LIVE

BREAKING NEWS:

Anime Recommender from TV Show Input

Daniel Wang presents a Capstone Project to General Assembly
DSI 123 on April 17, 2023

TABLE OF CONTENTS

LIVE

01

PROBLEM STATEMENT

IF THERE WAS A PROBLEM, YO, I'LL SOLVE IT.
CHECK OUT THE HOOK WHILE MY DJ REVOLVES IT.

ANIME RECOMMENDATIONS FROM TV SHOWS!

(The original slidedeck actually came with this already here. o_0)

WHOA

There are really people out there who don’t like anime??

GIVE ANIME A CHANCE!

- Yes, it's true: There ARE many people out there who don't like anime.
- But what about the artistic merits?
- Okay, IF someone were willing to give anime a chance, how do we make that attempt as smooth as possible?

JUST YOUR AVERAGE HATER...

Paulie/Polly Puppet ***(They/Them)***

“I don’t like anime. I do not like green eggs and ham.”

IN DEPTH

ARTISTIC MERITS OF ANIME

- Storied cultural history.
- Animated nature allows for fuller range of emotional display
- Can depict stories and situations (until recently) not possible with human actors.
- Wide variety of genres within “anime” umbrella; some even depicting realistic scenarios without overly cartoonish style. (Slice-of-life)

OPERATION WEEB-IFY (A PLAN OF ATTACK)

DEFINE THE PROBLEM

COLLECT & CLEAN DATA

Gather the data on shows (for input & output) that we could use to make recommendations on.

ANALYZE

BUILD RECOMMENDER

Use a combination of Data Science techniques to build a recommender engine.

??? (PROFIT)

Deploy our recommender and hope people take our suggestions into consideration.

02

LIVE

THE DATA

The data that powers the magic. Where and how did we get it?

DATA SOURCES

~~myanimelist.net~~

Good for anime, but any directly comparable info for live-action TV shows would need to come from elsewhere.

IMDb

Even with API, can't get wanted info.
Resort to webscraping.
Useful for top shows lists by genre.

Wikipedia

Publicly available API, but entries are not in standardized section formats.
Resorted to extracting whole pages.

TheMovieDB

Wish I'd just started here. This had nearly everything I needed.

FINAL DATA USED

IMDb

Name, href (URL for tv show pages), years, description, PG rating, genre tags, imdb rating, # votes, img thumbnail URL

TheMovieDB

Name, original name, original language, origin country, popularity, vote average, vote count, first air date, adult content (Y/N), poster img URL, (show) overview, tagline, genres, TV networks, keywords

Wikipedia

Returned search term, search URL, text of article

ALSO IN THE NEWS:...ATER STILL WET. PIGS STILL CAN'T FLY. COME ON SCIEN...

HOW DO WE KNOW IT'S ANIME?

CRITERIA

```
a = df['tmdb_keywords'].str.contains("anime") #191
b = df['origin_country'].str.contains("JP") #184
c = df['imdb_genre_tags'].str.contains("Animation") #481
d = df['wiki_text'].str.contains("anime")  #261
```

MY FINAL CRITERIA...

```
df.loc[a&c, 'is_anime'] = 1
```

(having tmdb_keyword of 'anime' & imdb_tag of 'Animation')

03

LIVE

MODELING

LDA/LSI learning model to generate similarities.
Also, simpler models.

DATA SCIENCE TECHNIQUES/TERMS

Concepts Used

Topic modeling	Finding natural groups/clusters of articles/documents/text
Unsupervised Learning/Classification	When you don't give the computer any 'labels' to train on
Latent Dirichlet Allocation (LDA)	Probabilistic model that infers a hidden topic structure
Latent Semantic Indexing (LSI)	Model based on SVD (Singular Value decomposition), principal component analysis
Saliency	Tells which WORDS are most impactful for identying topics of the documents
Perplexity	Measures how good the MODEL predicts words in a document/set

ALSO IN THE NEWS:
...ATER STILL WET. PIGS STILL CAN'T FLY. COME ON SCIEN...

