Textual analysis using Python

Session 1

Ties de Kok
University of Washington

UNIVERSITY of WASHINGTON

FOSTER SCHOOL OF BUSINESS

Workshop objectives

Primary objective:

Help you incorporate textual data into your projects

What I will not do:

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

UNIVERSITY of WASHINGTON

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!

UNIVERSITY of WASHINGTON

Workshop overview

> **Session 1:** foundational methods
  - "Current status quo"

> **Session 2:** state-of-the-art
  - "future" (?) → BERT & GPT

UNIVERSITY of WASHINGTON

Session #1 overview

> **Introduction to textual analysis**
  - What? Why? When?

> **Tour of fundamental methods (*how*):**
  - Text cleaning
  - Keyword counting
  - Regular expressions
  - Supervised machine learning
  - Unsupervised machine learning

UNIVERSITY of WASHINGTON

What is textual analysis?

Simple → the analysis of textual data

Similar inter-related names:

> Computational Linguistics

> Natural Language Processing

> Text Mining

UNIVERSITY of WASHINGTON

Why care about textual analysis?

- 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 to use textual analysis?

Problem: textual data is difficult to work with…

Why difficult?

- Lots of variations
- Reflects complex information
- Harder for computers to process
  - Requires sufficient storage & compute power

Do the benefits outweigh the cost?

→ If yes, use textual data

How to analyze text?

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

Options:

1. Perform manual analysis
2. Use a plug-and-play textual analysis tool
3. Code up your own program

   → **Our focus, using Python!**

UNIVERSITY of WASHINGTON

Textual analysis program

A textual analysis project consists of these steps:

1. Understand the textual data
2. Obtain the textual data
3. Clean the textual data
4. **Analyze** the textual data

My tool of choice!

The sky is the limit!

Aside #1 → How to get textual data?

Sources:

- Data provider
- 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

Aside #2 → Web scraping

Too much to cover here, interested?

Check out my Python course:

https://github.com/TiesdeKok/limperg_python

➔ Contains a session on web scraping

Case introduction

UNIVERSITY of WASHINGTON

FOSTER SCHOOL OF BUSINESS

Case: Glassdoor reviews

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

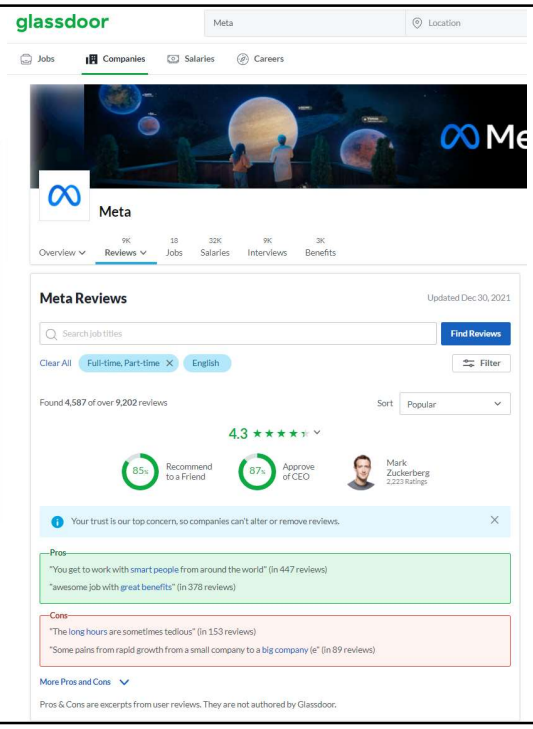
Pros

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.

