

# POLITICAL OPINIONS IN NON-POLITICAL SPACES

(OR HOW PROJECTS  
SOMETIMES TAKE  
UNEXPECTED TURNS)



Johanna zum Felde, Fabian Heindl, Katerina Turkova



# AGENDA

Project Idea & Developments

Data Collection

Analysis

Results

Discussion

# PROJECT IDEA & DEVELOPMENTS





# Initial idea

- **Find out how people share political opinions in non-political spheres (topic-driven)**
- **Case study: User reviews on platforms for books (e.g., Letterboxd, Goodreads) and movies (e.g., IMDb)**



# Issues

- 1. How should we differentiate between political and non-political spheres?**
- 2. How should we actually get the data? (Letterboxd, Goodreads & IMDb do not provide suitable APIs)**

# Change of plan

+


•

○

- 1. Let's focus on one media format (series)**
- 2. Let's compare a political (choice: House of Cards) and a non-political (choice: Orange is the New Black) case**
- 3. Let's compare grey (or probably even illegal) ways of data collection (= scraping IMDb reviews) with legal ways of data collection (= comments on YouTube through the official API) (= change from topic-driven to method-driven)**



# DATA COLLECTION



House of Cards (2013–2018)
User Reviews
+ Review this title

948 Reviews
☐ Hide Spoilers
Sort by: Featured
Filter by Rating: Show All

★ 7/10

**Stood firm for four seasons, collapsed completely in Season 5**  
TheLittleSongbird 15 November 2017

'House of Cards' much of the time was one of the most compelling shows. Sadly, it has also become one of the most frustrating. Not since 'Once Upon a Time' and 'The Walking Dead', and before that 'Lost' has such a brilliant show of great promise declined so rapidly.

Lets start with the many great things first. For the first four seasons, 'House of Cards' was seriously addictive, must-watch television and very quickly became one of my favorite shows. Thankback it

90 out of 97 found this helpful. Was this review helpful? [Sign in to vote.](#)

[Permalink](#)

★ 7/10

**Great start for the earlier seasons only to collapse in the final season**  
gavin-thelordofthefu-48-460297 31 August 2020

After being denied the position as Secretary of State, a congressman named Frank Underwood and his wife Claire Underwood work for each other in the white house in order for him to be elected as President of the United States. At the same time, he

```

Disallow: /_json/video/mon
Disallow: /_json/getAdsForMediaViewer/
Disallow: /list/ls*/_ajax
Disallow: /list/ls*/export
Disallow: /*/*/rg*/mediaviewer/rm*/tr
Disallow: /*/rg*/mediaviewer/rm*/tr
Disallow: /*/mediaviewer/*/tr
Disallow: /title/tt*/mediaviewer/rm*/tr
Disallow: /name/nm*/mediaviewer/rm*/tr
Disallow: /gallery/rg*/mediaviewer/rm*/tr

```

# Scraping IMDb

- Official API does not allow for collection of user reviews
- Policy does not allow for scraping of data we need

...

Let's do it anyway!





# Scraping IMDb: Attempt 1

Let's write a script that...

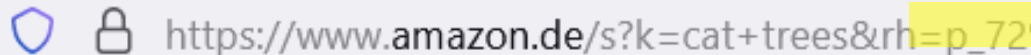
- extracts the necessary information through the `html.nodes()` function
- stores the information in a data frame

Result:

It works, but we only get the first 25 reviews (out of 948 for only HoC alone)

Idea for a solution:

We only have to add a function that extracts the reviews according to each page according to the count in the URL, right?

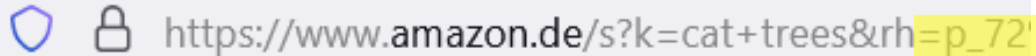


A screenshot of a web browser's address bar. It shows a lock icon, a shield icon, and the URL `https://www.amazon.de/s?k=cat+trees&rh=p_72'`. The text `p_72'` is highlighted in yellow.

→ Spoiler: IMDb is smarter than that

# Scraping IMDb: Attempt 2

Finding: IMDb uses a (hidden) pagination key system instead of a simple page count

 [https://www.amazon.de/s?k=cat+trees&rh=p\\_72](https://www.amazon.de/s?k=cat+trees&rh=p_72)

VS

```
> <div class="row text-center lister-working hidden" style="display: none;"> </div>
▼ <div class="load-more-data" data-key="g4w6ddbmqyzdo6ic4oxwjnzqrhrmuab63ykdx6piapd74ud5pjt6uds4oiyfjmjdb4dzah5d2jttflh6y73ipq5kkmd5o" data-ajaxurl="/title/tt1856010/reviews/_ajax">
  > <div class="ipl-load-more ipl-load-more--loaded"> </div>
```

# Scraping IMDb: Attempt 2

Let's write a new script (with generous help by ChatGPT) that...

- mimics a web browser
- automatically extracts the pagination key for each 25 reviews
- extracts the necessary information through the `html.nodes()` function
- stores the information in a data frame

Will it work?...

	Author	Rating	Title	Review
1	TheLittleSongbird	7/10	Stood firm for four seasons, collapsed completely in Season...	'House of Cards' much of the time was one of the most com...
2	gavin-thelordofthefu-48-460297	7/10	Great start for the earlier seasons only to collapse in the fin...	After being denied the position as Secretary of State, a cong...
3	armastusmaitse	6/10	Simply not worth it anymore after Kevin Spacey left...	Simply not worth it anymore after Kevin Spacey left... the ser...
4	gogoschka-1	9/10	5 Seasons Of Highly Addictive Political Drama And Darkly F...	I love spectacular TV-shows with amazing production values...
5	DiCaprioFan13	9/10	Great Political Drama!	House of Cards is about a ruthless politician (Kevin Spacy) a...
6	crescendo_1	10/10	I knew it... Pathetic...	The moment I heard they are gonna drop Kevin Spacey, I kn...
7	cncgjqbm	1/10	Robin Wright Destroyed this Show	This series was an exceptional political thriller, until Robin W...
8	Bigkingp007	8/10	What IDIOTS! Legacy destroyed!	How do you go from 10 to zero in sixty seconds? Kevin Spa...
9	Supermanfan-13	2/10	Worth Watching!	House of Cards really is as good as everyone says and I now...
10	Hitchcoc	2/10	What a Disappointment	I've reviewed all the episodes for Season 6 but feel the need...
11	guyzradio	2/10	Seasons are a steady slide from excellent to a complete wa...	I remember watching Season 1 of House of Cards and being...
12	m-m-mcadam	8/10	Terrible terrible ending	After committing so much time to what was once a great sh...
13	michelle_kummer	8/10	Kevin Spacey is House if Cards	It's not the same without Kevin, he is one of the best actors ...
14	HHTurkish	7/10	Disappointing Drivel	No way around it, Season 6 was awful ...in every respect that...
15	hdaytemur	1/10	There's no reason to watch without Kevin Spacey.	Claire's character is being portrayed as an important figure f...
16	contactgrod	7/10	WTF WAS THAT!? Terrible Season 6	Forgot a rating -I demand my 8 hours back! I want Netflix to...
17	Tweekums	1/10	Solid political melodrama that loses its way towards the end	This Netflix series, based on a BBC series, is centred on Fran...
18	noskins-23349	10/10	No Spacey no show	Don't watch it at all. These propagandas have ruined a great...
19	me-589-145643	1/10	Pioneering Perfection	I firmly believe that one of the major aspects of what makes...
20	larrytate-927-656516	9/10	Nothing without Spacey	House of Cards Seasons 1-5 was one of the best series ever ...

