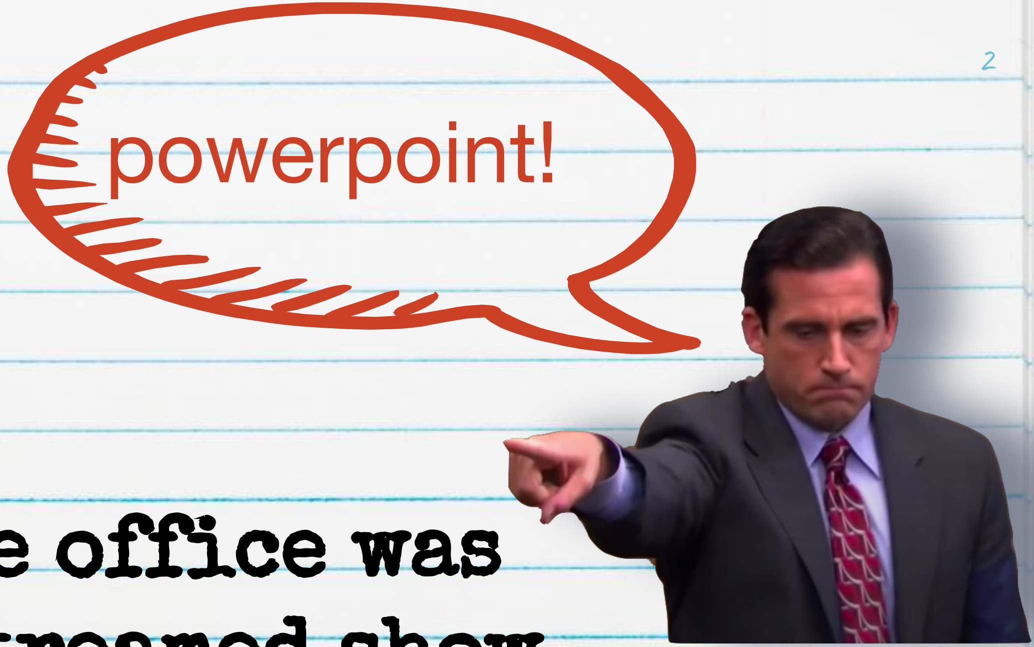
The Office:

modeling iconic character arcs

Aathira Chennat NLP, May 2022

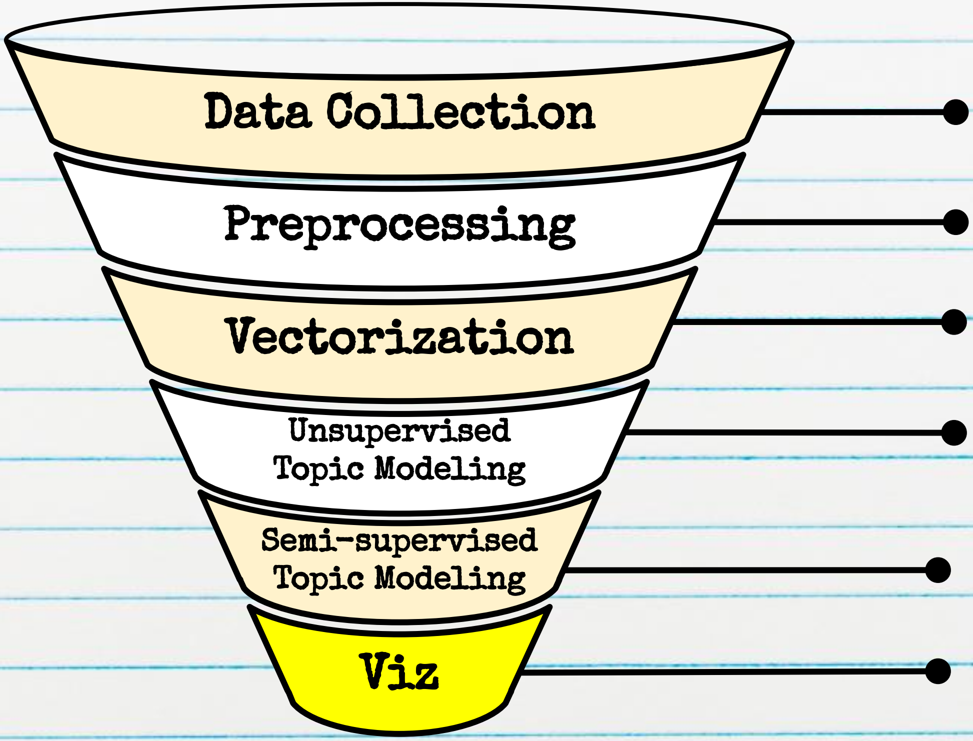


In 2020, the office was the most-streamed show in the US

According to Nielsen, we watched some
57 Billion minutes of this series.