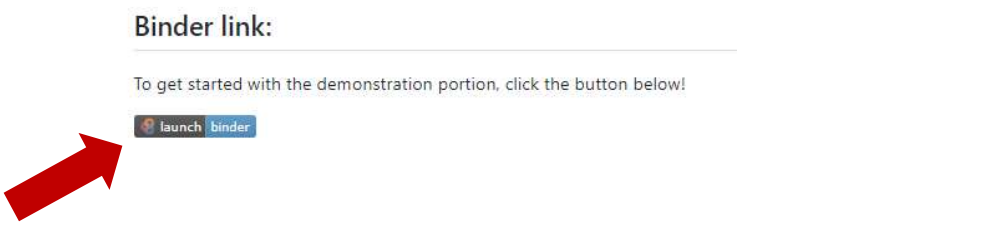
Demo dataset:
about 2,300 reviews



Notebook

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

https://github.com/TiesdeKok/eaa_2023_nlp_workshop



UNIVERSITY of WASHINGTON

Discussion point #1:

Cleaning data

UNIVERSITY of WASHINGTON

FOSTER SCHOOL OF BUSINESS

Discussion #1: cleaning data

Objective: get data clean enough that it won't mess up our analysis.

Discussion questions:

> *What types of data quality issues can you find with the text of these reviews?*

> *Anything we should be concerned about before analyzing the reviews?*

UNIVERSITY of WASHINGTON

slido

What data concerns do you have?

ⓘ Start presenting to display the poll results on this slide.

Cleaning data → how?

Usually a combination of three methods:

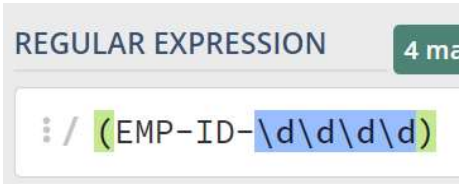
- **Python string operations**
  - E.g., remove whitespace
- **Python libraries**
  - Spacy, Textblob, NLTK
- **Regular expressions**
  - E.g., extract or replace specific patterns

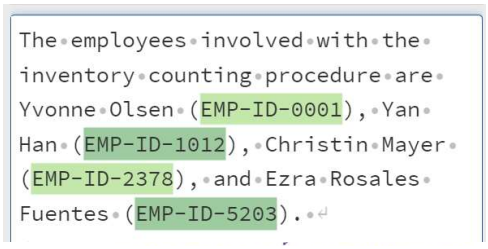
UNIVERSITY of WASHINGTON

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**).*





Special mention – Regular expressions

Helpful resources:

- **Regexr.com & Pythex.org**
  - → help to test, debug, and understand regular expressions
- **Using Python for Text Analysis in Accounting Research** *(Anand et al., 2020 – Foundations)*
  - → includes an entire chapter on Regular Expressions
- **ChatGPT**
  - → Regex is less hard to machines

Pre – text cleaning

> Let's check the notebook!

UNIVERSITY of WASHINGTON

Analyzing textual data

UNIVERSITY of WASHINGTON

FOSTER SCHOOL OF BUSINESS

Analysis – Foundational methods

> **Keyword counts**

> **Classification (e.g., sentiment)**
- Rule-based
- Supervised machine learning

> **Clustering (e.g., topic modeling)**
- Rule-based
- Unsupervised machine learning

UNIVERSITY of WASHINGTON

#1 – Keyword counts

Objective: count frequency of words of interest

Hardest part? → getting the keyword list

Task: how much does a review talk about compensation

> *What keywords would you add to your list to quantify how much a review discusses employee compensation?*

slido

What are your compensation keywords?

ⓘ Start presenting to display the poll results on this slide.

#1 – Keyword counts

> Let's check the notebook!

UNIVERSITY of WASHINGTON

#2 – Classification: rule-based

Objective: classify documents on some dimension

Hardest part? → getting a good keyword list

Example: identify reviews that discuss compensation

Rule-based:

```
if num_compension_words > 0:
    compensation = 1
else:
    compensation = 0
```

#2 – Classification: supervised ML

Objective: classify documents on some dimension

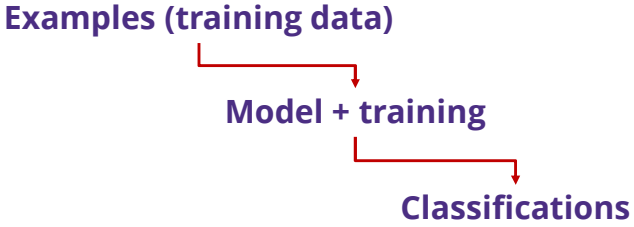
Hardest parts:

(1) *Getting good/enough examples*

(2) *Representing text numerically*

Example: identify reviews that discuss compensation

Steps:



slido

What types of dimension would be interesting to classify?