Showing 1 to 20 of 948 entries, 4 total columns

YES!



# Collecting YouTube comments

- Fairly easy process due to the existence of an official API (YouTube Data API V3)
- Requires prior registration and enabling through Google Cloud services

(instructions can be found on:

[https://bookdown.org/paul/apis\\_for\\_social\\_scientists/youtube-api.html](https://bookdown.org/paul/apis_for_social_scientists/youtube-api.html))

- Downsides:
  - No direct way to collect answers to comments by other users
  - Potential issues with encoding of emojis and other symbols used in comments (specifically when converting data into .csv files)

# Result:

Total data	IMDb reviews	YouTube comments
House of Cards (official trailers for S1-6)	948	4833
Orange is the New Black (official trailers for S1-7)	538	11439

Let's analyse!

# ANALYSIS



# Analysis

## Preprocessing

1. Check for missing values and remove them (not necessary in all of the datasets)
2. Create a corpus from the 'Review' or 'Comments' columns
3. Tokenize corpus and remove unnecessary elements
4. Clean out stopwords and words with 2 or fewer characters





# Analysis

## FCM Model

We created a co-occurrence matrix `fcf()` indicating co-occurrences of words within a window of 5

# Analysis

## Word Embeddings - GloVe Model

- created a GloVe model object (rank = 100, x\_max = 10)
- fit model and return embeddings (n\_iter = 10, n\_threads = 8)
- Vector model

Dataset: House of Cards YouTube Comments

politics	real	life	What	US	says	Vind	American	original
1.0000000	0.5072032	0.4611152	0.4370924	0.4322750	0.3779404	0.3730363	0.3610259	0.3591885
is	trailers	Tiram	actor	female	he			
0.3541167	0.3450757	0.3397124	0.3380687	0.3265951	0.3216673			

# Analysis

..but also

	V1	V2	feature
	<dbl>	<dbl>	<chr>
1	-0.0350	-0.00713	Netflix
2	-0.0393	-0.0298	what
3	-0.0638	-0.00265	a
4	0.000293	0.00761	bunch
5	-0.0631	0.0410	of
6	0.0112	0.0137	idiots

## Words as Vectors

	[,1]	[,2]	[,3]	[,4]	[,5]
Netflix	-0.4668306	-0.2331532	-0.14830390	-0.15021185	0.006563899
what	-0.5352057	-0.2144846	-0.01960057	0.43268122	-0.410759689
a	-0.5980483	-0.7189317	-0.16591887	0.13268030	0.018109101
bunch	0.1434041	-0.2253619	-0.12707456	0.28419248	0.416129689
of	-0.5066481	-0.7635489	0.35342469	-0.83787931	-0.172223629
idiots	0.1906750	-0.3159724	0.41689757	0.02448499	0.294078509

Dataset: House of Cards YouTube Comments

# Analysis

## Preparation for Visualization

- Dimension reduction using irlba
- Define the keywords and patterns for visualization

("political", "economy", "ideology", "affairs", "religion", "philosophy", "democracy", "policy", "demographics", "nationalism", "imperialism", "geography", "socialist", "wing", "culture", "democratic", "relations", "socialism", "politically", "society", "communism", "economic", "diplomacy", "parties", "presidency", "overview", "opposition", "liberal", "governments", "politicians", "sociology", "history", "business", "nationalist", "foreign", "judicial", "elections", "conflict", "government", "background", "president", "communist", "marxist", "marxism", "parliament", "constitution", "parliamentary", "disputes", "constitutional", "matters", "politic\*", "econom\*", "ideolog\*", "affair\*", "religion\*", "philosoph\*", "democ\*", "polic\*", "demograph\*", "national\*", "imperial\*", "geograph\*", "socialis\*", "wing\*", "cultur\*", "relations", "societ\*", "communis\*", "diploma\*", "parties", "presiden\*", "overview\*", "opposition\*", "liberal\*", "government\*", "sociolog\*", "histor\*", "business\*", "foreign", "judicial", "election\*", "conflict\*", "background\*", "marxis\*", "parliament\*", "constitution\*", "disput\*", "matters")

