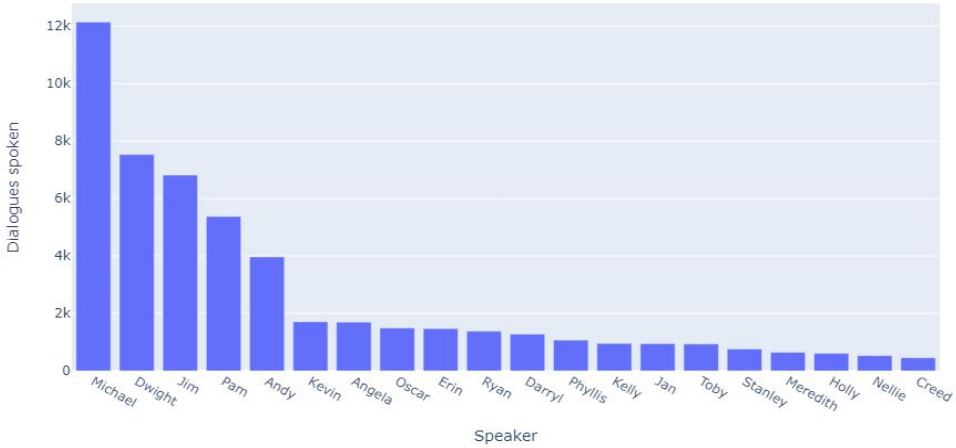
That's what she said!

~ The Office (US) script analysis

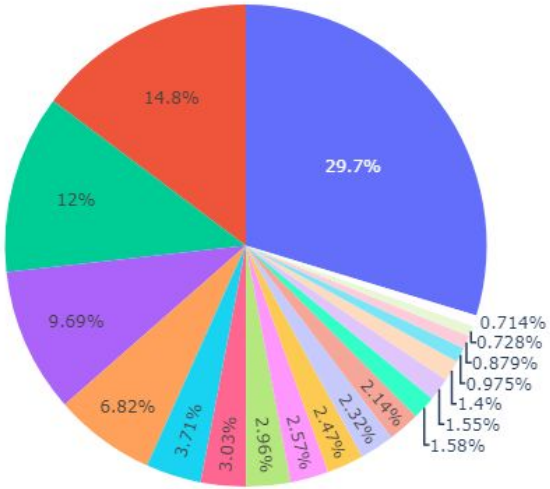
Luuk Boekestein, Shantanu Motiani and Eline Westerbeek

