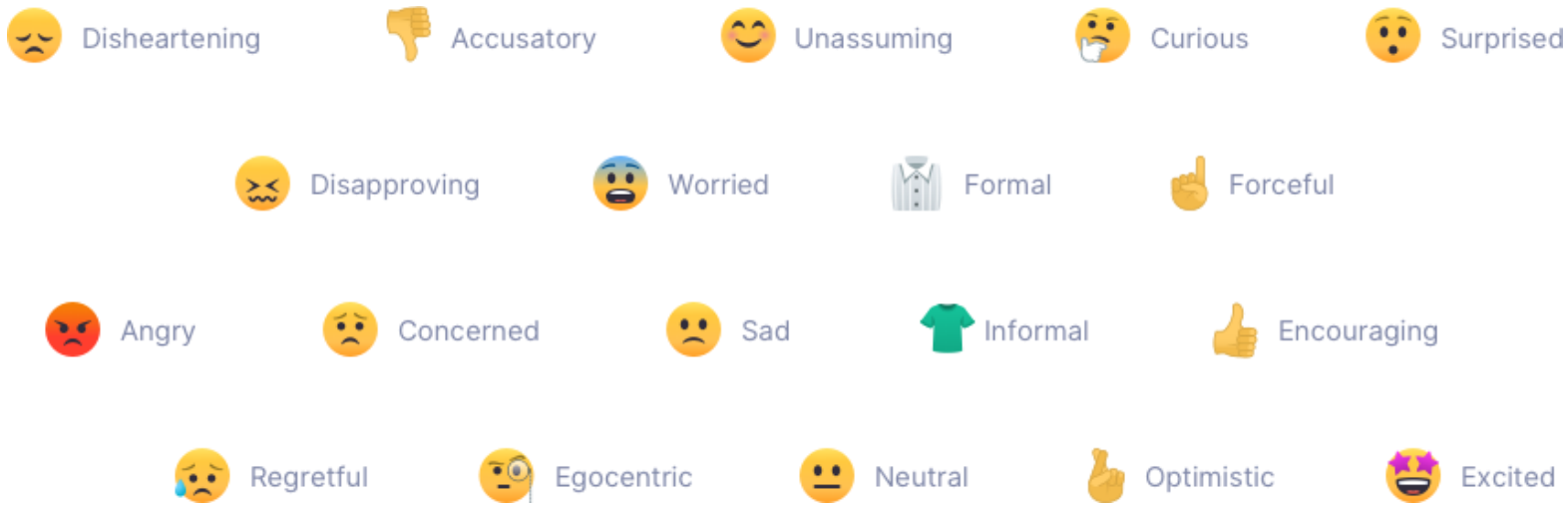
Dear PhD student,

Correct my writing; I expect it on my desk by tomorrow morning. Make sure to include an analysis of the emotion in the document, although you will probably forget.

Please don't disappoint me.

Your supervisor





Examples of plug-and-play tools

Microsoft Word

→ Readability

Azure

Text analytics

A collection of features from Cognitive Service for Language that extract, classify, and understand text within documents

Grammarly

→ Readability

→ Emotion

Broad entity recognition

Identify important concepts in text, including key phrases and named entities such as people, events, and organizations.

Powerful sentiment analysis

Examine what customers are saying about your brand and analyze sentiments around specific topics through opinion mining.

Document summarization

Extract sentences that collectively convey the essence of a document.

Cloud services

e.g., Azure

End-to-end textual analysis

An end-to-end project consists of:

1. Understanding the textual data
2. Obtaining the textual data
3. Cleaning the textual data
4. Analyzing the textual data

My tool of choice!

This can be time consuming, but the sky is the limit!

Let's get ourselves set up!

You don't need to install anything, just go to:

https://github.com/TiesdeKok/MAS_2022_textual_analysis

MAS 2022 session on Textual Analysis

Instructor: Ties de Kok

Binder link:

To get started with the demonstration portion, click the button below!

launch binder

Let's work through an example!

