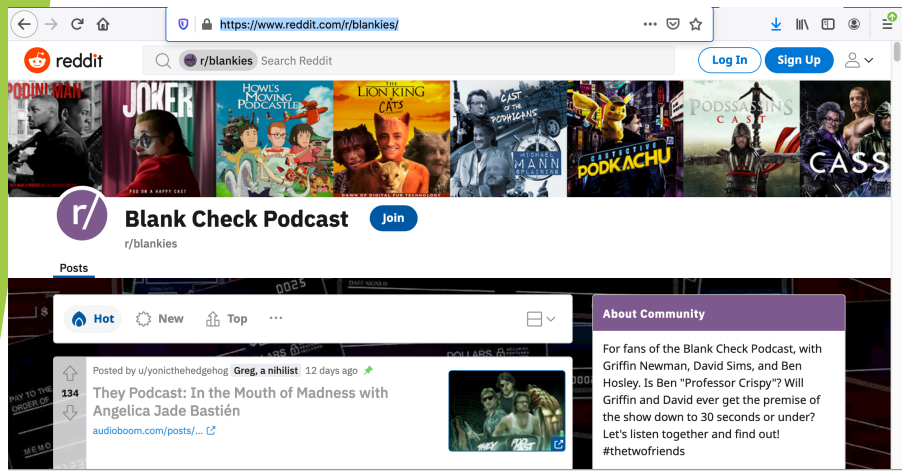


Finding Topics in a Podcast Subreddit

Background

- Podcast: Blank Check with Griffin & David
  - “Blank Check reviews directors' complete filmographies episode to episode. Specifically, the auteurs whose early successes afforded them the rare ‘blank check’ from Hollywood to produce passion projects.”
- Subreddit: www.reddit.com/r/blankies/
  - Created June 2016
- Do members talk about other topics?

Data

Reddit Posts

- Single phrase titles
- Multi-paragraph post
- Ads

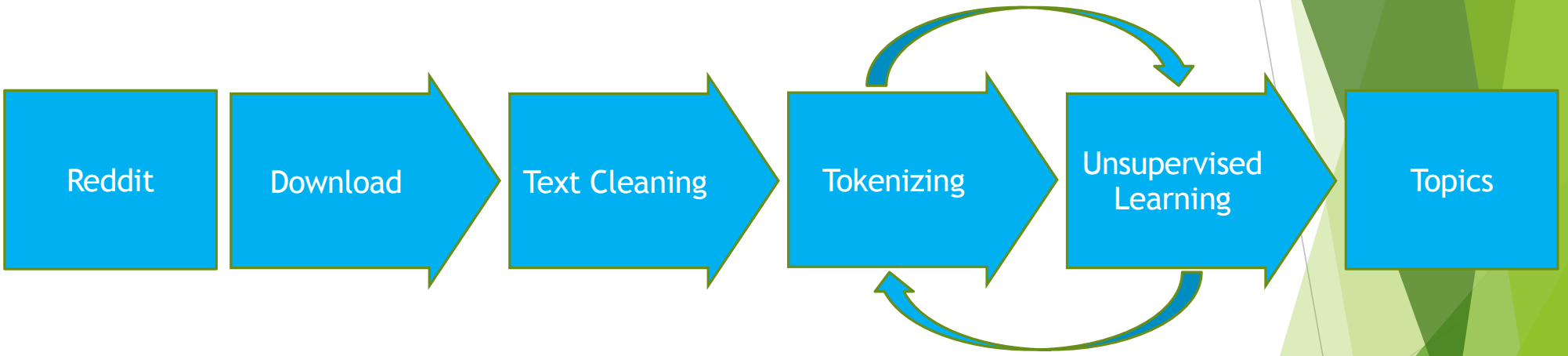
- “Upvotes”
  - Proxy to check if real post