1.Project Design

- ➔ Randomly select scripts from most popular seasons (2-7)
- ➔ Parse into character-specific dataframes
- ➔ Preprocess and tune individual topic models



Data Collection — 66 scripts scraped from ForverDreaming.Com

Preprocessing — Regex and SpaCy pipeline

Vectorization — Count & Term/Inverse Document Frequency

Unsupervised Topic Modeling — PCA, NMF, & LDA

Semi-supervised Topic Modeling — CorEx

Viz — Matplotlib & Scattertext

Characters Modeled

Michael:
392,762

Dwight:
149,501

Jim: **113,022**
Pam: **88,920**

Domain-Specific Preprocessing

Custom Stopwords

Time-specific and office-specific words: *yesterday, morning, week, month, water, coffee, lunch, elevator, floor,* etc.

Nicknames

For Dwight, consolidate *Monkey* and *Angela.* For Michael, consolidate *Pam, Pammy, Spamster.*

SpaCy NER

Recognizing companies, characters, and cities, i.e. New York, Scranton, Pennsylvania, Dunder Mifflin, Ed Truck, David Wallace

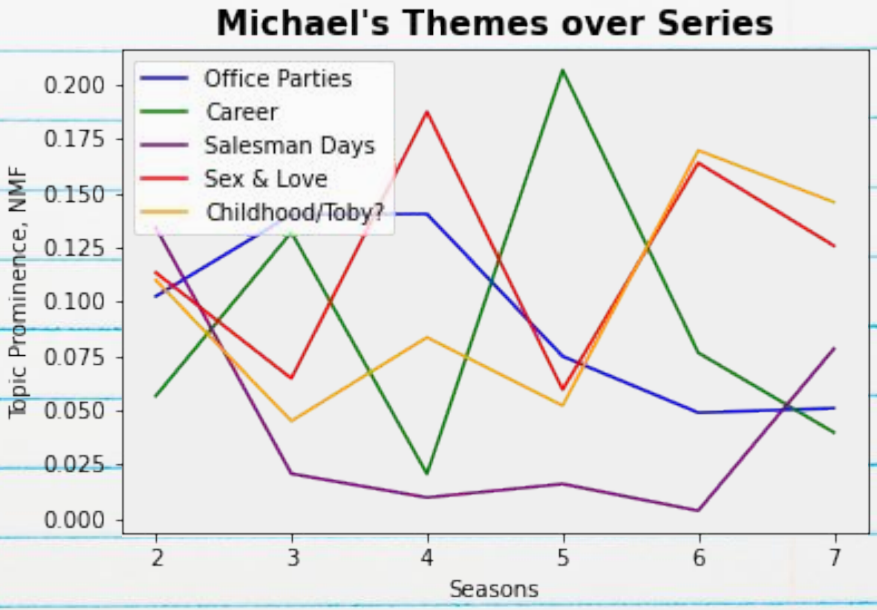
Results

Image Courtesy of NBC

Michael

Top Words:

Pam	(224)
Dwight	(218)
Michael	(173)
Friend	(74)
Toby	(71)
Paper	(71)
Party	(62)



Dwight

Top Words:

Michael (206)
Jim (99)
Dwight (86)
Office (49)
Pam (41)
Desk (41)
Hay (22)

Jim

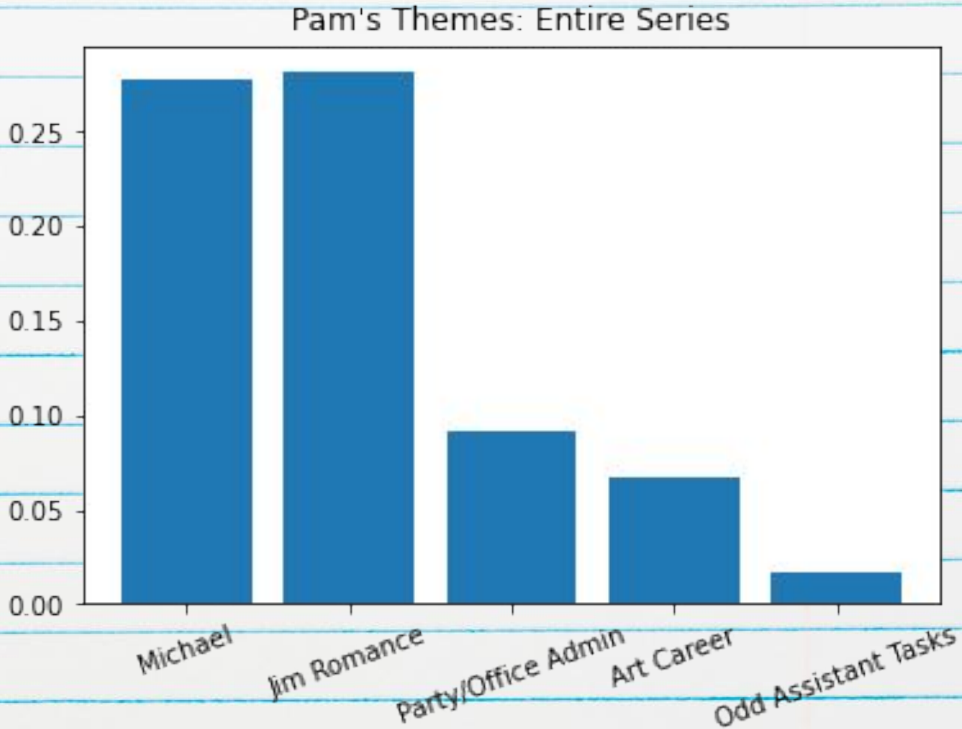
Top Words:

Pam	(122)
Dwight	(118)
Michael	(114)
Jim	(53)
Office	(41)
Desk	(32)

Pam

Top Words:

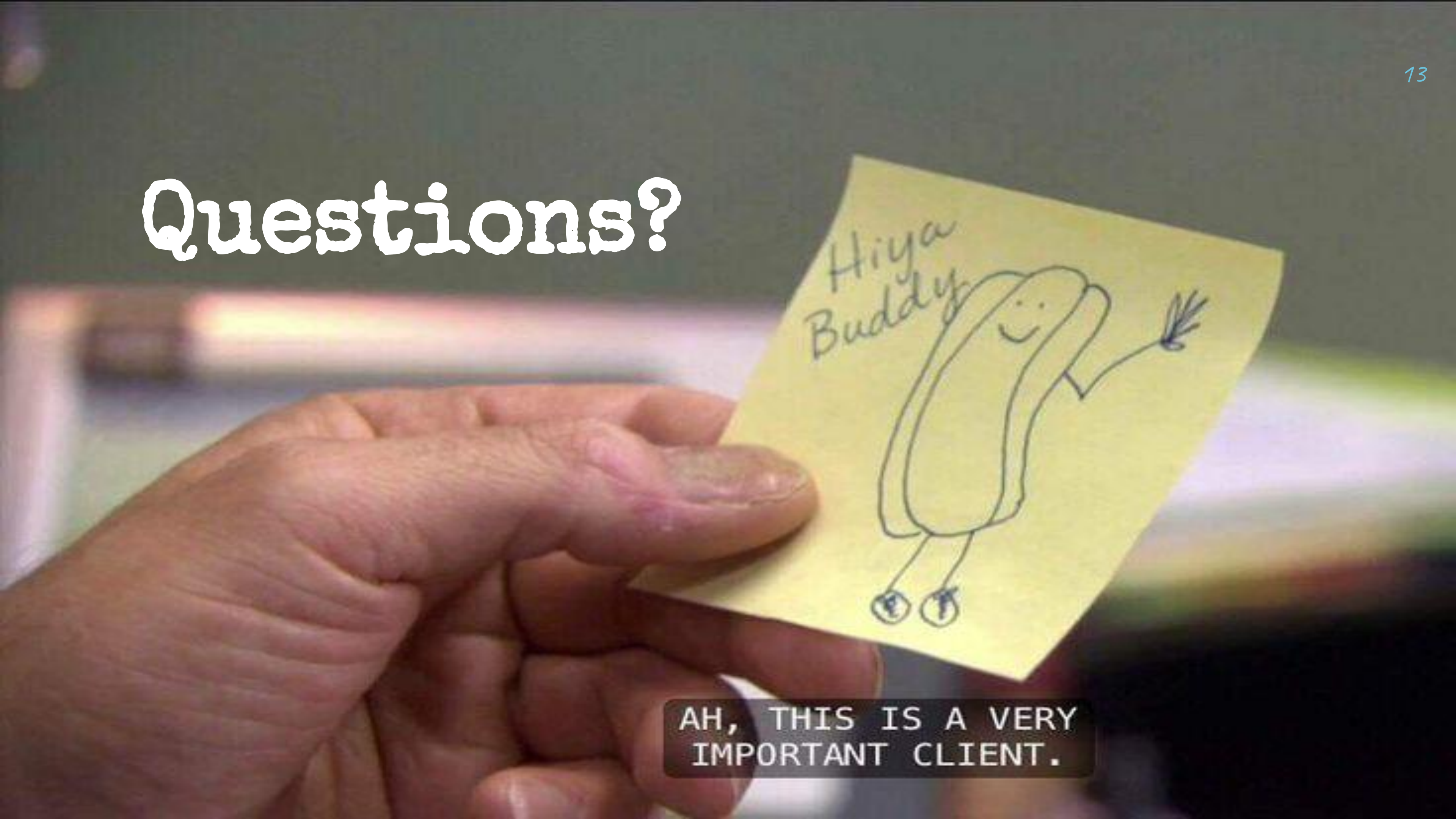
Michael	(177)
Jim	(98)
Pam	(40)
Dwight	(38)
Mom	(27)
Angela	(25)
Chair	(21)