Initial Analysis and Visualizations

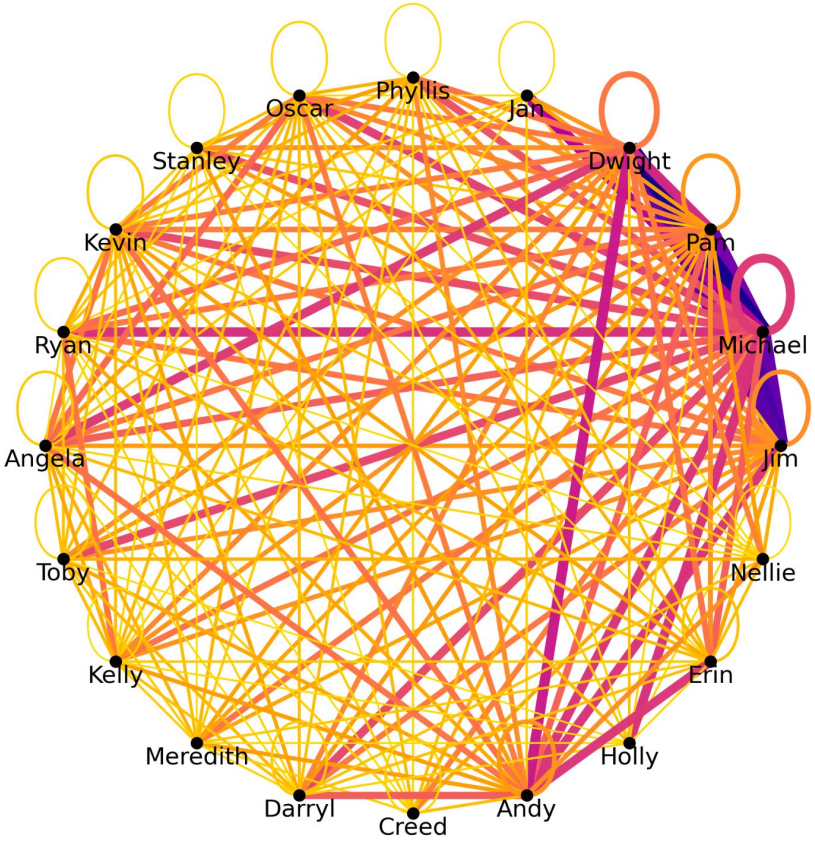
Lines spoken by popular characters

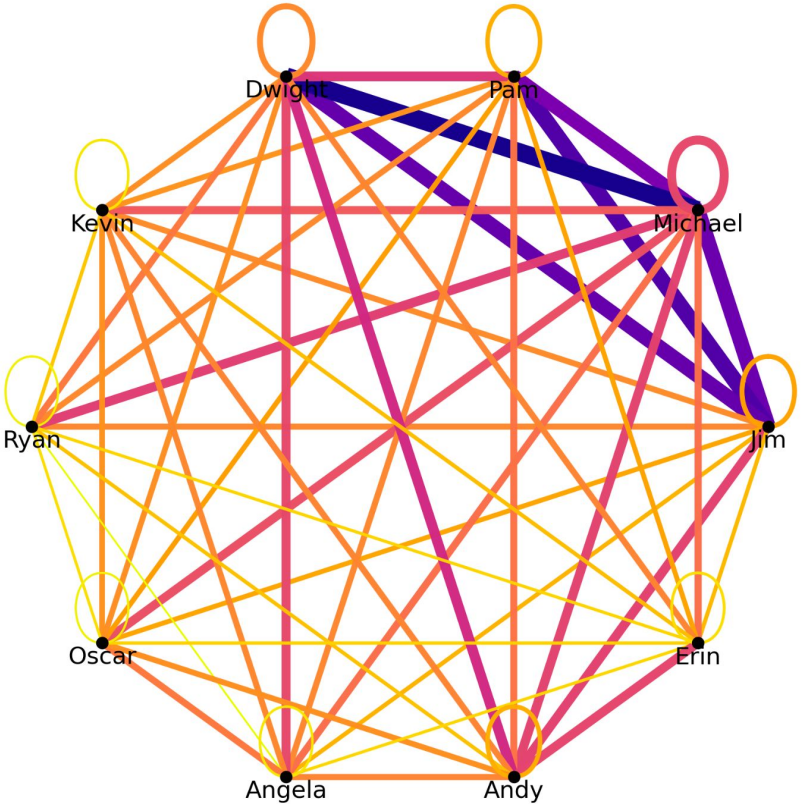


Season 5



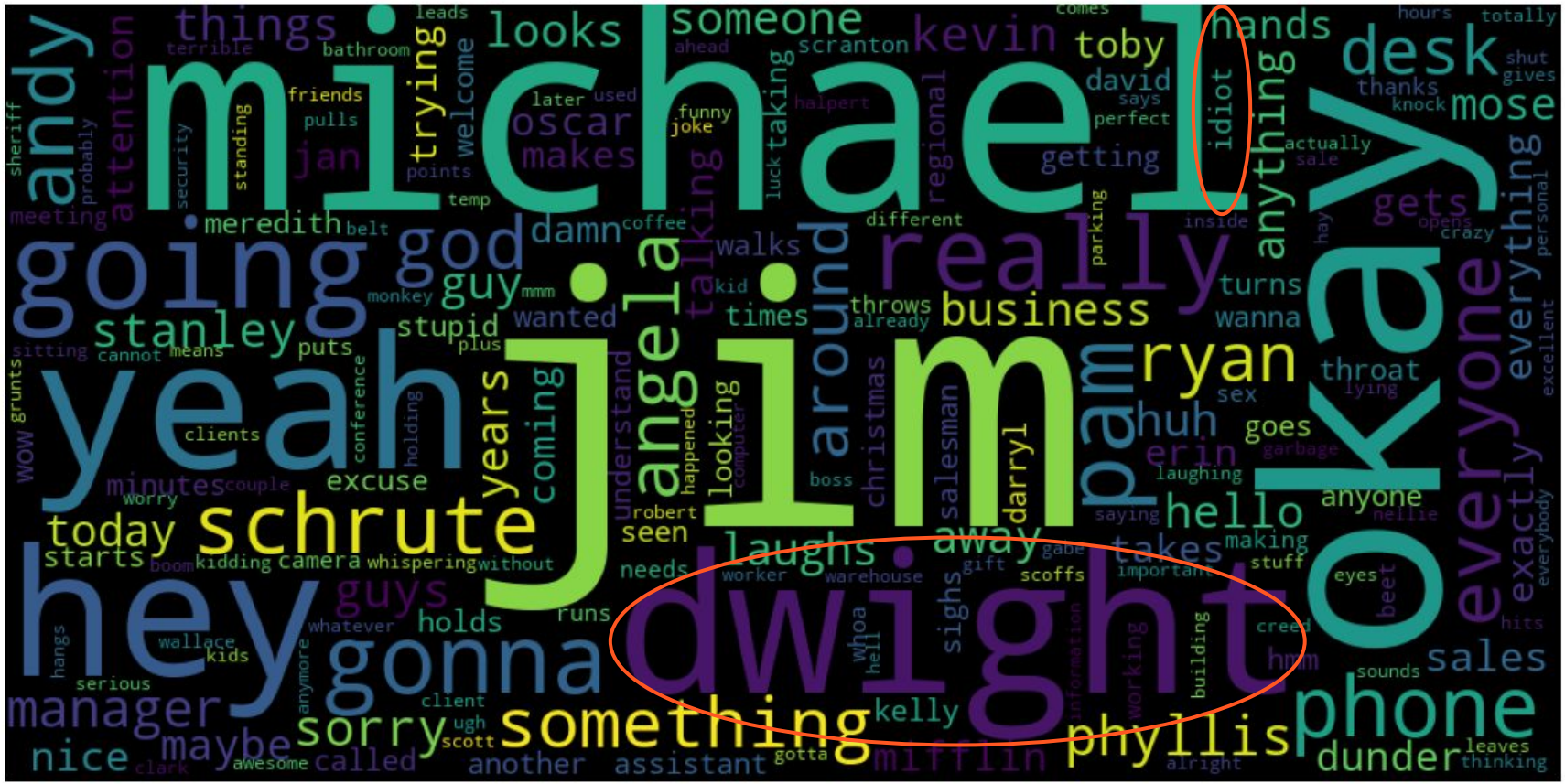
Character interactions





Dialogues by season

Dwight's Word-cloud



Sentiment Analysis

Sentiment analysis

The problem:

- **Classify short lines of text on sentiment**

Made test set by annotating 300 lines manually

line_text	Sentiment
All right Jim. Your quarterlies look very good...	?
Oh, I told you. I couldn't close it. So...	?
So you've come to the master for guidance? Is ...	?
Actually, you called me in here, but yeah.	?
All right. Well, let me show you how it's done.	?
[on the phone] Yes, I'd like to speak to your ...	?
I've, uh, I've been at Dunder Mifflin for 12 y...	?
Well. I don't know.	?
If you think she's cute now, you should have s...	?
What?	?
Any messages?	?
Uh, yeah. Just a fax.	?
Oh! Pam, this is from Corporate. How many time...	?
You haven't told me.	?
It's called the wastepaper basket! Look at tha...	?