Data

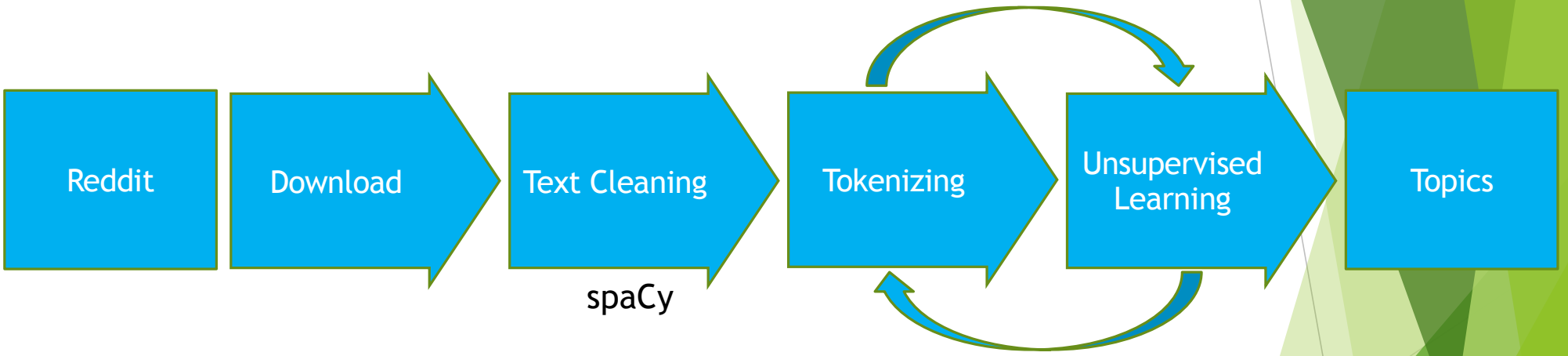
- Upvote of 18 or higher
- 2070 posts

Design

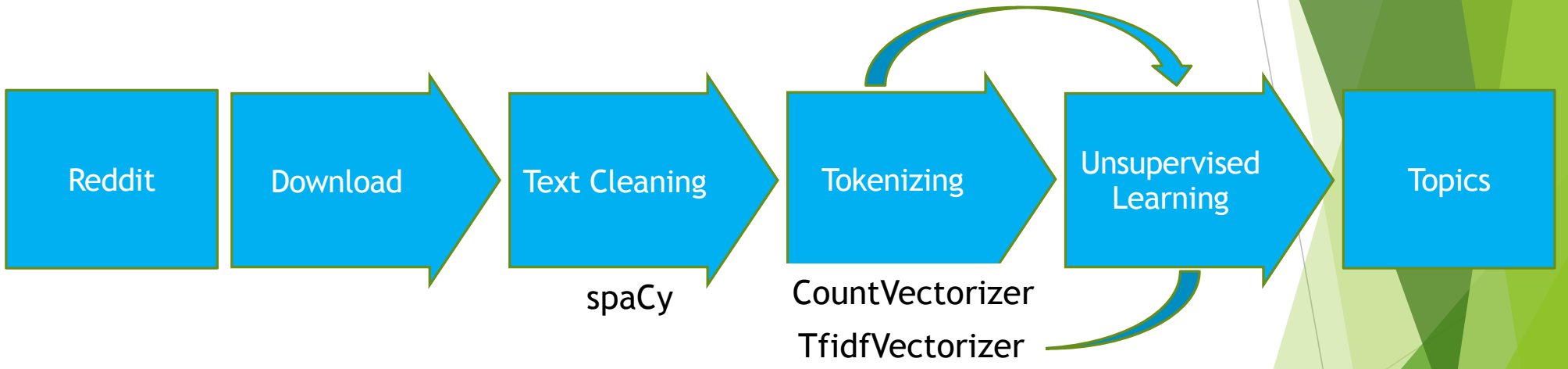
Workflow


Reddit
Download
Text Cleaning
Tokenizing
Unsupervised Learning
Topics

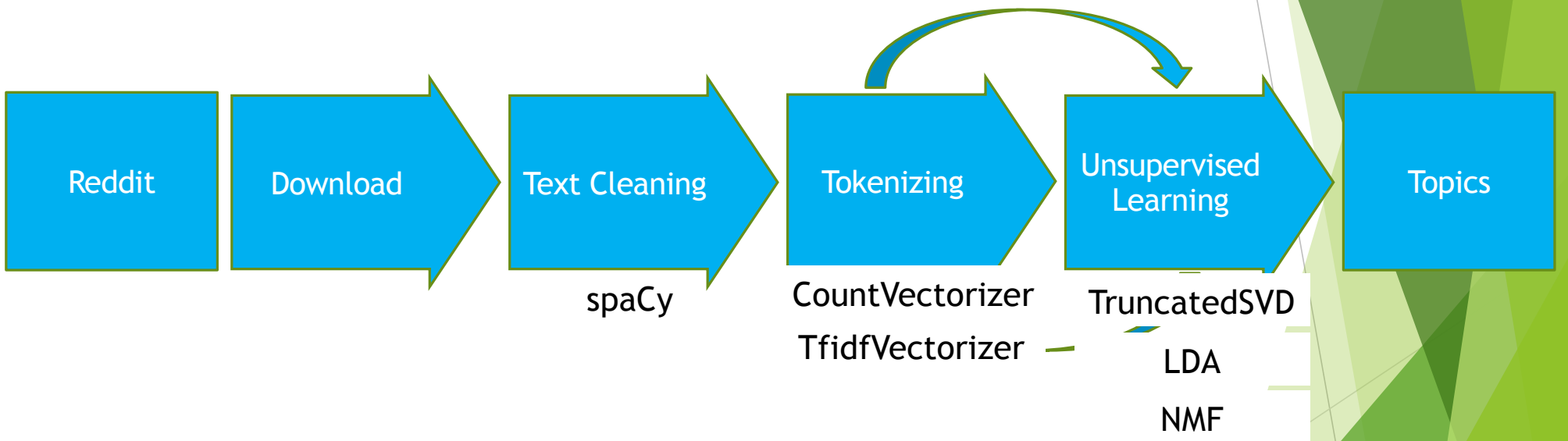
Workflow


Reddit
Download
Text Cleaning
spaCy
Tokenizing
Unsupervised Learning
Topics

Workflow


Reddit
Download
