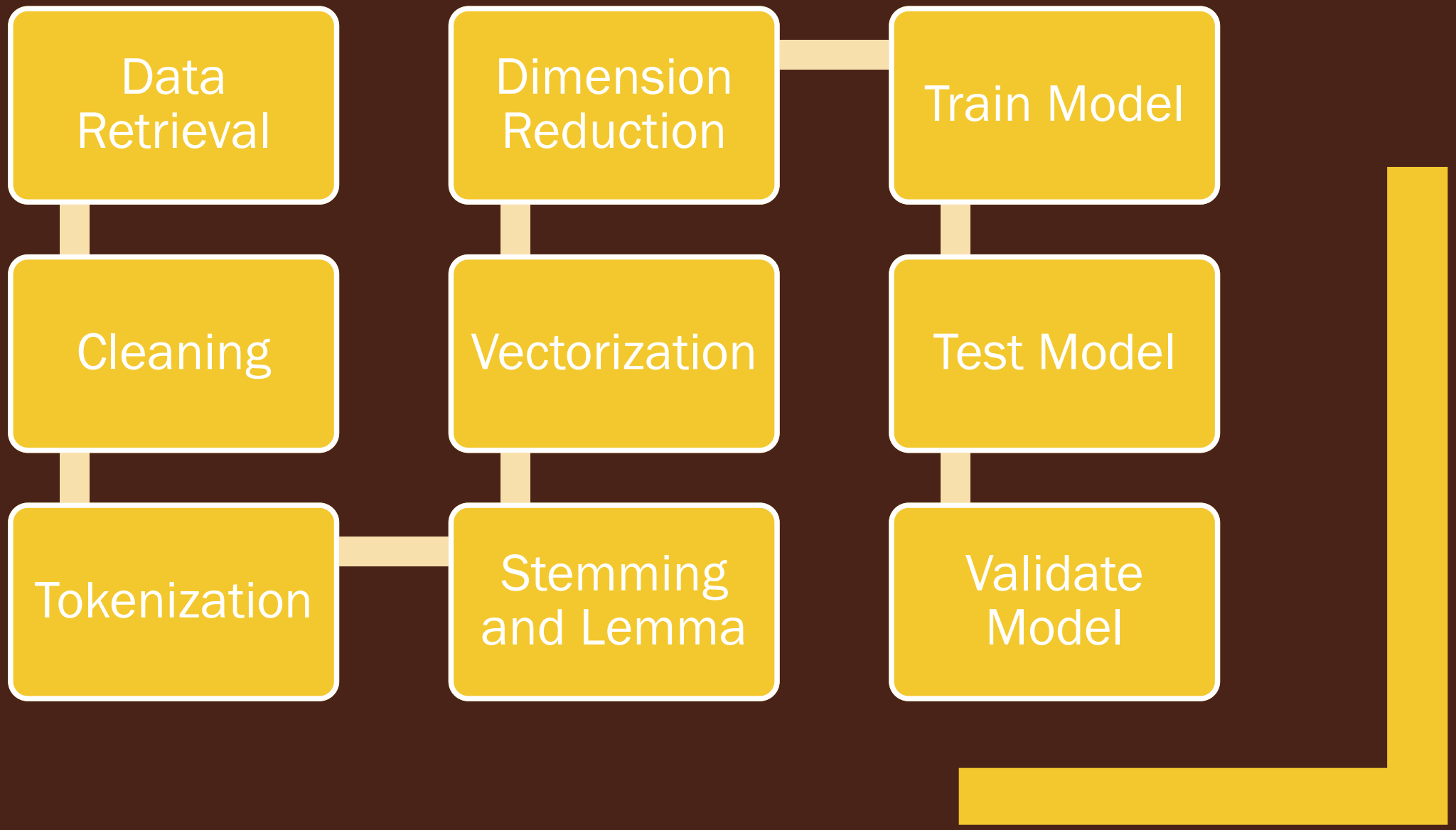
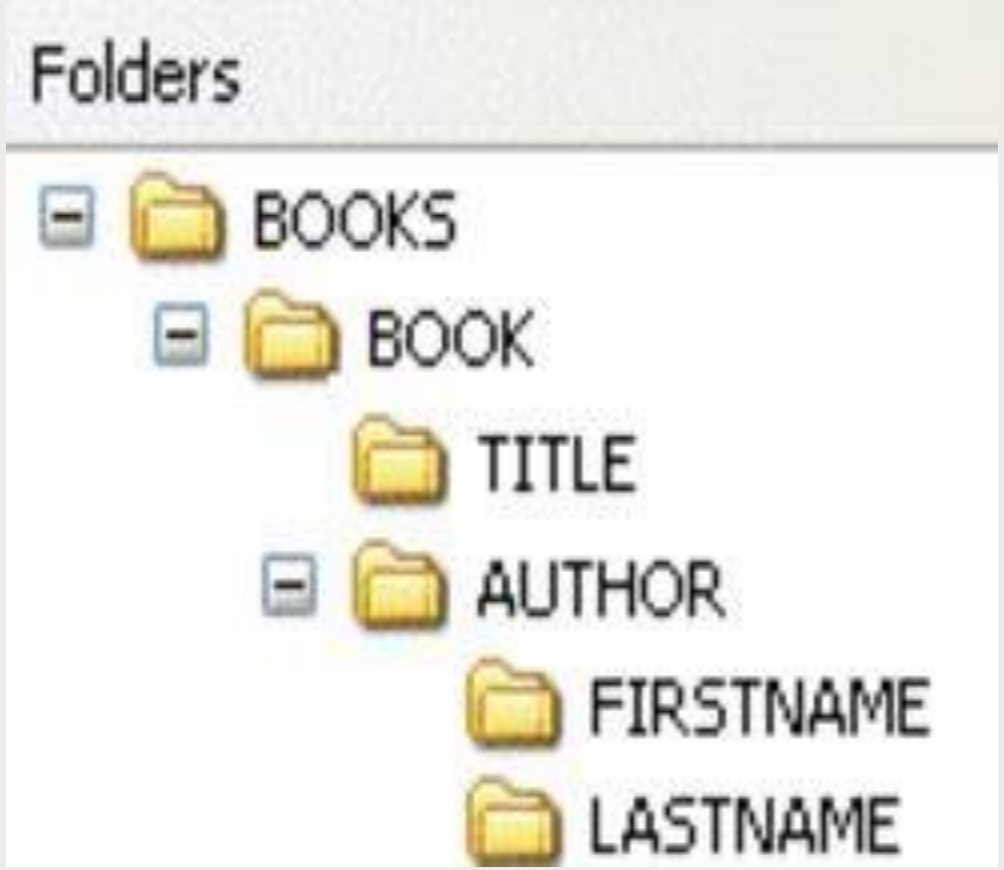
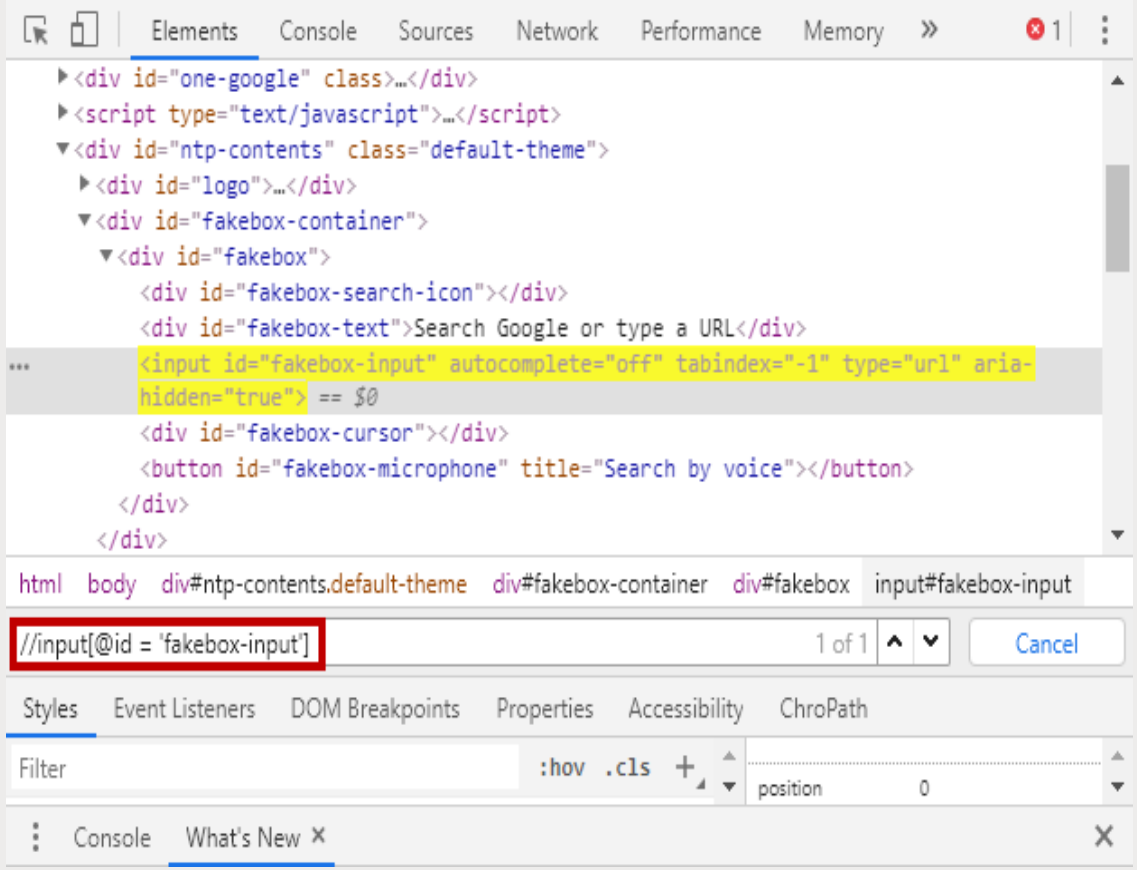
PREDICTION OF SEARCH QUERIES USING VIDEO COMMENTS

Lemon Lin Reimer