ⓘ Start presenting to display the poll results on this slide.

#2 – Representing text numerically

Problem: statistical models require numbers as input

Text =/= numbers

Necessary to represent text in numerical form:

- Term Frequency (or TF-IDF)
- Embeddings (e.g., word2vec)
- BERT

➔ **More on this next session**

#2 – Classification

> **Let's check the notebook!**

UNIVERSITY of WASHINGTON

#3 – Clustering: rule-based

Objective: cluster similar documents together

Hardest part? → getting good keyword lists

Example: create a keyword list for:

> **Compensation**
> **Culture**
> **Management**
> **Etc.**

#3 – Clustering: unsupervised ML

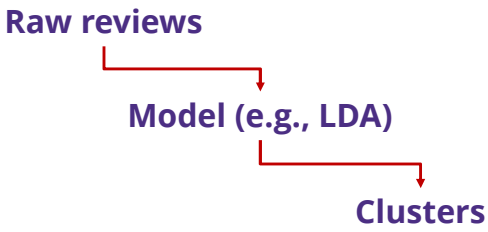
Objective: cluster similar documents together

Hardest parts: ***(1)*** *Conceptual definition of clusters*

(2) *Evaluating and labeling clusters*

Example: identify reviews that discuss compensation

Steps:

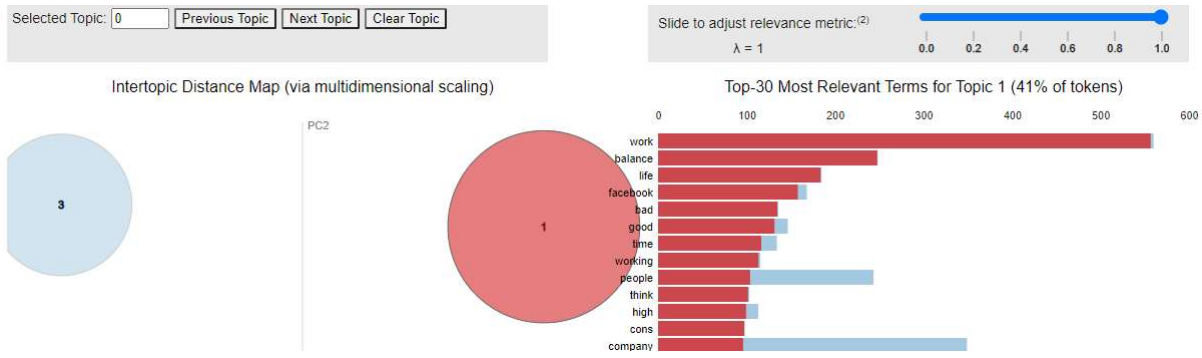


#3 – Clustering – discussion

What categories come to mind for the "cons"?

- Item #1
- Item #2

#3 – Clustering – demo



UNIVERSITY of WASHINGTON

#3 – Clustering

> Let's check the notebook!

UNIVERSITY of WASHINGTON

Conclusion

I hope this session makes you excited about the possibilities of working with textual data!

Want to learn more?

- *Learn Python for Research*
- *Python NLP tutorial*
- Limperg Python course recordings
- Textual Analysis in Accounting: What's Next
- Using Python for Textual Analysis in Accounting Research

Next session ➔ **State-of-the-art: BERT & GPT**

Thank you!

UNIVERSITY of WASHINGTON

FOSTER SCHOOL OF BUSINESS