Hypothetical research objective:

Extract insights(s) from of employee reviews

Pros

Facebook deeply cares about its employees and has built a compelling culture around support and growth. Career growth opportunities are plentiful. If you don't like the team you're on or don't get the support you want from your manger, Facebook empowers you to find new teams or projects. Facebook wants its employees to be invested in their work and to feel connected to its larger mission. If large scale opportunities and growth are important to you, Facebook is a fantastic place to work.

Cons

Facebook's culture is demanding and fast paced. The greatest aspect of working at Facebook is that everyone is very motivated and very smart. The problem with this is that they all expect the very same of you. Holding a very high bar for excellence can certainly be demanding so it's important to make sure you're always carefully paying attention to your own personal work/life balance.

The data

5.0 ★★★★★

Current Employee, more than 3 years

People Focused

May 24, 2020 - Engineering Manager in San Francisco, CA

✓ Recommend ✓ CEO Approval ✓ Business Outlook

Pros

Facebook deeply cares about its employees and has built a compelling culture around support and growth. Career growth opportunities are plentiful. If you don't like the team you're on or don't get the support you want from your manger, Facebook empowers you to find new teams or projects. Facebook wants its employees to be invested in their work and to feel connected to its larger mission. If large scale opportunities and growth are important to you, Facebook is a fantastic place to work.

Cons

Facebook's culture is demanding and fast paced. The greatest aspect of working at Facebook is that everyone is very motivated and very smart. The problem with this is that they all expect the very same of you. Holding a very high bar for excellence can certainly be demanding so it's important to make sure you're always carefully paying attention to your own personal work/life balance.

Demo dataset: about 2,300 reviews

glassdoor | Meta | Location

Jobs | Companies | Salaries | Careers

Meta

Overview | Reviews 9K | Jobs 18 | Salaries 32K | Interviews 9K | Benefits 3K

Meta Reviews Updated Dec 30, 2021

Search job titles | Find Reviews

Clear All | Full-time, Part-time ✕ | English | Filter

Found 4,587 of over 9,202 reviews | Sort: Popular

4.3 ★★★★½

85% Recommend to a Friend | 87% Approve of CEO | Mark Zuckerberg 2,223 Ratings

Your trust is our top concern, so companies can't alter or remove reviews.

Pros

"You get to work with smart people from around the world" (in 447 reviews)

"awesome job with great benefits" (in 378 reviews)

Cons

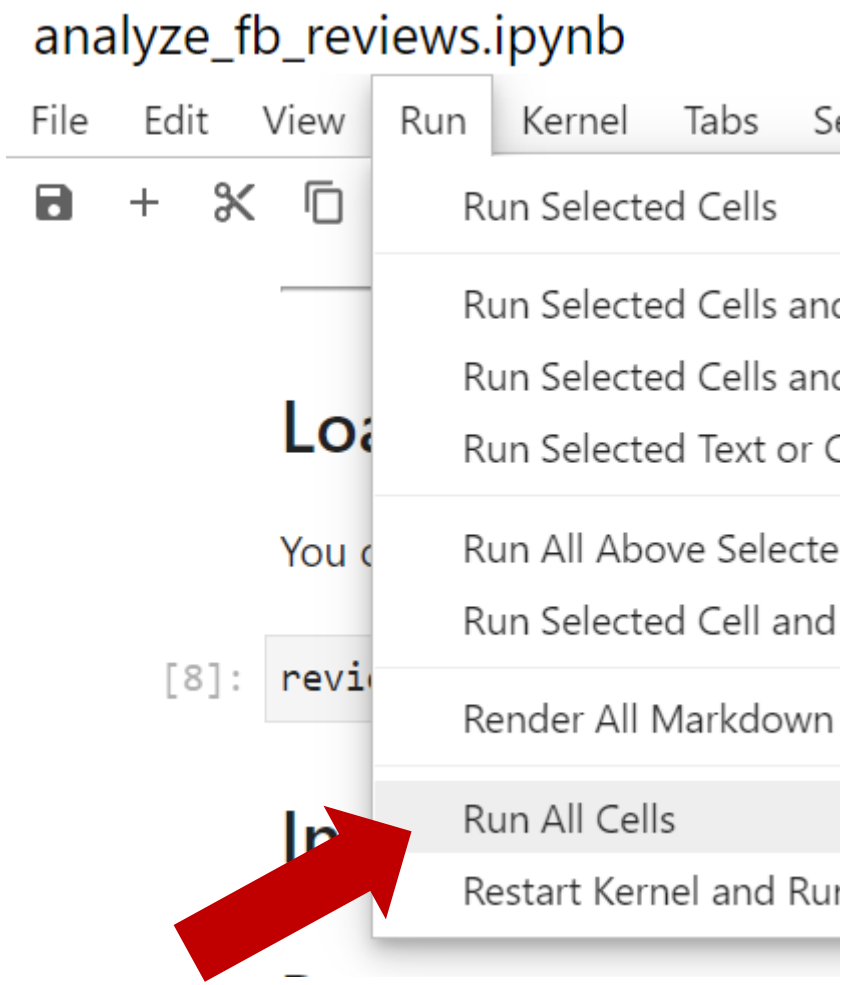
"The long hours are sometimes tedious" (in 153 reviews)