DATA

5 tensors found  
Word2Vec 10K

Label by word Color by No color map

Edit by word Tag selection as

Load Publish Download Label

☒ Sphereize data

Checkpoint: Demo datasets

Metadata: oss\_data/word2vec\_10000\_200d\_labels.tsv

UMAP T-SNE PCA CUSTOM

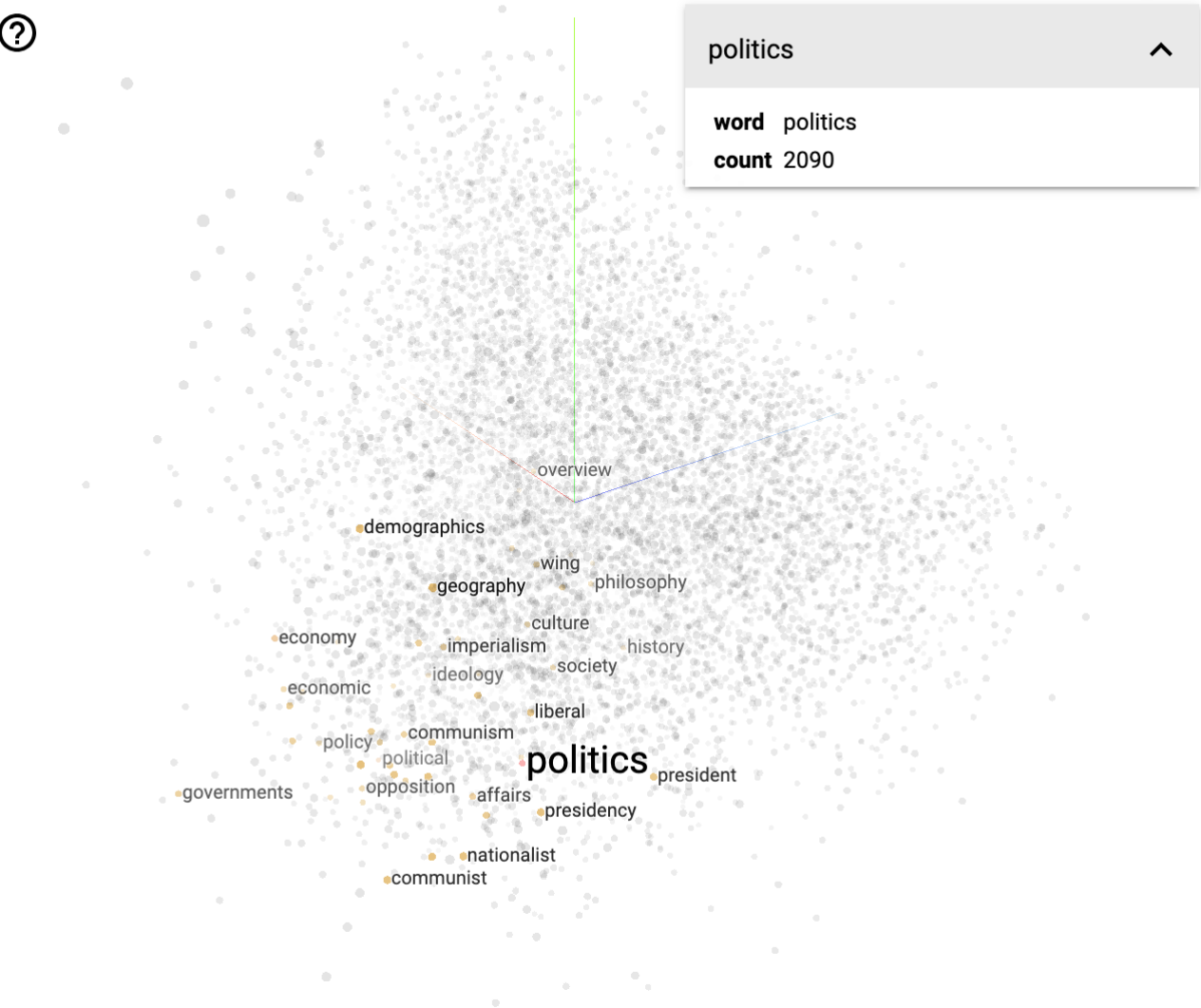
X Component #1 Y Component #2

Z Component #3 ☒

PCA is approximate.

Total variance described: 8.5%.

Points: 10000 | Dimension: 200 | Selected 51 points



Show All Data Isolate 51 points Clear selection

Search politics by word

neighbors 50

distance COSINE EUCLIDEAN

Nearest points in the original space:

political	0.523
economy	0.549
ideology	0.587
affairs	0.591
religion	0.591
philosophy	0.598
policy	0.603
democracy	0.607
demographics	0.622
nationalism	0.626
imperialism	0.628
geography	0.630

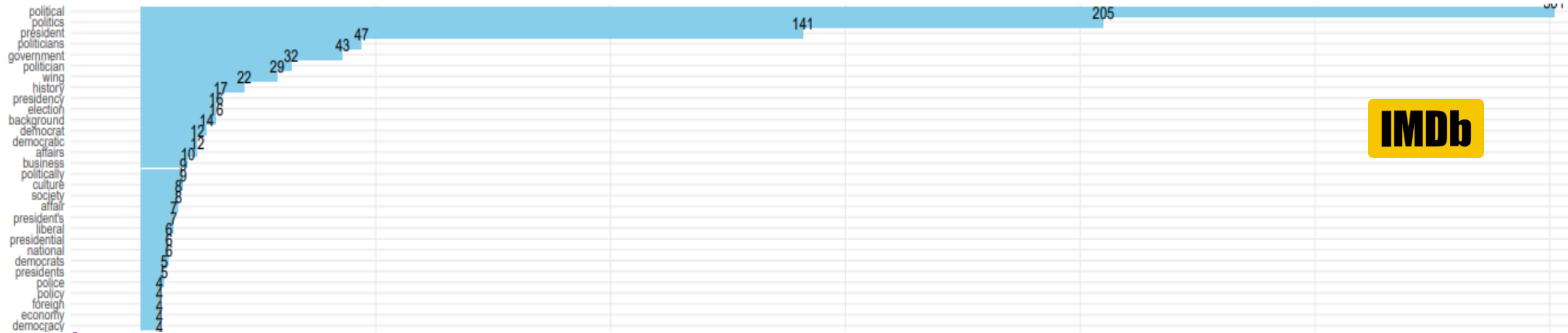
BOOKMARKS (0)