For more info:
SLIDESGO | SLIDESGO SCHOOL | FAQs

LDA vs. LSI

(Based on different mathematical models)

Latent Dirichlet Allocation (LDA)

- Based on a probabilistic generative model
- Goal: discover main topics and distribution of words within those topics.
- Supposedly better for “messy”/noisy data
- Less prone to overfitting
- More interpretable topics(?)

Latent Semantic Indexing (LSI)

- Based on SVD (Singular Value decomposition), principal component analysis
- Can directly output similarity
- More computationally efficient (FAST!)

MODEL VARIATIONS

FORCED # TOPICS

And multiple re-runs

SUBSETS OF DATA

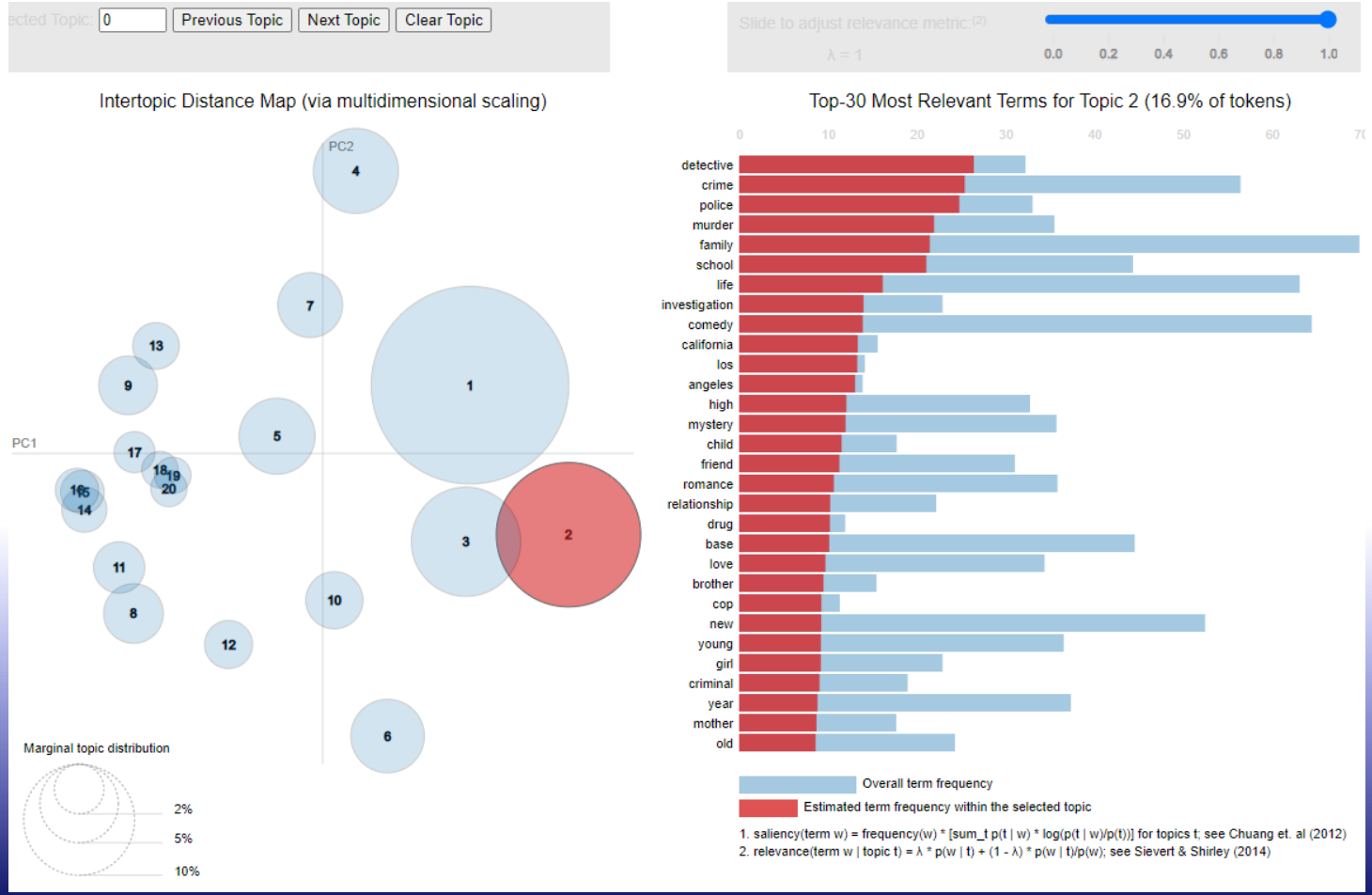
With/without certain data columns (e.g., ALL of Wikipedia)