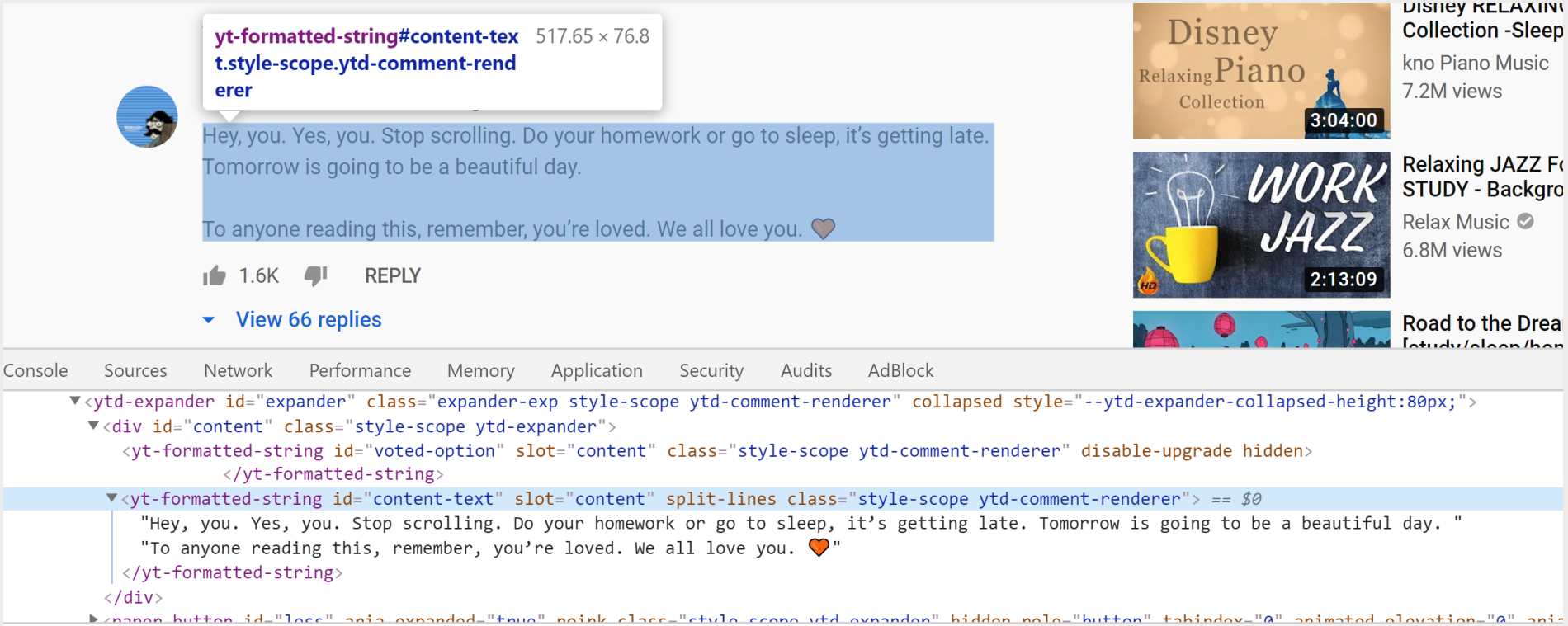

Data Retrieval
Cleaning
Tokenization
Stemming and Lemma
Vectorization
Dimension Reduction
Train Model
Test Model
Validate Model