Models used for Sentiment analysis

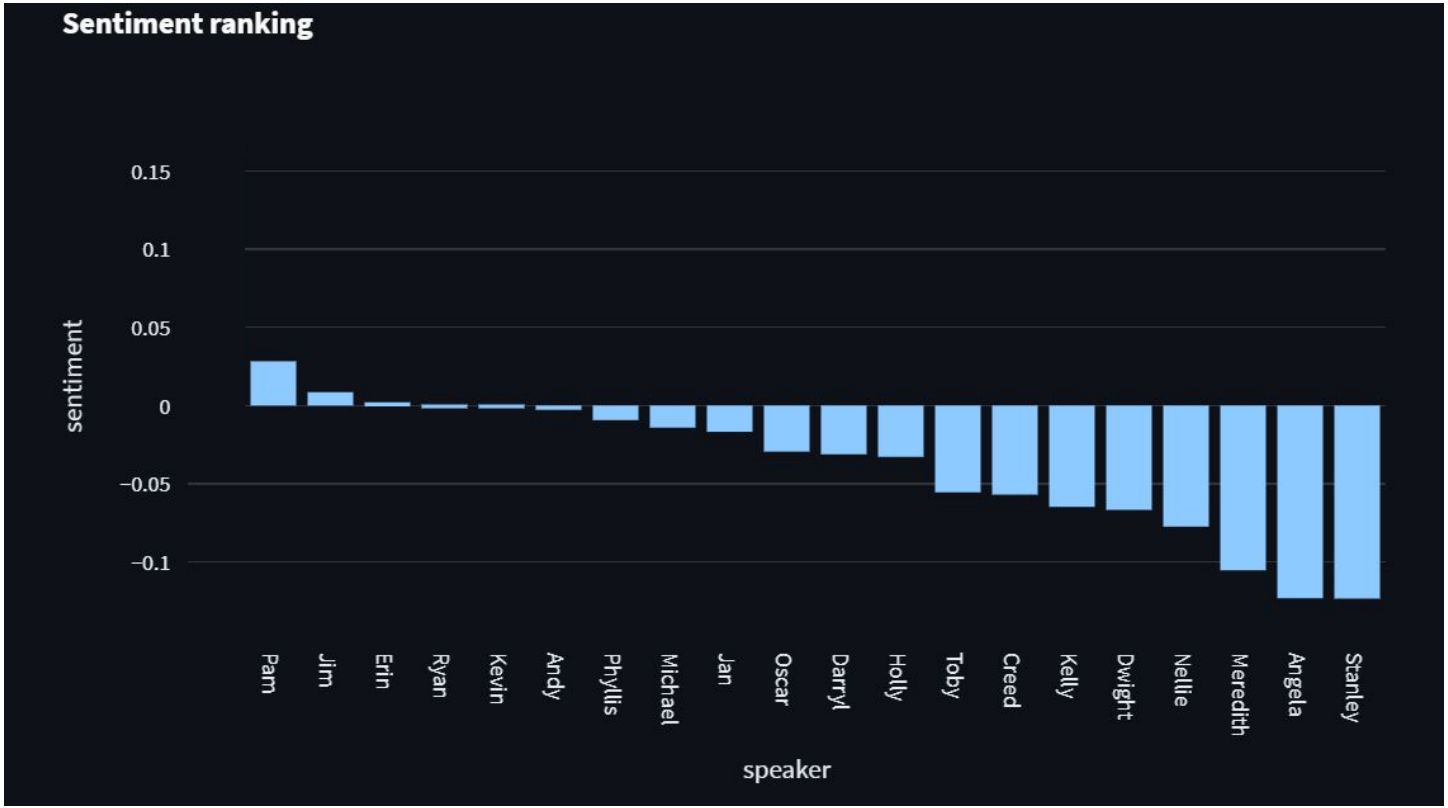
- **BERT cased**
- **BERT uncased**
- **Roberta**
- **Distilbert**

Model	Preprocessing	Accuracy	Precision	MSE
Roberta large	Remove descriptions	0.67	0.69	1.29
Roberta large	Keep descriptions	0.65	0.67	1.36
Bert uncased	Remove descriptions	0.63	0.63	0.46
Bert uncased	Keep descriptions	0.62	0.63	0.47
BERT	Remove descriptions	0.54	0.54	0.65
BERT	Keep descriptions	0.55	0.55	0.66

Limitations

- **Lines of text are very short**
- **“Test” data is small, potentially biased**
- **Accuracy of 67% for 3-class classification**
- **Lines are missing a lot of context**
    - **Sarcasm, expressions, etc.**

Nevertheless, results:



Results: Andy

Starts dating Angela **Becomes manager**

grouped by scene!