DIFFERENT MODEL TYPES

LDA, LSI, cosine similarity on keyword only


A PICTURE IS WORTH A
THOUSAND WORDS

pyLDAvis


ected Topic
0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (16.9% of tokens)
0
10
20
30
40
50
60
detective
crime
police
murder
family
school
life
investigation
comedy
california
los
angeles
high
mystery
child
friend
romance
relationship
drug
base
love
brother
cop
new
young
girl
criminal
year
mother
old
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

RESULTS EXPLORATION

LDA results in Jupyter Notebook via pyLDAvis

(*Alt-Tab here*)

04

LIVE

RECOMMENDER

Demonstrating our recommender in action*

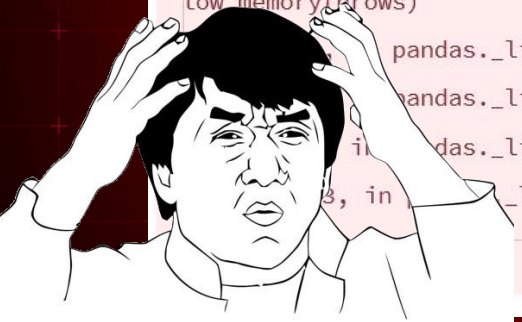
EPIC FAIL. *RAGE QUIT*

localhost:8502

ost Visited fb YouTube GMaps Gmail Y! Mail Yelp Wikipedia OneLook Zhongwen

```
ffer, kwds)
kages\pandas\io\parsers\readers.py", line 611, in _read

kages\pandas\io\parsers\readers.py", line 1778, in read
ype: ignore[attr-defined]
kages\pandas\io\parsers\c_parser_wrapper.py", line 230, in read
low_memory(nrows)
                pandas._libs.parsers.TextReader.read_low_memory
                andas._libs.parsers.TextReader._read_rows
            i   das._libs.parsers.TextReader._tokenize_rows
3, in           _libs.parsers.raise_parser_error
```



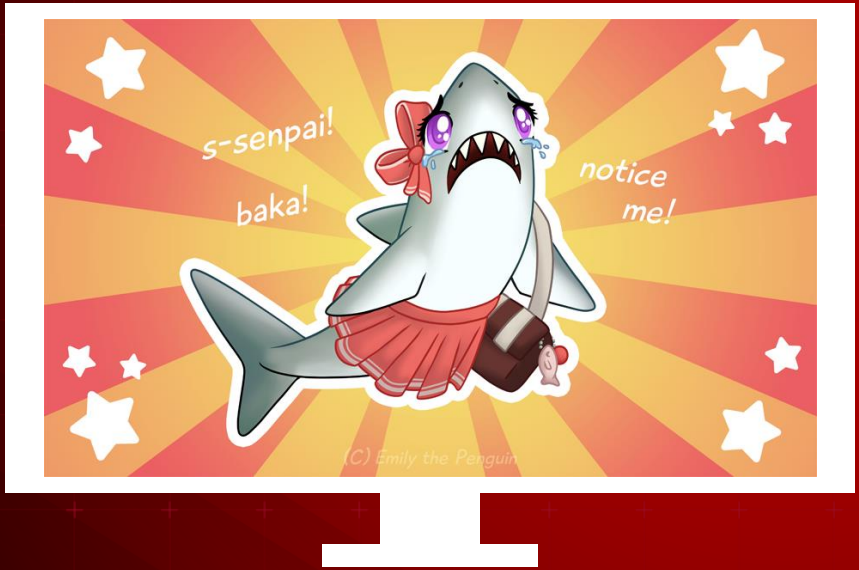
Streamlet.io

You had ONE job...