Text Cleaning
spaCy
Tokenizing
CountVectorizer
TfidfVectorizer
Unsupervised Learning
Topics

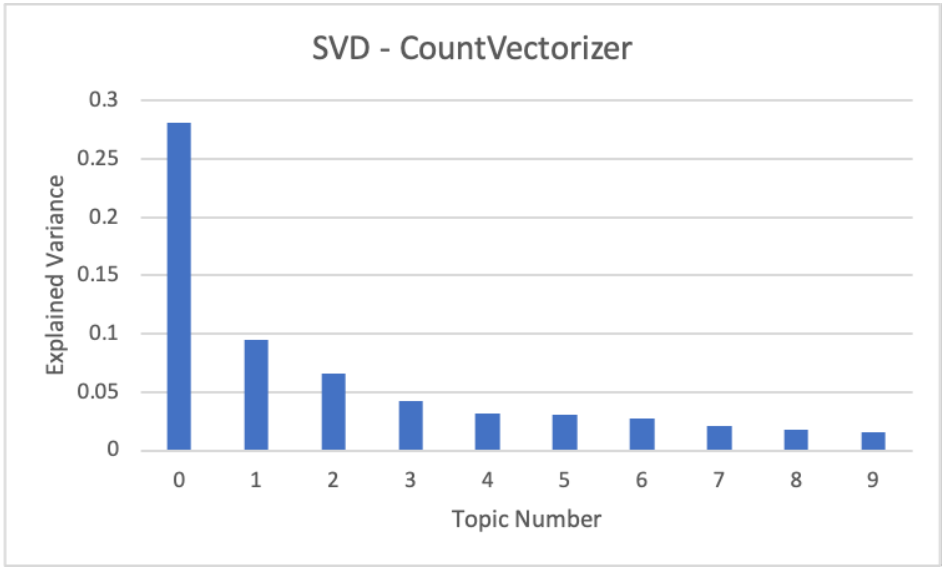
Workflow


Reddit
Download
Text Cleaning
spaCy
Tokenizing
CountVectorizer
TfidfVectorizer
Unsupervised Learning
TruncatedSVD
LDA
NMF
Topics

Modeling


THE MAGNIFICENT
MR. TANGENT
BLANK
CHECK
GRIFFIN
NEWMAN
&
DAVID
ISMS
WHY SO SERIOUS?