# RESULTS

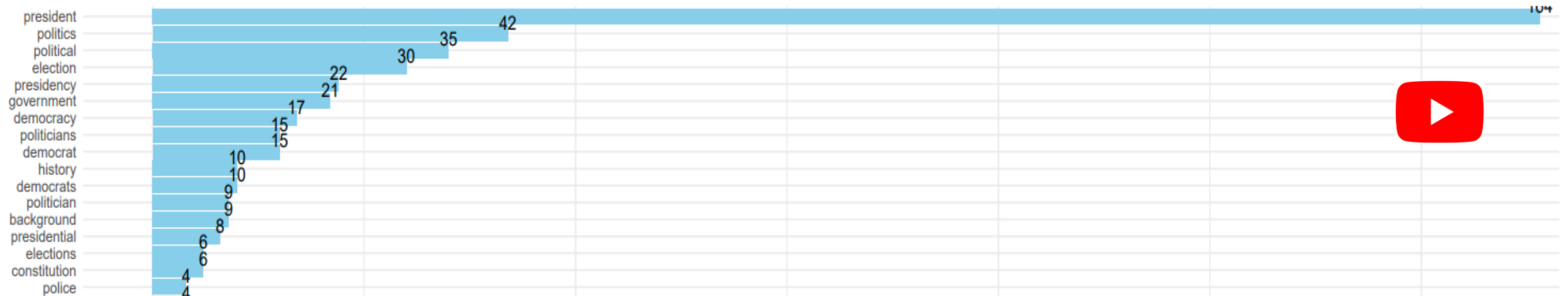


# Results HoC – Total keyword counts

Total keyword counts, HoC, IMDB

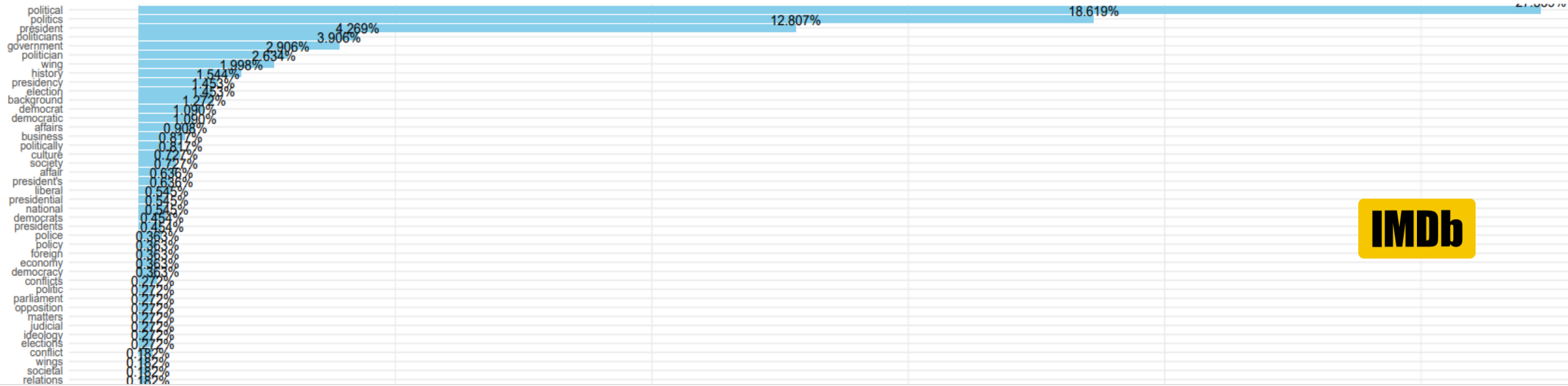


Total keyword counts, HoC Trailer 1-6, YouTube



# Results HoC – Relative keyword counts

Relative keyword counts, HoC, IMDB