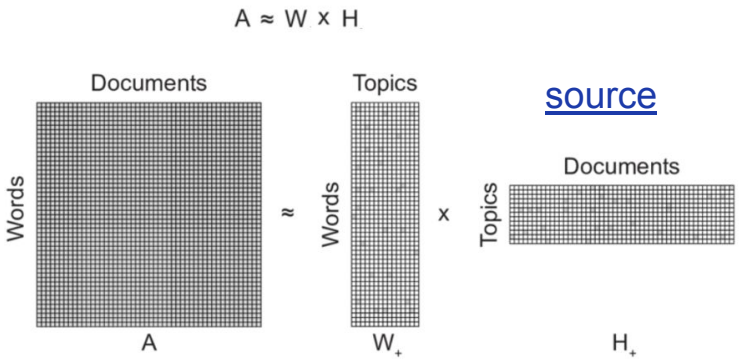
Topic modelling

LDA

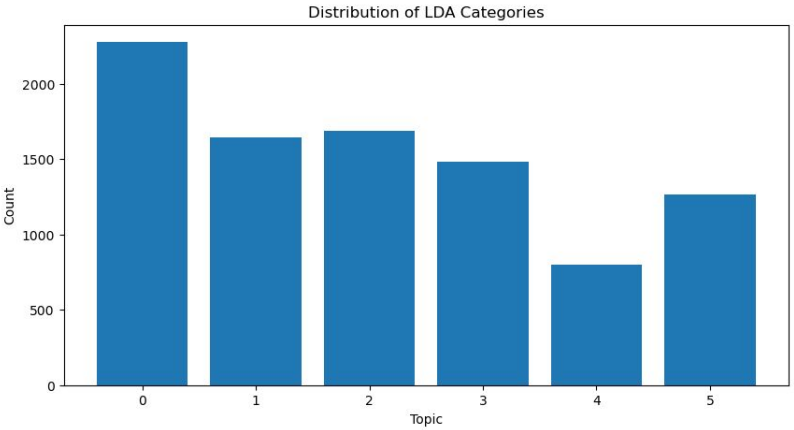
- probabilistic model
- assumes documents are made up of a mixture of topics
- each topic is characterized by a distribution of words

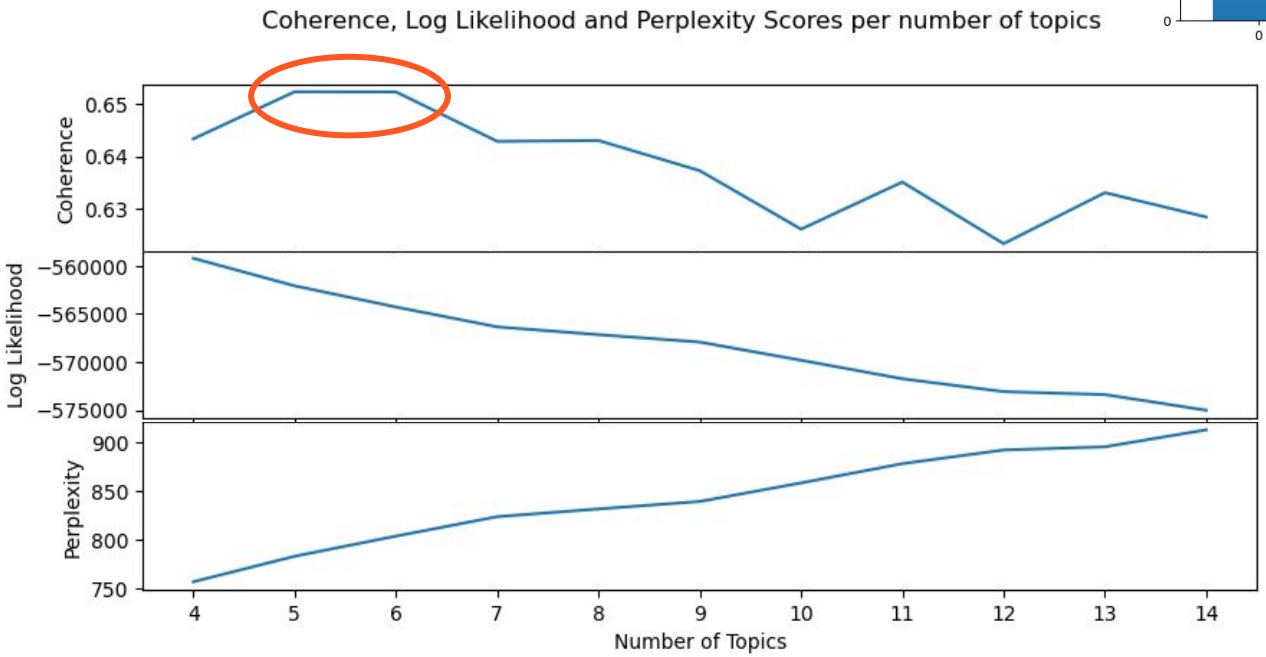
NMF

- decomposes a matrix into two non-negative matrices
- each document seen as a linear combination topics
  - each topic is represented by word weights