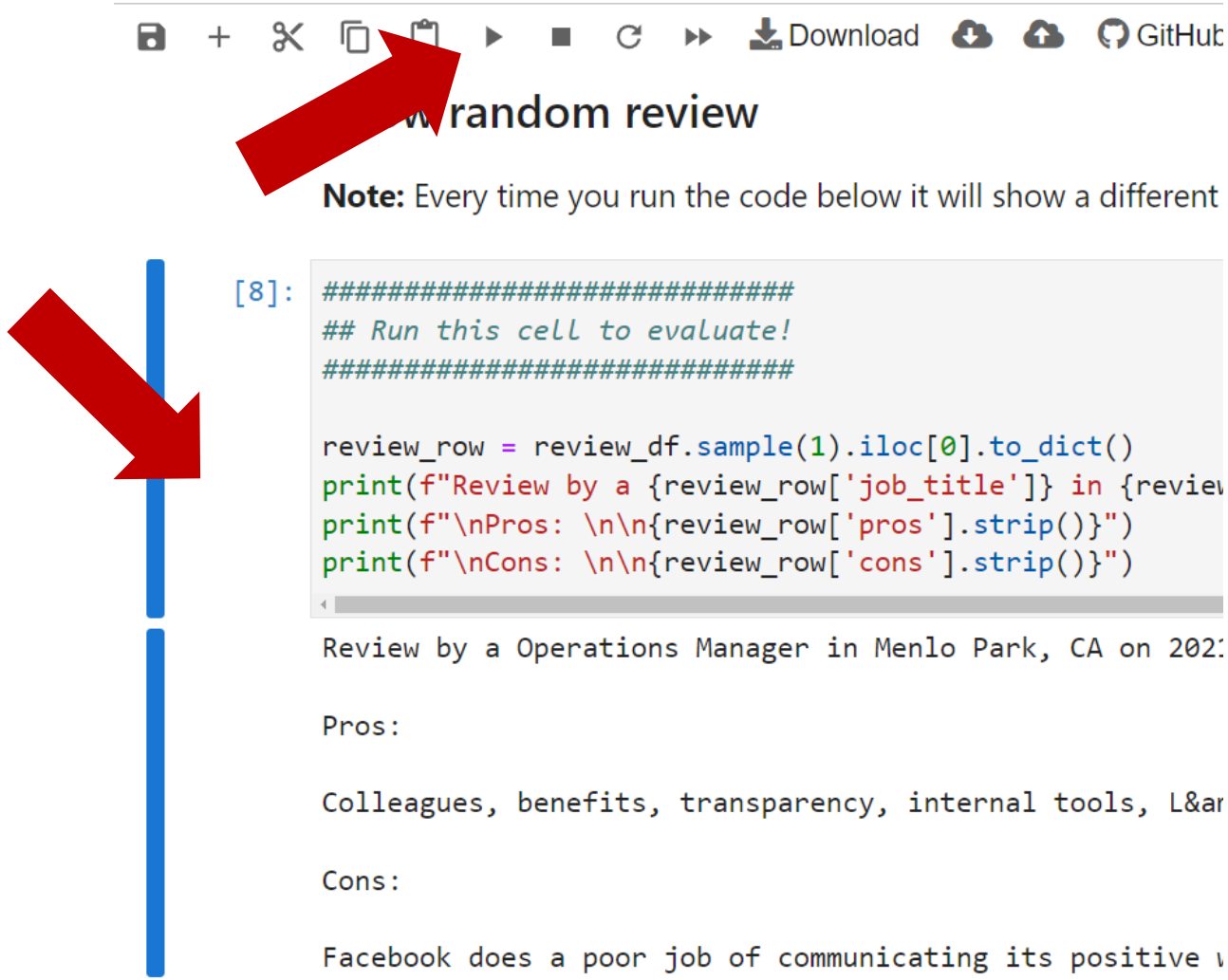
"Some pains from rapid growth from a small company to a big company (e" (in 89 reviews)