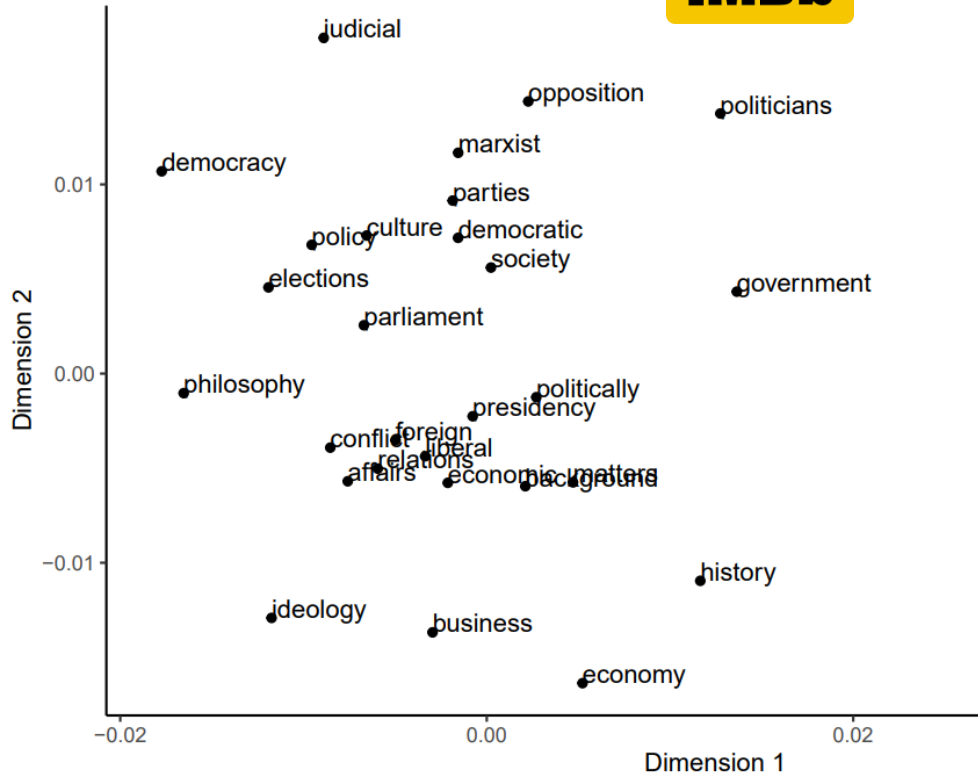


Relative keyword counts, HoC Trailer 1–6, YouTube





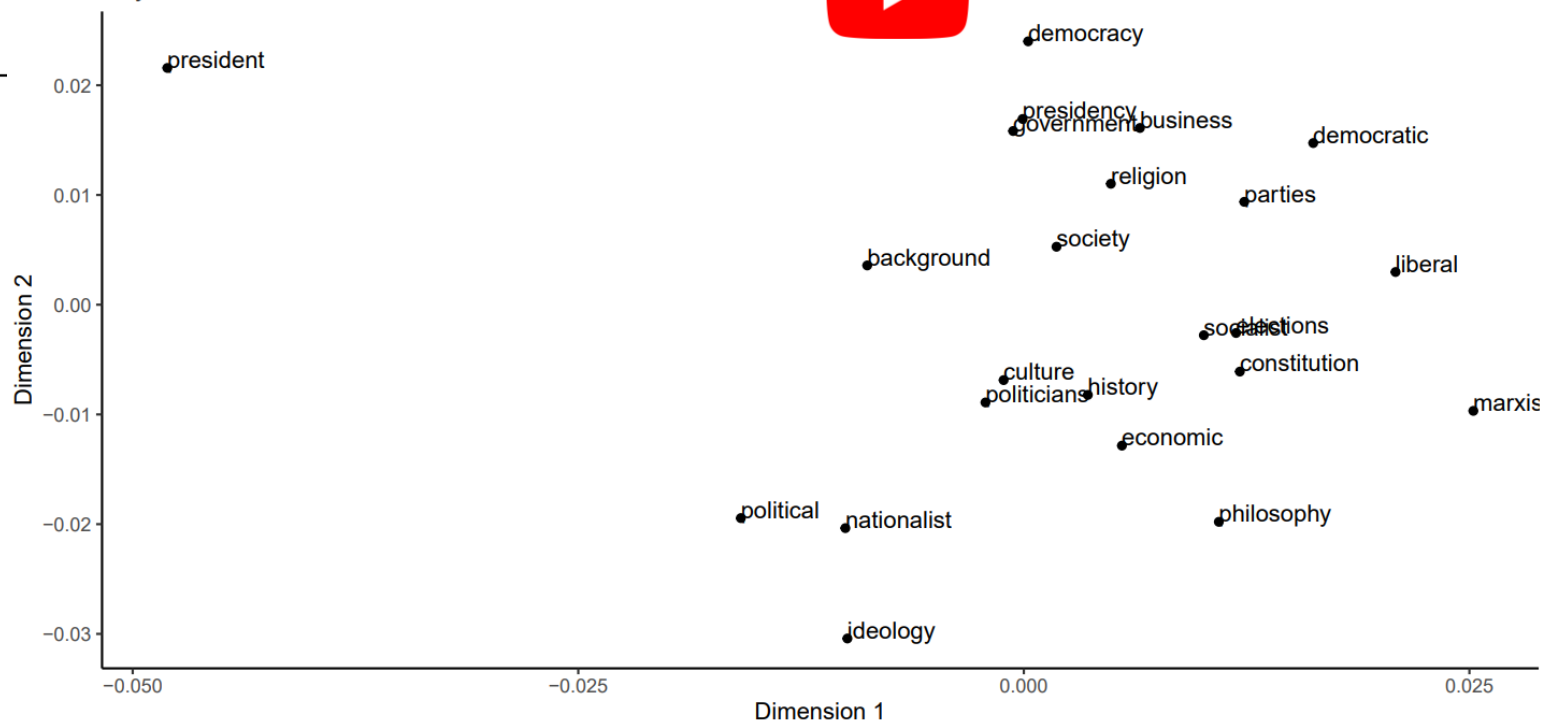
Keywords, HoC, IMDB, 100 dimensions

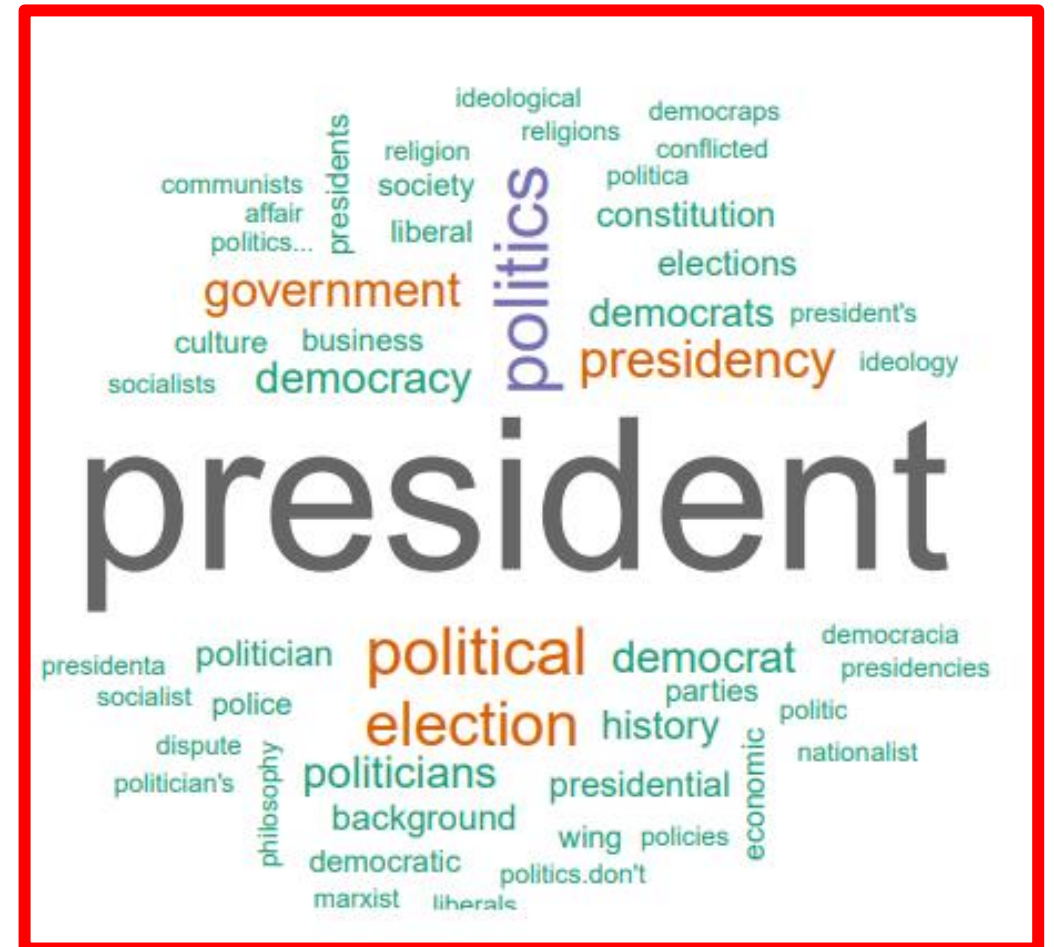
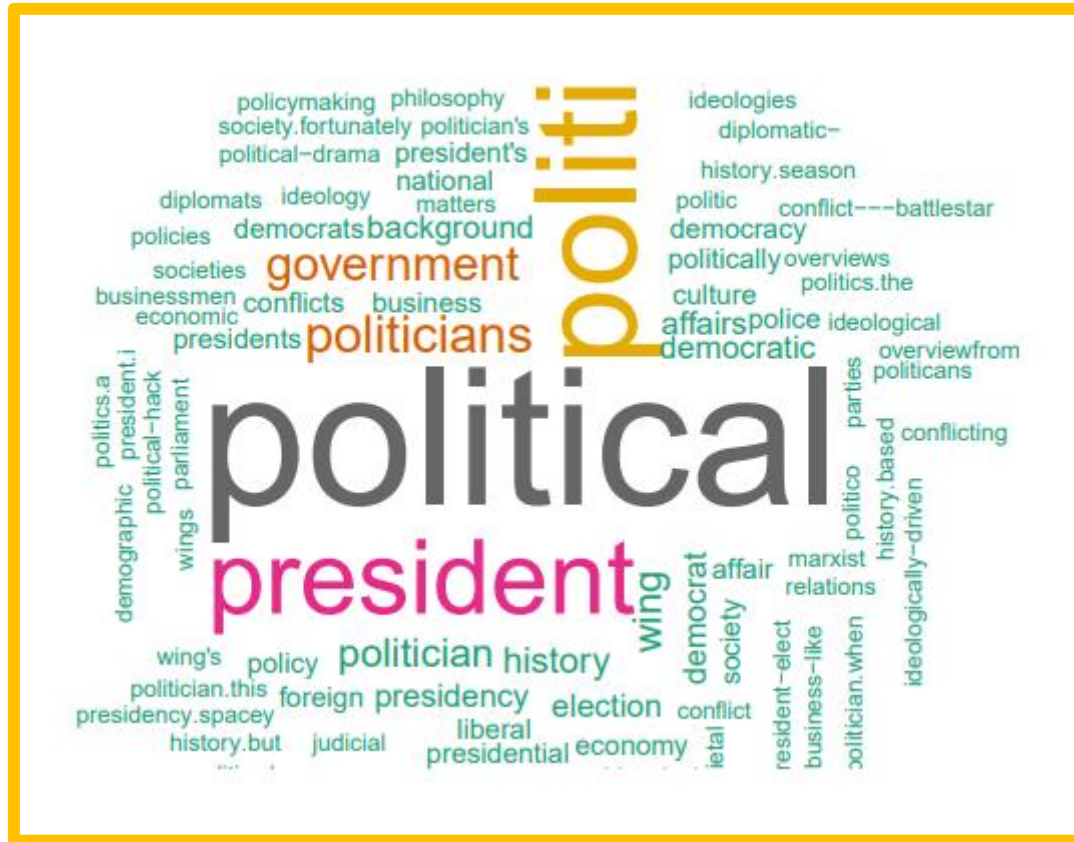