LDA: Choosing no. of topics



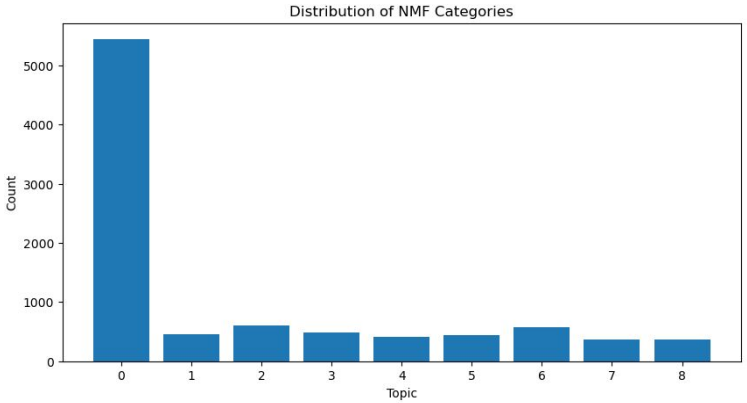


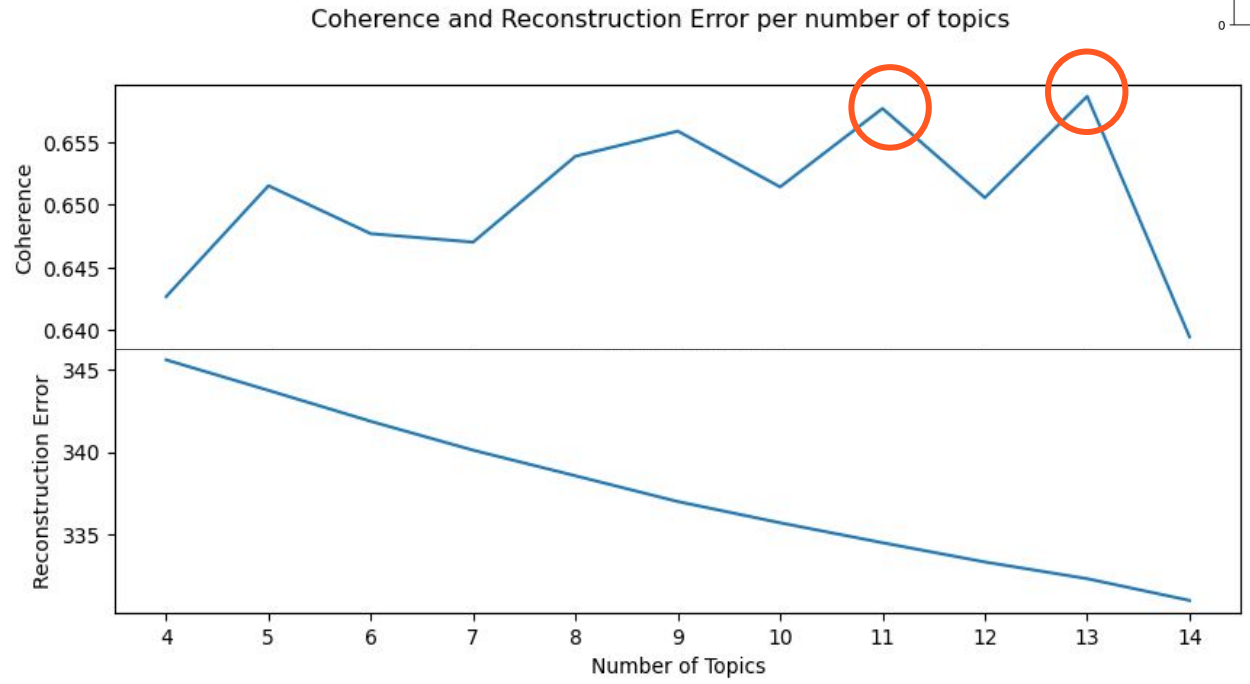
LDA Topics

No	Topic name	Top 10 words
0	??	phone sorry paper dunder mifflin woman company scranton listen black
1	breaks/meeting?	friend meeting break sorry singing welcome weird lunch family voice
2	parties/holidays?	party money christmas everybody today check throw movie tonight dream
3	??	night year question laugh manager watch got ready happened exactly
4	the office?	getting office month picture excuse sorry talking computer person table
5	humour?	office laugh place better crazy sound remember pretty number funny

quite unclear...