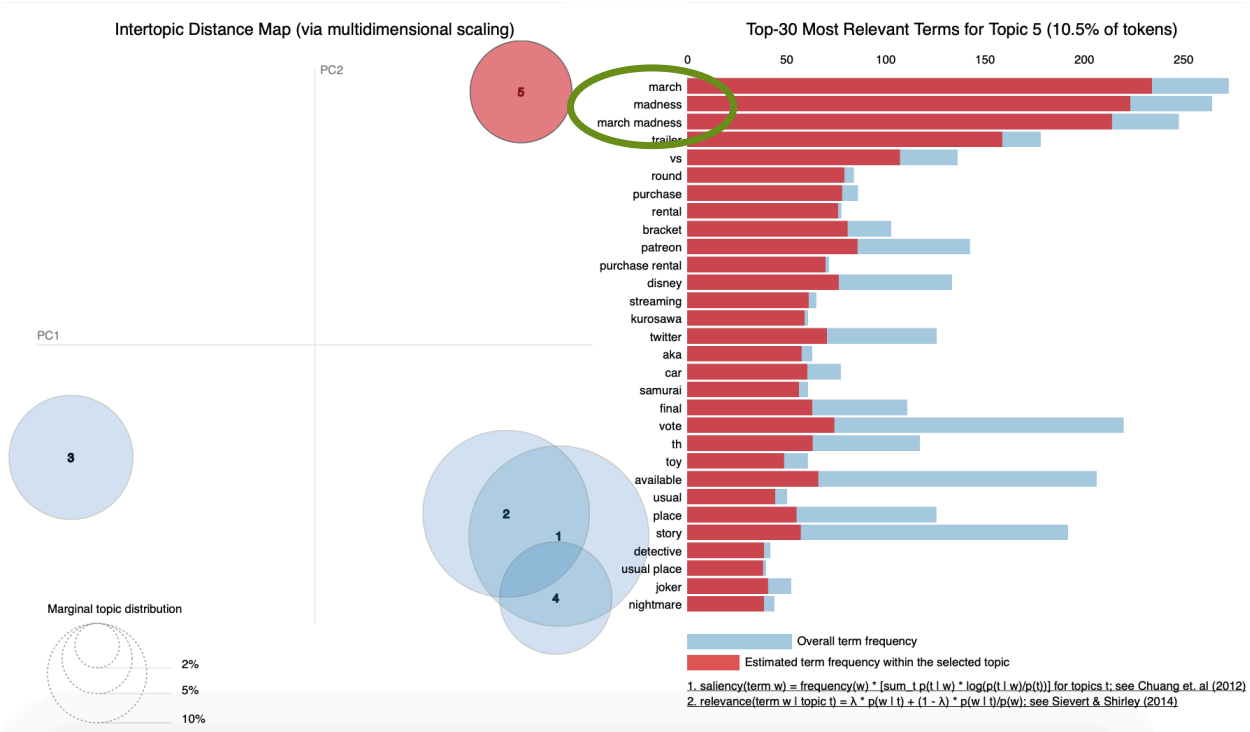
Results – Count Vectorizer



Topic 0	game, watch, youtube, youtube watch, play
Topic 1	feature, director, number, good, number feature
Topic 2	feature, number feature, number, age feature, age
Topic 3	set, brickset, millennium, falcon, brickset set
Topic 4	record, batman, big, director, charlie
Topic 5	record, batman, charlie, watch, big
Topic 6	star, good, war, star war, record
Topic 7	people, thing, way, think, star
Topic 8	lion, king, lion king, new, date
Topic 9	purchase, rental, purchase rental, available, place