# Results HoC – Vectors



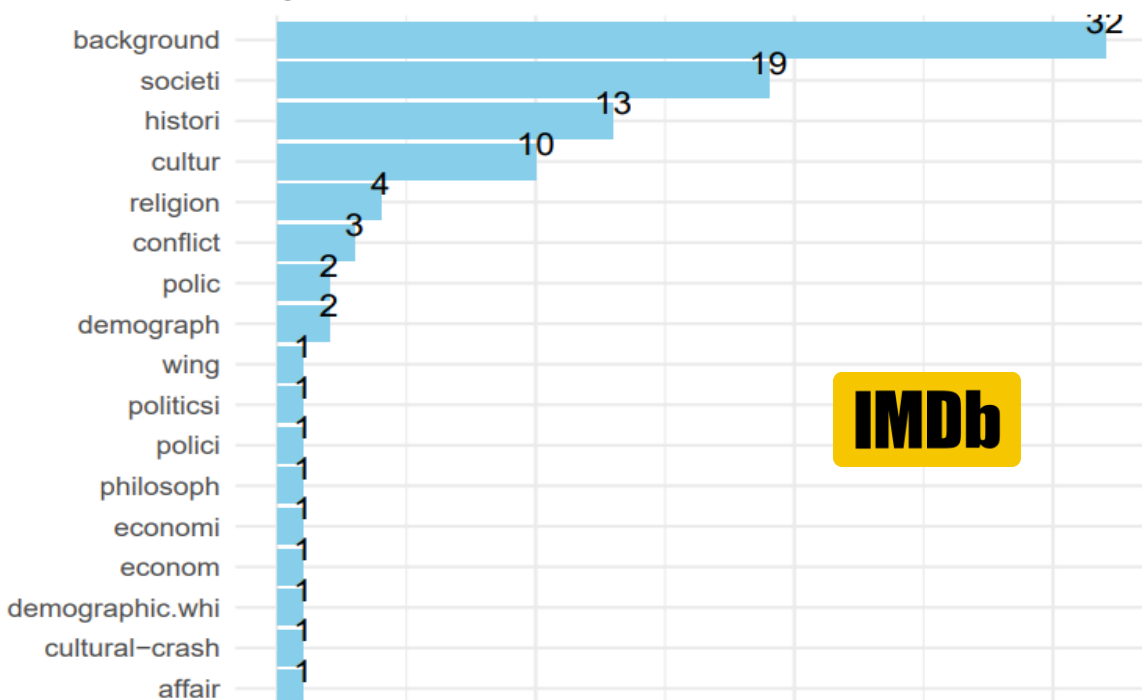
Keywords, HoC, YouTube comments, 100 dimensions