Future Work

1. Compile scripts from all episodes for larger corpus, stronger results.
2. Create a content-based recommender for certain characters and relationships using finalized topics models.

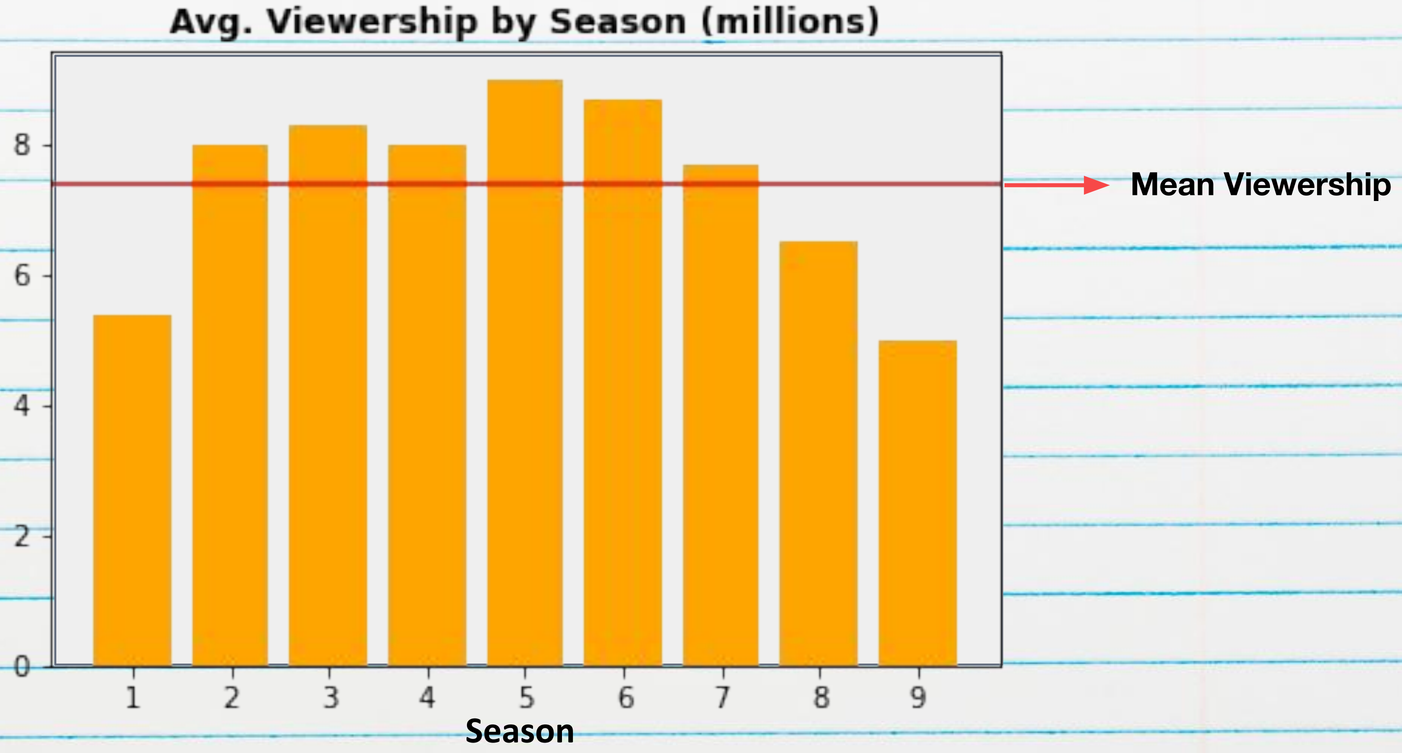
Questions?



Appendix

1. Spangler, Todd. “‘The Office’” Was By Far the Most-Streamed TV Show in 2020, Neilsen Says.” Variety, 01/12/2021. Retrieved 04/29/2021 at: https://variety.com/2021/digital/news/the-office-most-streamed-tv-show-2020-nielsen-1234883822
2. Slides Carnival for Presentation Slides Template

Avg. Viewership by Season


Avg. Viewership by Season (millions)
Season
Mean Viewership