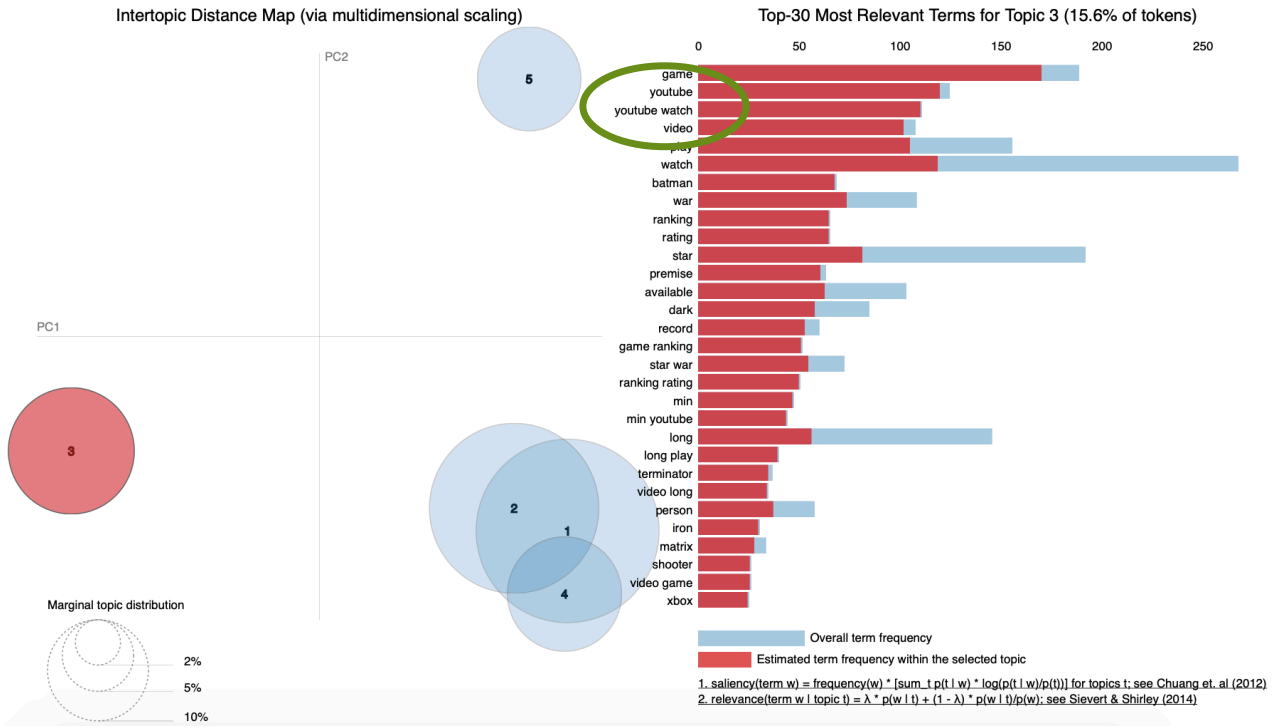
CountVectorizer-SVD

Top 10 Topics by Variance Explained



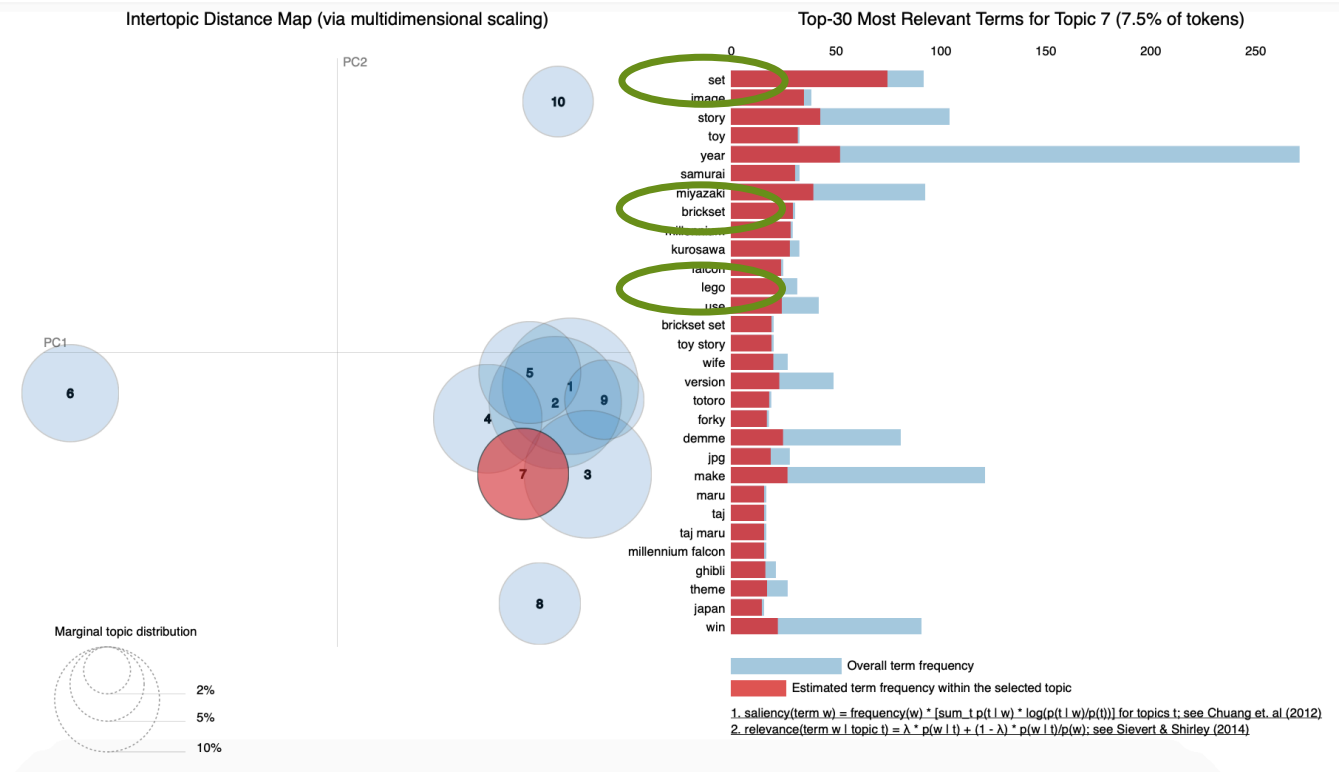
CountVectorizer-LDA Five Topics

"March Madness"



CountVectorizer-LDA Five Topics

"YouTube"