NMF: Choosing no. of topics





NMF Topics quite coherent :)

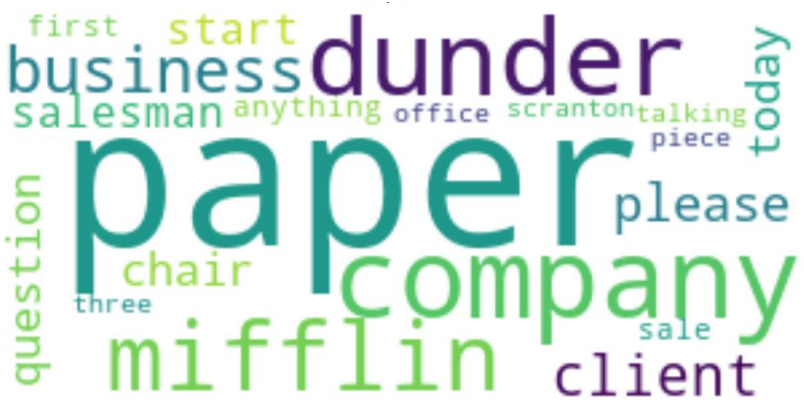
No	Topic name	Top 10 words
0	general dialogue	everybody today talking start better woman meeting night minute place
1	phone calls	phone hang ring answer client calling listen imitating transfer sound
2	the office	office camera chair place knock parking conference walking lunch building
3	the company and business	paper dunder mifflin company business client scranton sale salesman question
4	parties and party planning	party christmas committee planning start throw break everybody starting birthday
5	apologising?	sorry voice stupid excuse hand apology client doing probably pretty
6	humour and jokes	laugh funny laughing pretty tonight year place camera welcome drink
7	roles/positions in the company	manager regional assistant branch scranton wallace person position better sale
8	friendships and relationships	friend smart happy totally thinking different understand girlfriend sitting voice