More Pros and Cons

Pros & Cons are excerpts from user reviews. They are not authored by Glassdoor.

How to use the notebook





Task #1 – Understand the data

Your tasks:

1) Come up with three constructs that you could extract from the reviews

2) What problems do you foresee when looking through the texts?

Download GitHub

random review

Note: Every time you run the code below it will show a different

[8]:
```
############################
## Run this cell to evaluate!
############################

review_row = review_df.sample(1).iloc[0].to_dict()
print(f"Review by a {review_row['job_title']} in {revie
print(f"\nPros: \n\n{review_row['pros'].strip()}")
print(f"\nCons: \n\n{review_row['cons'].strip()}")
```

```
Review by a Operations Manager in Menlo Park, CA on 202

Pros:

Colleagues, benefits, transparency, internal tools, L&a

Cons:

Facebook does a poor job of communicating its positive
```

Constructs & Challenges

Constructs:

Challenges:

An aside #1 → How to get such data?

Ways to obtain Textual data:

- Download from archive
- Hand collection
- Automatic collection:
  - API access
  - Web scraping

Types of files:

- Text (*.txt)*
- PDF (.pdf)
- Web pages (*.html)*
- Machine readable formats (e.g., *.json*)
- Proprietary files

An aside #2 → Web scraping

Too much to cover in one session, interested?

Check out my Python course:

https://github.com/TiesdeKok/limperg_python

➔ Contains a session on web scraping

Task #2 – Resolving data issues

Text is unstructured and messy; data cleaning is very important!

Common cleaning steps:

- Remove invalid characters
- Remove excess whitespaces
- Extract words and sentences
- Filter text on language
- Normalize text to be more comparable

Task #2 – cont.

Your tasks:

1) Compare the “dirty” vs. the “clean” text and find the differences.

2) Can you still find remaining issues?

[12]:
```
############################
## Run this cell to evaluate!
############################

review_row = review_df.sample(1).iloc[0].to_dict()
print(f"Review by a {review_row['job_title']} in {review
print(f"\nPros: \n\n{review_row['pros'].strip()}")
print(f"\nPros clean: \n\n{review_row['pros_clean'].stri
print('\n' + '-'*50)
print(f"\nCons: \n\n{review_row['cons'].strip()}")
print(f"\nCons clean: \n\n{review_row['cons_clean'].stri
```

```
Review by a Information Security Specialist in Menlo Par

Pros:

* Great colleagues_x000D_
* Great perks/benefits_x000D_
* Interesting projects_x000D_
* Ability to to move laterally

Pros clean:

Great colleaguesxD. Great perksbenefitsxD. Interesting p
rally.
```

Special mention – Regular expressions

How do we get all the employee IDs?