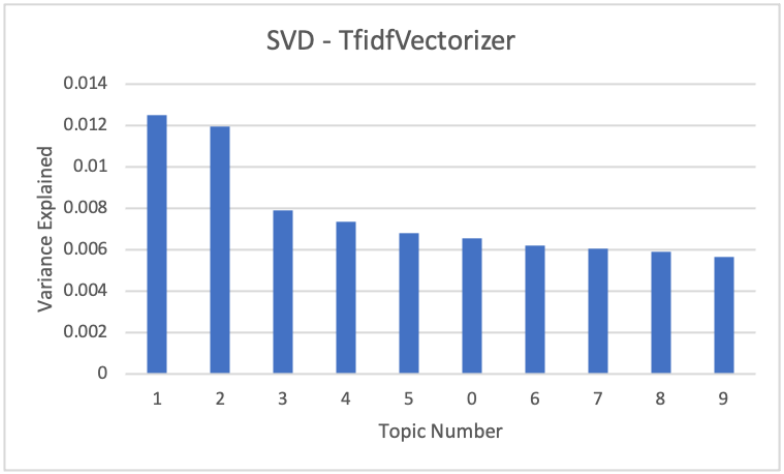
CountVectorizer-LDA Ten Topics

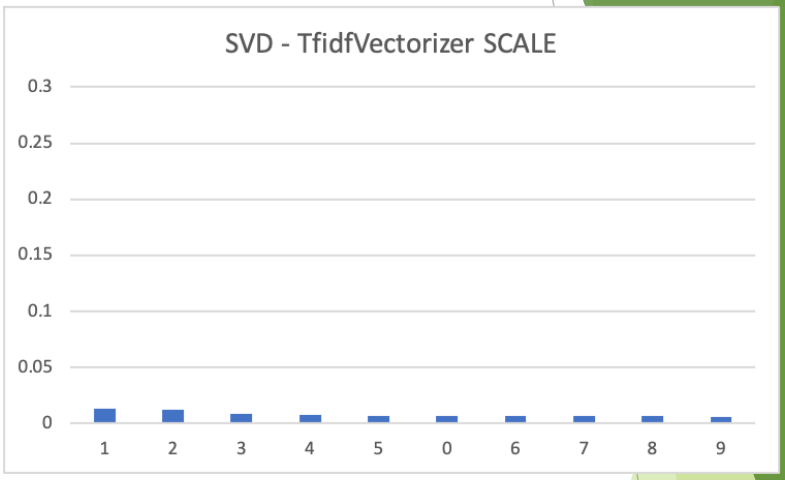
"Lego"