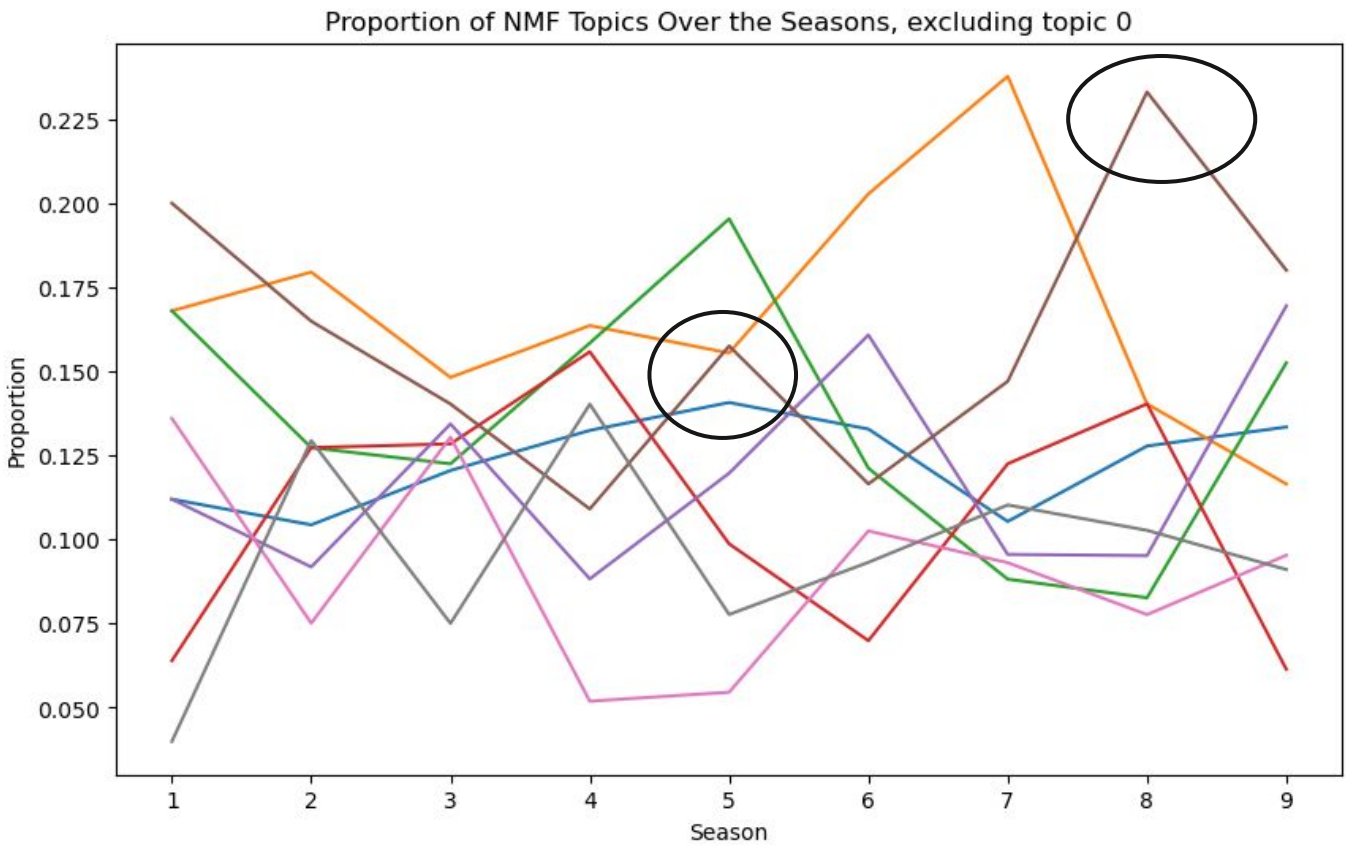
Word Cloud Examples

1: Phone Calls

4: Parties + Party Planning

3: The Company and Business





No	Topic name
1	phone calls
2	the office
3	the company and business
4	parties and party planning
5	apologising?
6	humour and jokes
7	roles/positions in the company
8	friendships and relationships

Dialogue generation with Markov Chains

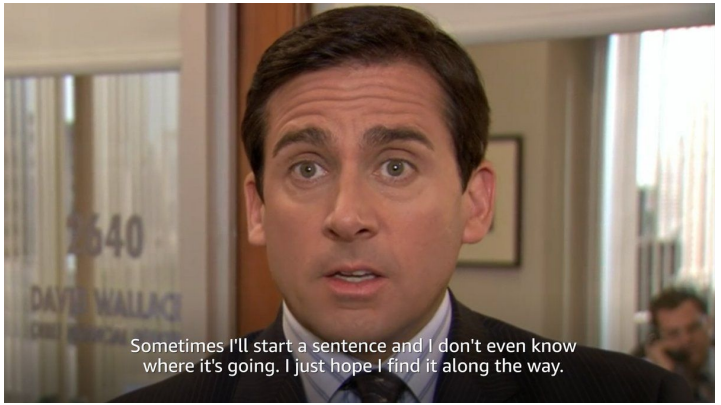
Markov Chains

A Markov chain, named after Andrey Markov, is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules.

The defining characteristic of a Markov chain is that no matter how the process arrived at its present state, the possible future states are fixed. In other words, the probability of transitioning to any particular state is dependent solely on the current state.

Examples: words, weather conditions, football score-lines, or stock performances.

How did it perform for us?



“Congratulations. All right. It's time to carbo-load.” ~ Michael

“It's a lot going on, so what you do to it?” ~ Dwight

“No. That is cool. Well, I'll see you in a meeting at 7:30 with a very classy event, a night in.” ~ Jim

“Ah damn. That sounds like he's mafia though”

~ Michael

“Uh, yes, I'm in pain. I love pain. To me, pain is the office board not having deniability”

~ Dwight

“You hadn't noticed she's a part-time frozen yogurt chef.”

~ Andy

Web App

resources

https://brilliant.org/wiki/markov-chains/

https://towardsdatascience.com/text-generation-with-markov-chains-an-introduction-to-using-markovify-742e6680dc33

https://www.kaggle.com/code/nhuhduong/interaction-in-the-office

https://www.kaggle.com/datasets/fabriziocominetti/the-office-lines

https://www.analyticsvidhya.com/blog/2021/06/part-15-step-by-step-guide-to-master-nlp-topic-modelling-using-nmf/