*The employees involved with the inventory counting procedure are Yvonne Olsen (**EMP-ID-0001**), Yan Han (**EMP-ID-1012**), Christin Mayer (**EMP-ID-2378**), and Ezra Rosales Fuentes (**EMP-ID-5203**).*

REGULAR EXPRESSION 4 ma

```
/ (EMP-ID-\d\d\d\d)
```

```
The employees involved with the
inventory counting procedure are
Yvonne Olsen (EMP-ID-0001), Yan
Han (EMP-ID-1012), Christin Mayer
(EMP-ID-2378), and Ezra Rosales
Fuentes (EMP-ID-5203).
```

Special mention – Regular expressions

regexr.com

Untitled Pattern Save (ctrl-s)

Expression

/(\d+[,\.]*\d*)/g

(**Capturing group #1.** Groups multiple tokens together and creates a capture group for extracting a substring or using a backreference.

- \d **Digit.** Matches any digit character (0-9).
  - + **Quantifier.** Match 1 or more of the preceding token.
- [**Character set.** Match any character in the set.
  - , **Character.** Matches a "," character (char code 44).
  - \. **Escaped character.** Matches a "." character (char code 46).
  ]
  - * **Quantifier.** Match 0 or more of the preceding token.
- \d **Digit.** Matches any digit character (0-9).
  - * **Quantifier.** Match 0 or more of the preceding token.

)

Task #3 – Analyzing the data

Analyzing the data
is where the fun begins!

We will cover:

- Keyword analysis
- Sentiment analysis
- Topic modeling

Process + Clean text

Represent text numerically
- Bag of Words
- Word Embeddings

Statistical Models
- "Traditional" machine learning
  - Supervised → Classification (E.G. Naïve Bayes, SVM)
  - Unsupervised → Dimensionality reduction (E.G. LDA)
- Neural Networks (not covered)

Direct feature extraction
- Feature Search
  - E.G.
    - Entity recognition
    - Pattern search
- Text evaluation
  - E.G.
    - Dictionary counting
    - Readability

Task #3a – Keyword Extraction

What is it? → Count the number of times a keyword occurs

Simple principle but has many use cases!

Your tasks:

1) Check the frequency of the provided keywords
2) Try some of your own keywords

Any interesting patterns you can find?

Check whether each pros and cons contains a keyword

```
keywords_of_interest = ['traffic', 'culture',
                        'salary', 'diversity',
                        'balance']
##                     ^^^^^^^^^^^^^^ <-- Change this!

text_types = ['pros', 'cons']
contains_columns = []
for text_type in text_types:
    for keyword in keywords_of_interest:
        review_df[f'{text_type}_contains_{keyword}'] = re
        contains_columns.append(f'{text_type}_contains_{k
```

Task #3b – Sentiment analysis

What is it? → A score that indicates whether a text reflects a positive, neutral, or negative sentiment.

Your tasks:

1) Evaluate whether the sentiment scores make sense.