You failed me!

JUPYTER NOTEBOOK?!?

(yes, really)

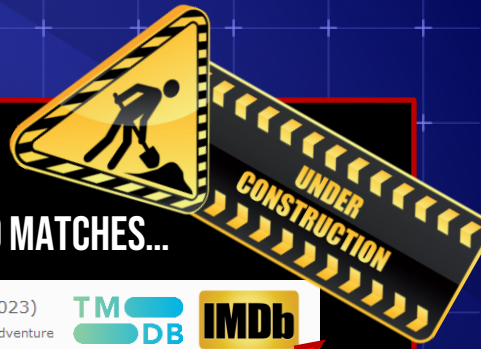


In lieu of a demonstration of a functional Streamlit app,

...a demonstration inside Jupyter notebook!

(*Alt-Tab here*)

How it SHOULD have looked...



YOU INPUT: “ GAME OF THREES ”

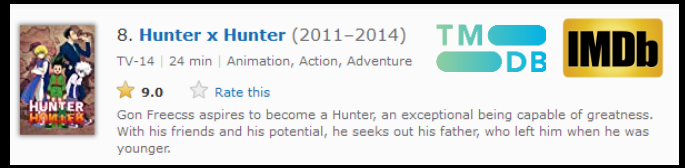
ANIME OUTPUT ONLY? ☑

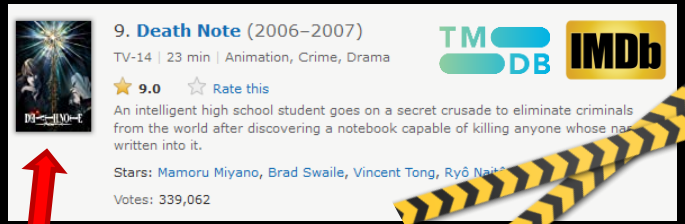
NUMBER OF OUTPUT ENTRIES (DEFAULT: 5) 3

NO EXACT MATCHES FOR INPUT “ GAME OF THREES ”.
1 RESULTS STARTING THE SAME FOUND. DID YOU MEAN:

NAME	YEARS	HREF
“GAME OF THRONES	(2011-2019)	/TITLE/TT0944947/”

OUR TOP RECOMMENDED MATCHES...





05

LIVE

AWESOME NEWS* (*NOT)

...ROOM FOR IMPROVEMENT

A lot of meat left on the bone.

What else could we do?

AREAS FOR IMPROVEMENT

1. TUNE WHICH INPUTS

Final models all exclude Wikipedia data. (Too much unhelpful garbage data.)

2. MORE OUTPUT FILTERS

I personally disagree with some of the recommendations.