CountVectorizer-NMF

Topics 0 and 3, of 5

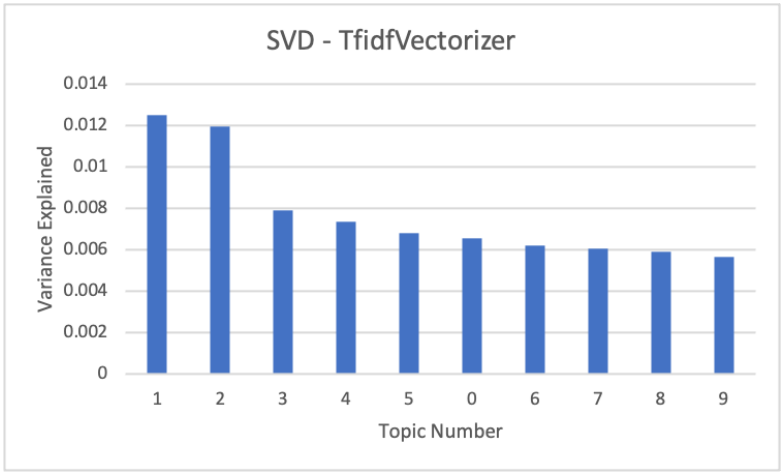
Results – TfidfVectorizer





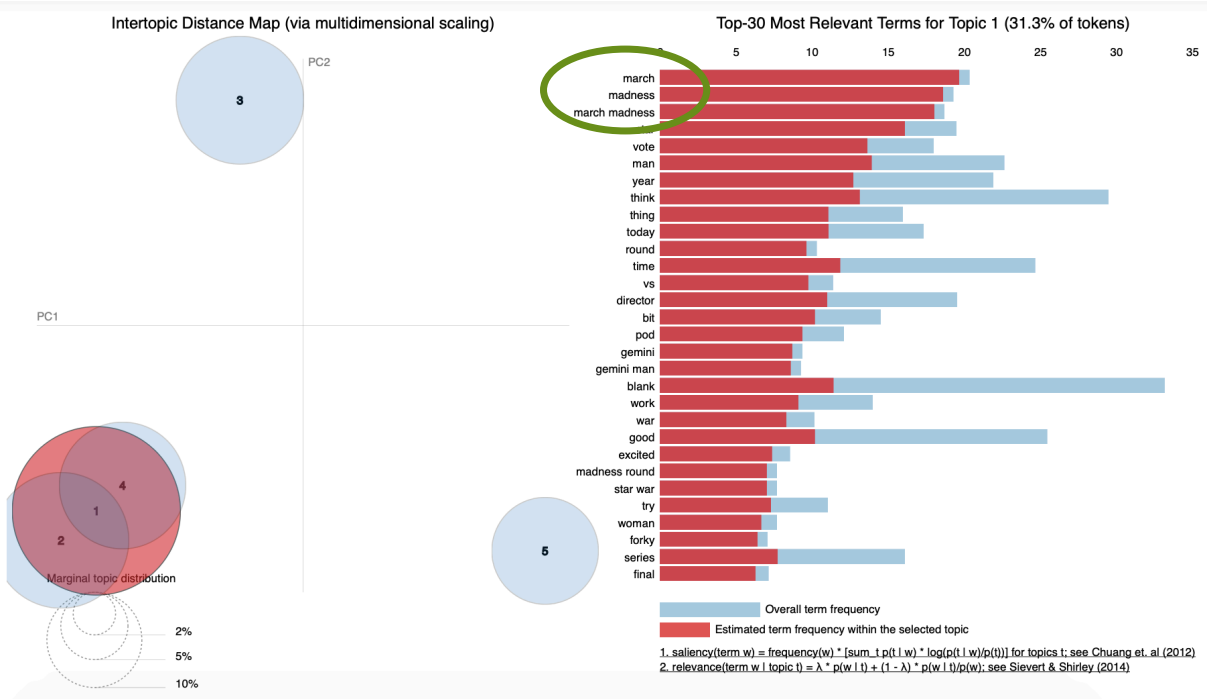
TfidfVectorizer-SVD

Much less variance explained



Topic 0	blank, think, march, madness, march madness
Topic 1	march, madness, march madness, round, vs
Topic 2	trailer, march madness, march, madness, round
Topic 3	man, gemini, gemini man, sim, look
Topic 4	blank, man, gemini, gemini man, sim
Topic 5	new, look, joker, star, director
Topic 6	vote, new, blankie, today, blank
Topic 7	vote, look, car, today, cast
Topic 8	look, blankie, think, new, blank
Topic 9	cast, mann, michael, michael mann, pod

TfidfVectorizer-SVD



TfidfVectorizer-LDA – 5 topics

"March Madness"

TfidfVectorizer-NMF

Topics 1 and 3, of 5

Summary


BLANK
CHECK
GRIFFIN
NEWMAN
WITH
&
DAVID
SIMS
A PAUL VERHOEVEN PODCAST
PODSHIP
CASTERS

Recap

- Used both CountVectorizer and TfidfVectorizer
- Used SVD, LDA, and NMF with both tokenizations
- SVD and NMF found “Lego” and “YouTube” when tokenized with CountVectorizer
- LDA found “March Madness” both when tokenized with CountVectorizer and when tokenized with TfidfVectorizer
- No one topic was found by all six combinations

Takeaway

- Members of the subreddit mostly stay on the topic of the podcast

Cautions

- Topics were too sparse for TfidfVectorizer
- Sample size was small

Future Work

- Find top *n* bigrams for more in-depth topic analysis
- Try other subreddits
  - Do most stay on topic the way this one does?

Thank you!