Results HoC  
—  
Wordclouds

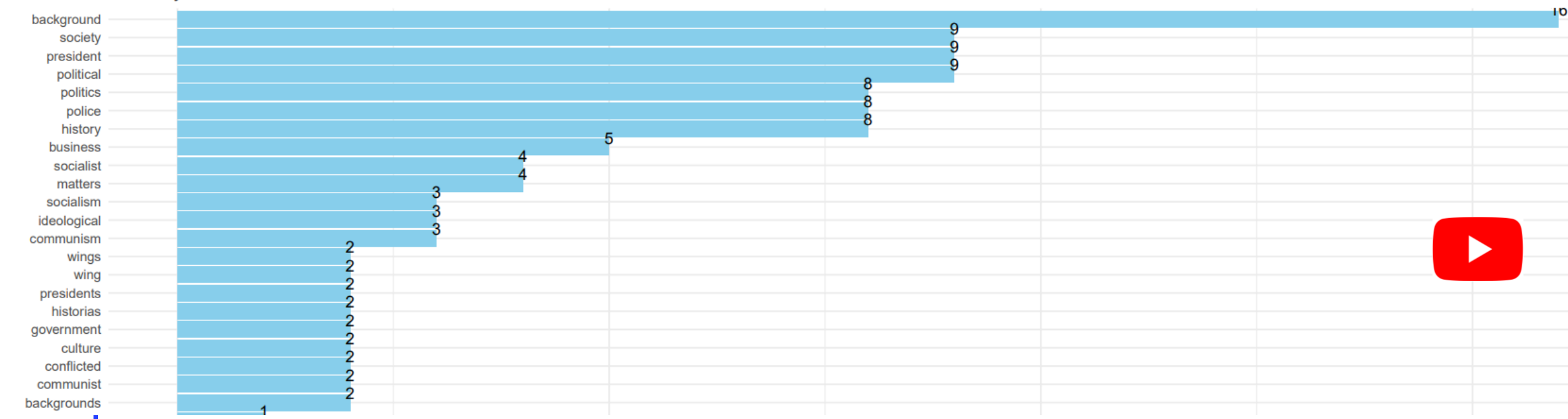
Keyword Counts in IMDB Reviews



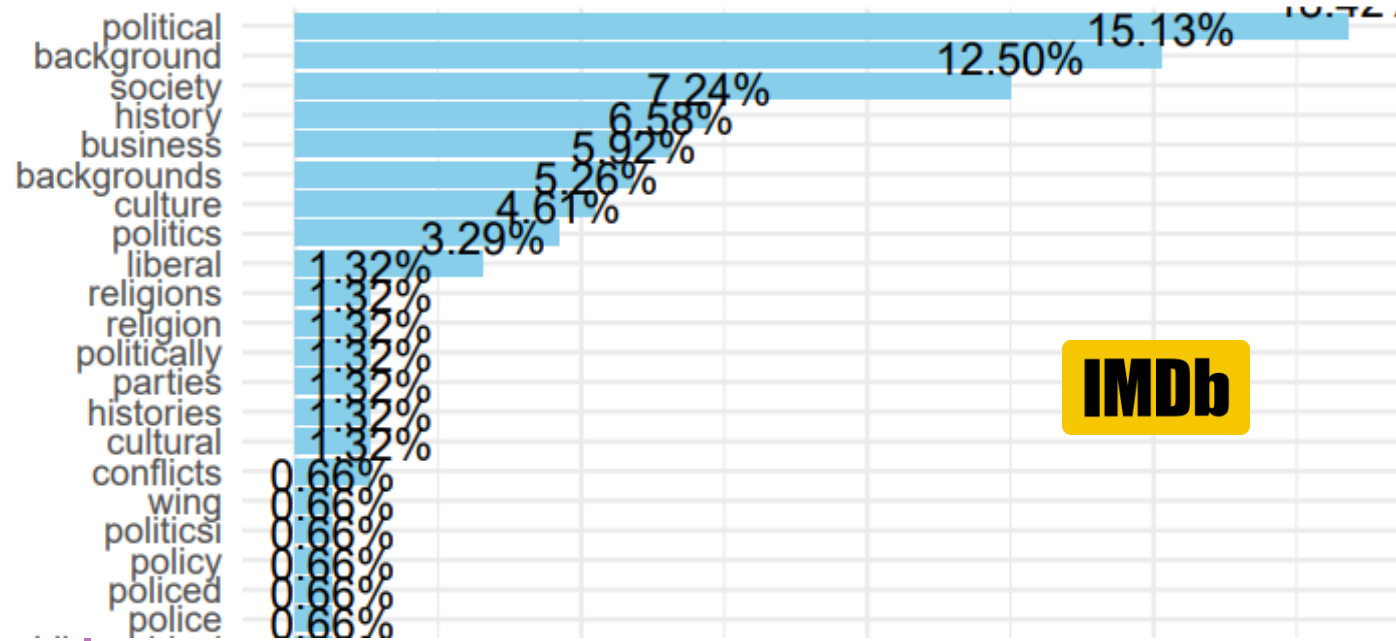
**IMDb**

Results **OitNB** –  
Total keyword  
counts

Total keyword counts, OitNB Trailer 1–6, YouTube



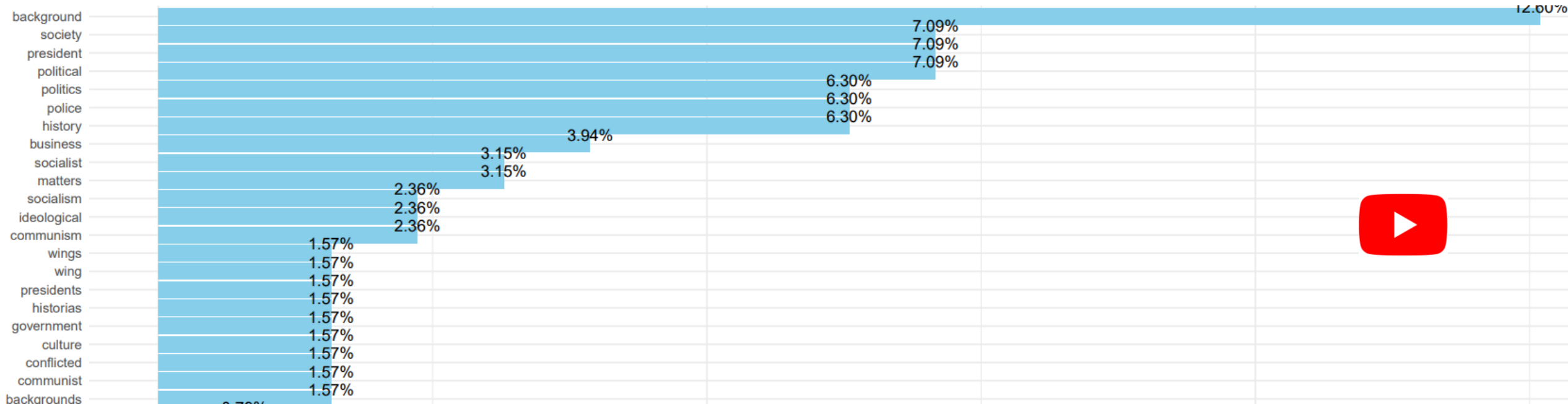
## Relative Frequency of Keywords in IMDB Reviews