Try: no shows under ___ score; no shows under ___ reviews

3. FUNCTIONING STREAMLIT

Without a public-facing tool, available to the masses– what was it all for? =(

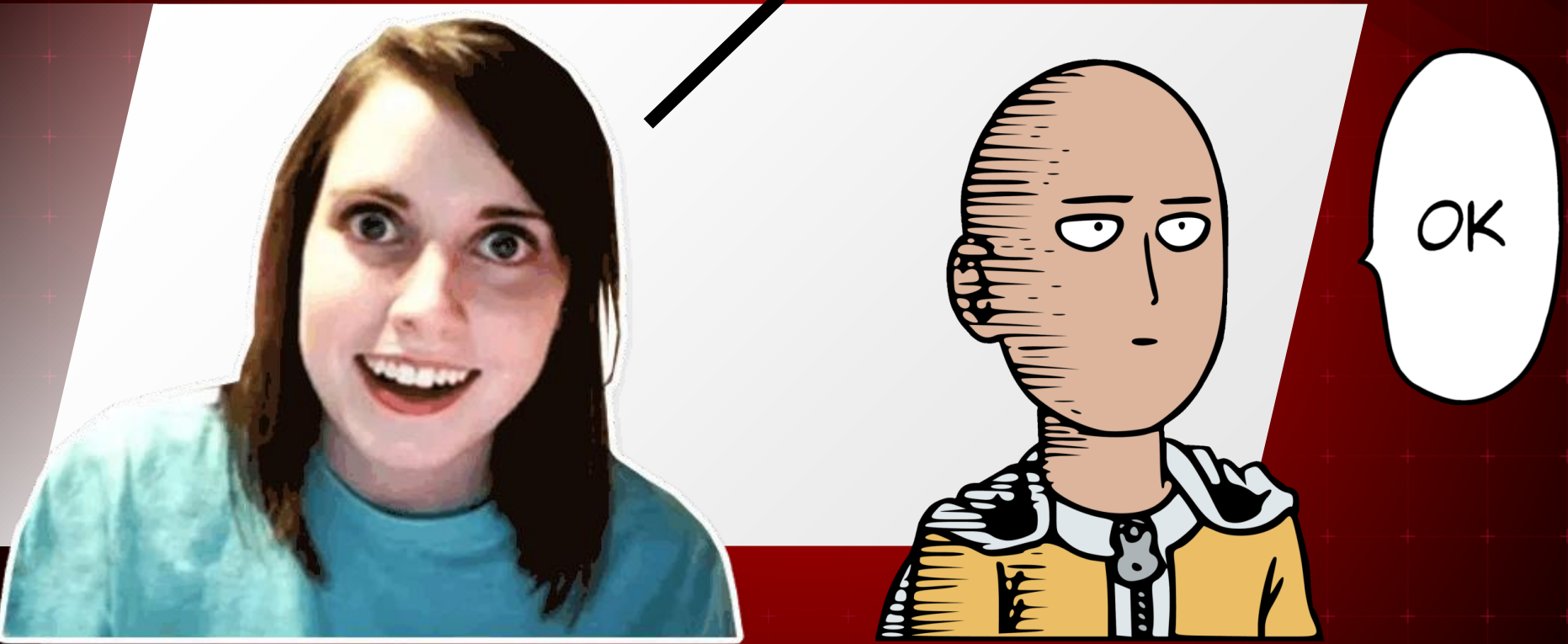
4. JUST USE SQL QUERY

06

CONCLUSION

Thanks for your attention- let’s wrap this up!

So, you're going to try anime, right?



“You miss 100% of the shots you don’t take.”

~ Wayne Gretzky

—MICHAEL SCOTT

“Fail Fast. Fail Forward”

YOLO. For Science.

RESOURCES

Useful webpages I referenced:

Data Sources

- IMDb - www.imdb.com
- Wikipedia - en.wikipedia.org
- TheMovieDB - www.themoviedb.org

Data Science Technique References

- General Assembly DSI Lesson 705-lesson-recommender-systems
- https://www.geeksforgeeks.org/nlp-gensim-tutorial-complete-guide-for-beginners/
- (Jiamei Wang) https://medium.com/swlh/sentiment-analysis-topic-modeling-for-hotel-reviews-6b83653f5b08
- (Mimi Dutta) https://www.analyticsvidhya.com/blog/2021/07/topic-modelling-with-lda-a-hands-on-introduction/
- (Avinash Navlani) https://machinelearninggeek.com/latent-semantic-indexing-using-scikit-learn/
- https://stackoverflow.com/questions/21285380/find-column-whose-name-contains-a-specific-string
- https://www.statology.org/pandas-check-if-column-contains-string/
- (Raiyan Quaium) https://medium.com/@raiyanquaium/how-to-web-scrape-using-beautiful-soup-in-python-without-running-into-http-error-403-554875e5abed

A SPECIAL THANKS TO...

Instructors

Tim, Katie, Loren, Rowan... and Sim too!

...AND VIEWERS LIKE YOU!

Classmates

Class of 2023! WOOoooOOO!!!

Best BFF forevar!

Stay cool. Never change.

K.I.T.! H.A.G.S.!

LIVE

THANKS

Do you have any questions?

wangdj3@gmail.com
+1 469 438 5735
https://github.com/wangdj3/DSI_Capstone/

https://slidesgo.com/theme/tv-news-report-digital-newsletter