-1 is most negative and +1 is most positive

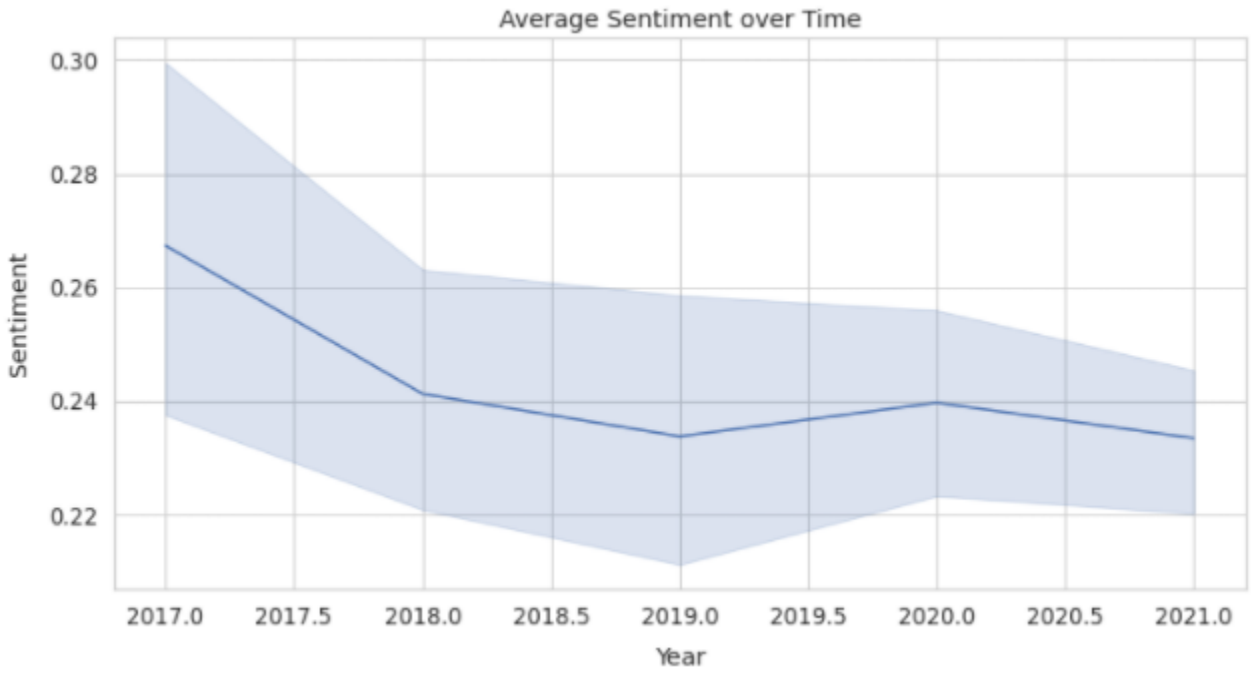
Inspect sentiment

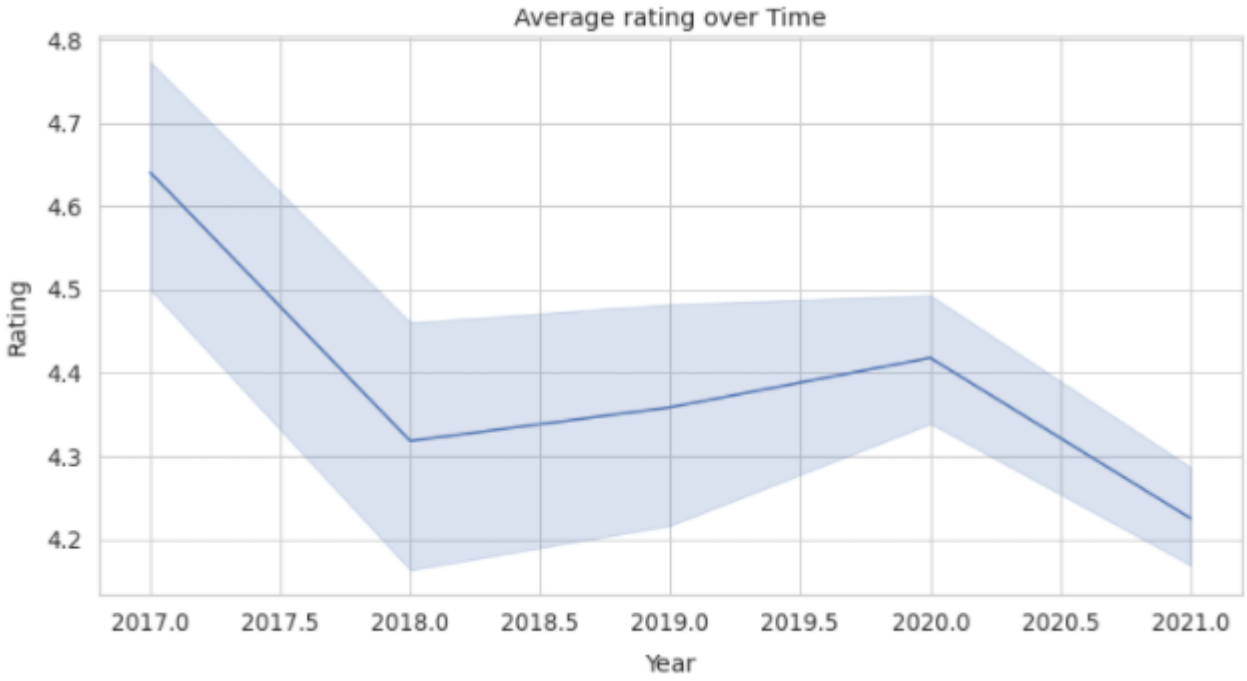
```
############################
## Run this cell to evaluate!
############################

review_row = review_df.sample(1).i
print(f"Review by a {review_row['j
print(f"\nPros clean (sentiment: {
print('\n' + '-'*50)
print(f"\nCons clean (sentiment: {
```

```
Review by a Product Manager in Mer
0 00:00:00

Pros clean (sentiment: 0.57):
```