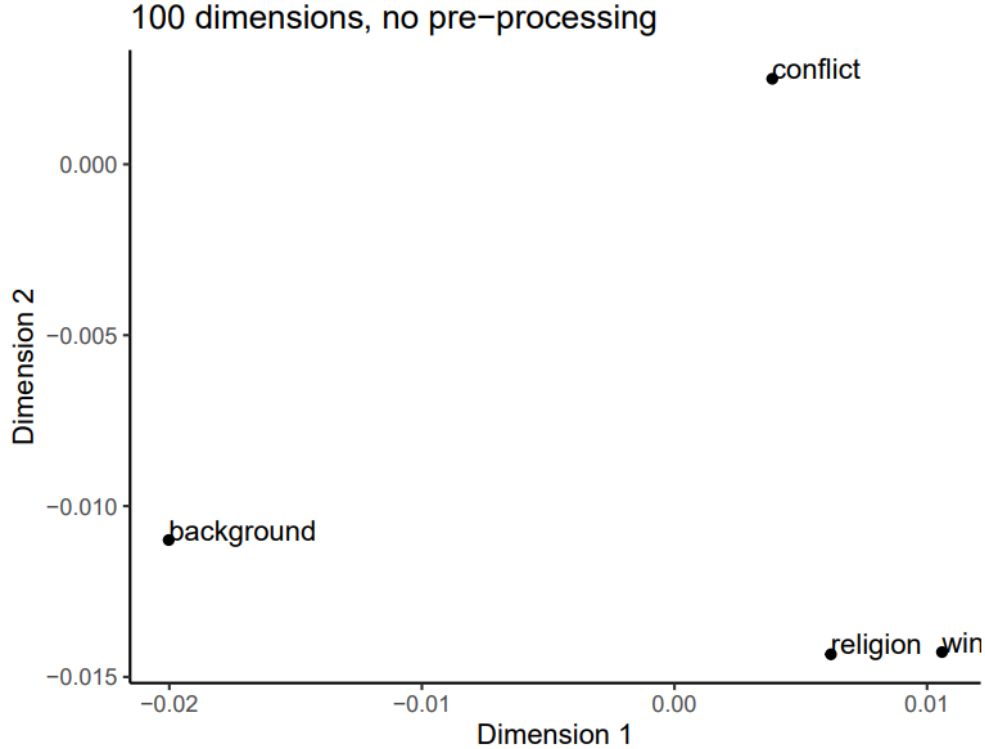


IMDb

Results •  
OitNB – ○  
Relative  
keyword  
counts

Relative keyword counts, OitNB Trailer 1-6, YouTube



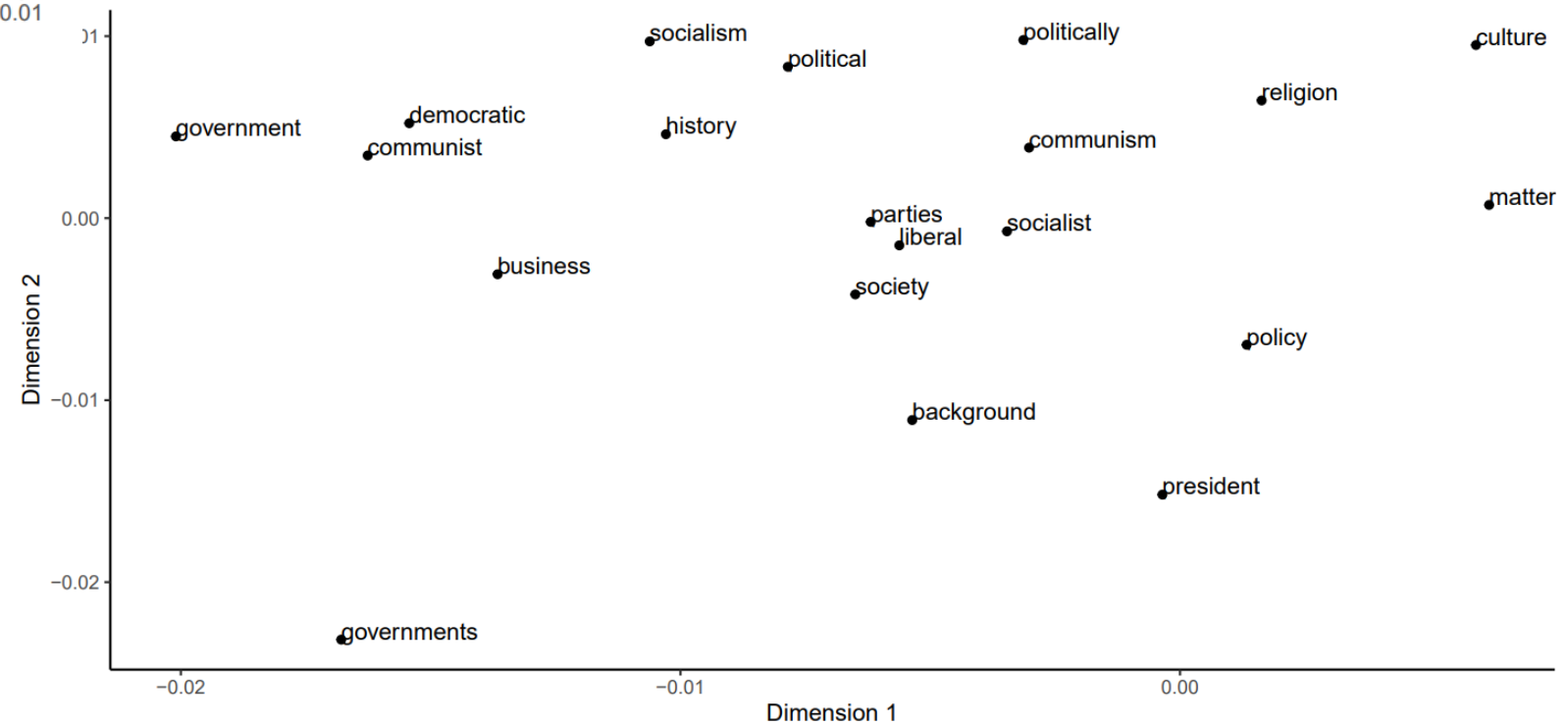


# Results

## OitNB – Vectors



Keywords, OitNB, YouTube comments, 100 dimensions

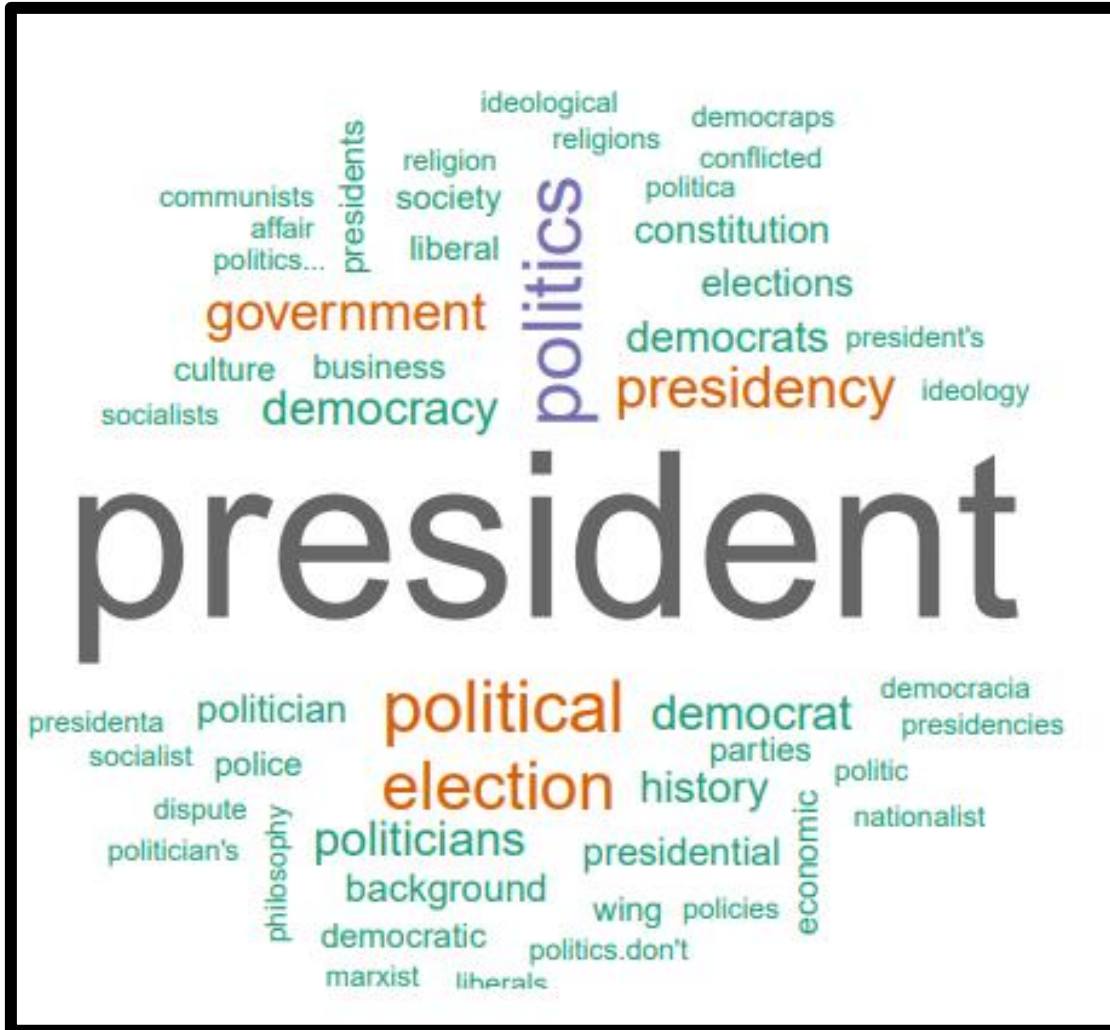




# Comparison HoC vs OitNB – Wordclouds

HOUSE  of CARDS

|ORANGE|  
is the new BLACK|







# DISCUSSION



# Our thoughts



- Both data sources (IMDb & YT) generally show slight differences, but also strong overlaps in comparison
- Political keywords can indeed be found in non-political series (therefore, probably also in book or movie reviews)
- Current analysis does not really answer the question on how people share political opinions (additional context-based analyses needed)
- Legal restrictions can be circumvented, but the cost-benefit ratio (given the legal dimension and the technical adjustments needed) does not seem to justify it, especially when other resources are available
- BUT: YouTube comments tend to be much shorter and often less informative or elaborate than reviews on IMDb



A vertical bar on the left side of the slide, transitioning from orange at the top to purple at the bottom.

# Your thoughts?

+

•

○

+



○



●



# THANK YOU

All scripts and data sets available, just message us  
on Slack 😊