Sources

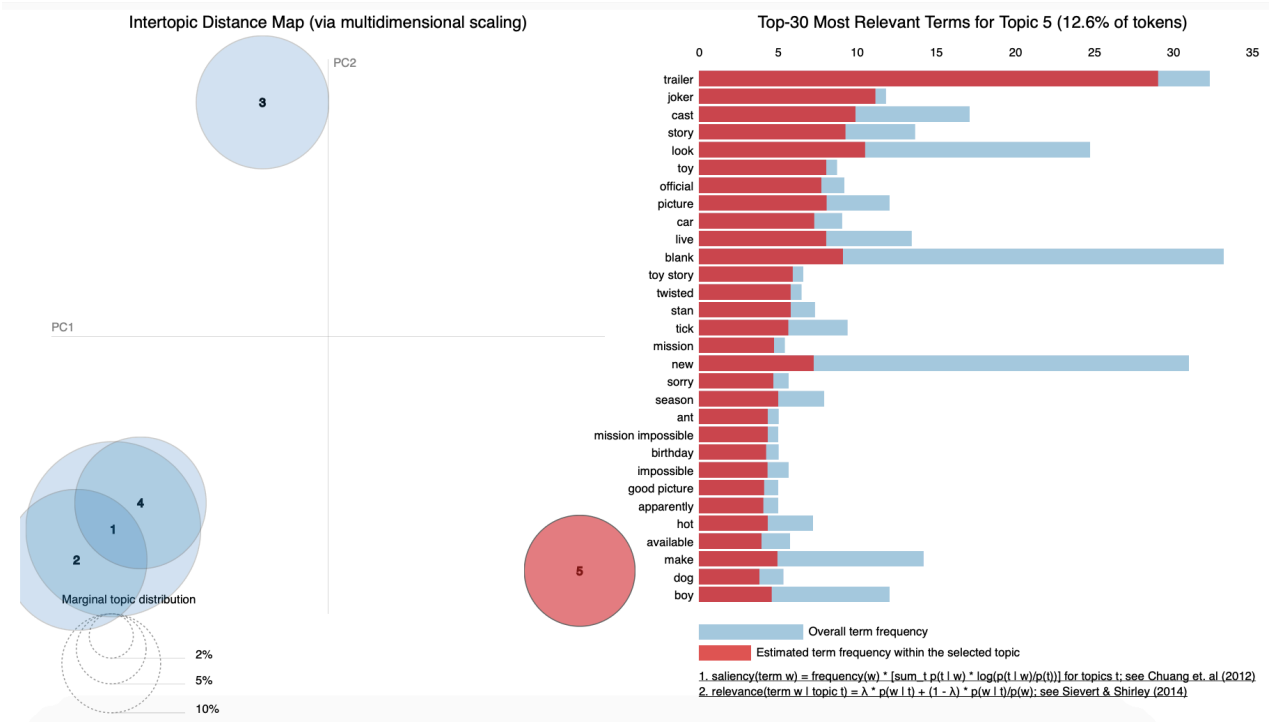
- https://audioboom.com (images)
- https://www.reddit.com/r/blankies/ (image; data)

Appendix

Tokenizers

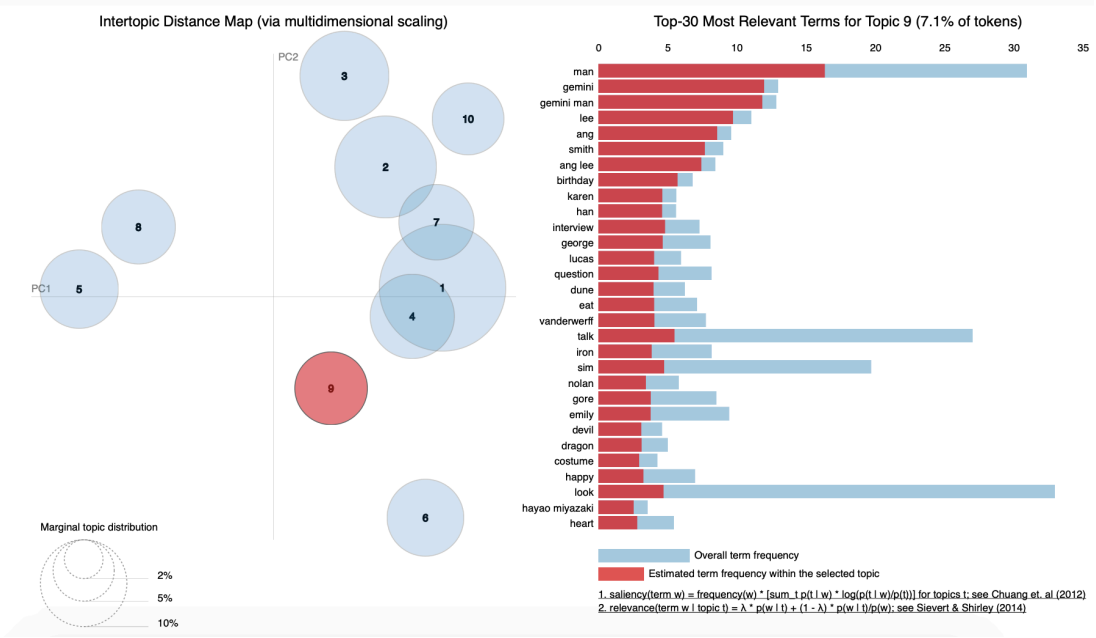
- ngram: 1- and 2-grams
- max_features: 1000

- CountVectorizer: max_df = .6
- TfidfVectorizer: max_df = .75



TfidfVectorizer-LDA 5 Topics

??????



TfidfVectorizer-LDA Ten Topics

"Gemini Man"

Word Cloud – No Processing