Task #3b – Sentiment analysis

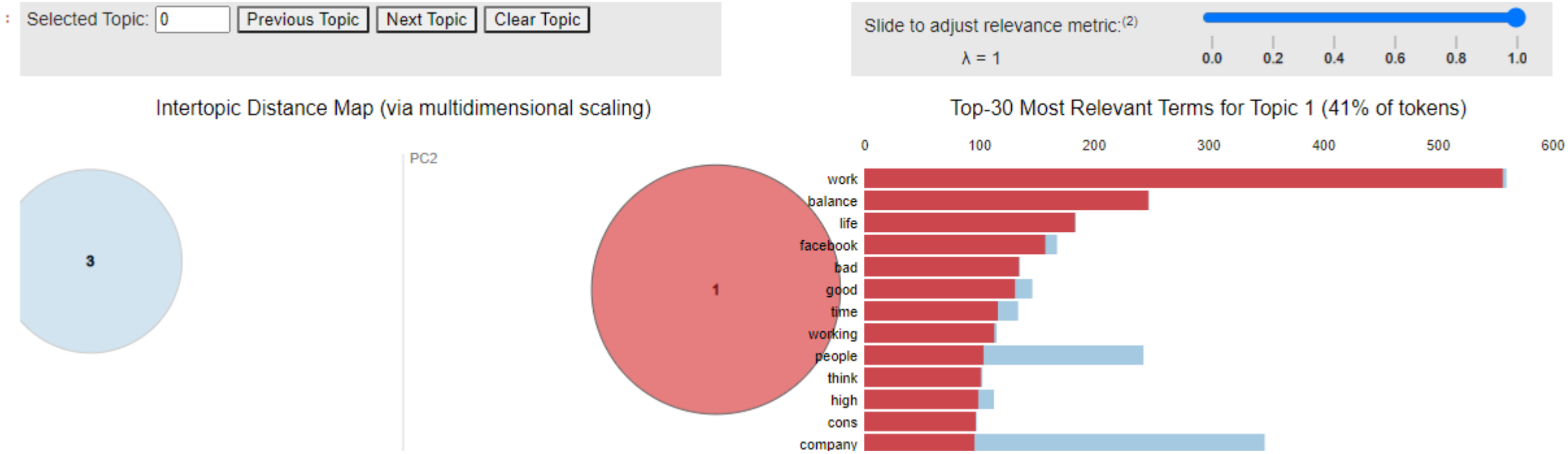
Does the text sentiment vary predictably with the rating score?





Task #3c – Topic modeling

What is it? → Attempt to identify clusters of text that are like each other but separate to other text clusters



Conclusion & next steps

I hope this session and demonstration gave you a taste of what you can do with textual data!

Want to learn more (specifically *end-to-end*)?

> *Learn Python for Research*

> *Python NLP tutorial*

> Limperg Python course recordings

Finally, my inbox is always open!

➔ tdekok@uw.edu

Thank you!

UNIVERSITY of WASHINGTON

FOSTER SCHOOL OF BUSINESS