xpaths in HTML code





Data collection through selenium



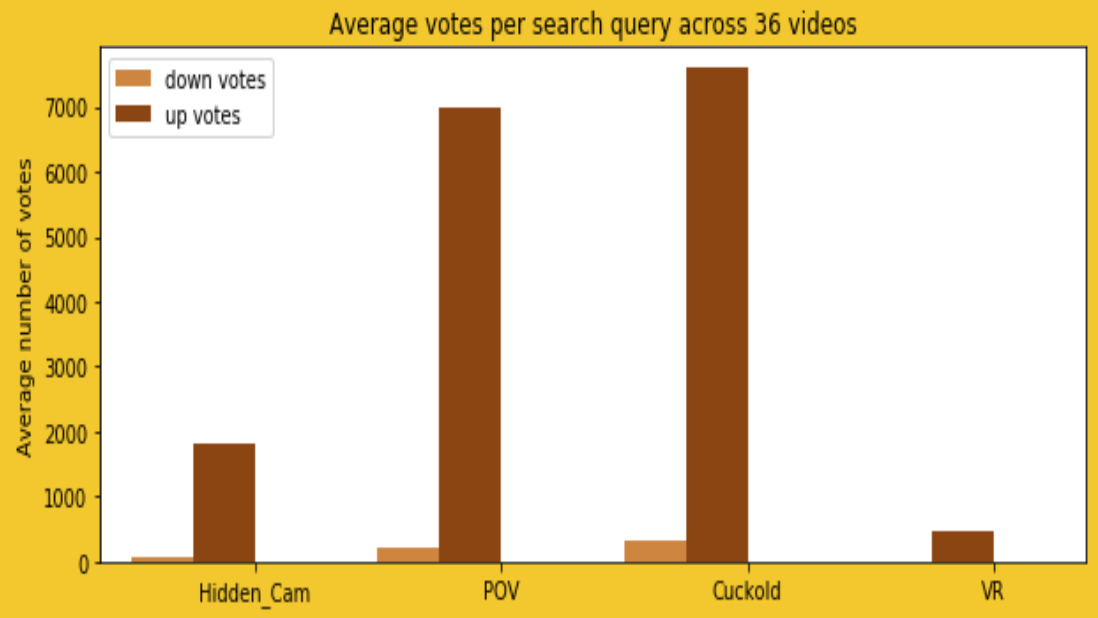

```
self.driver = webdriver.Chrome(chrome_options = Options().add_argument("--headless"))
self.driver.get("https://www.youtube.com/")
comment = self.driver.find_element(s)_by_xpath('//*[@id="content-text"]')
text = comment.text
slot_name = comment.get_attribute("slot")
```

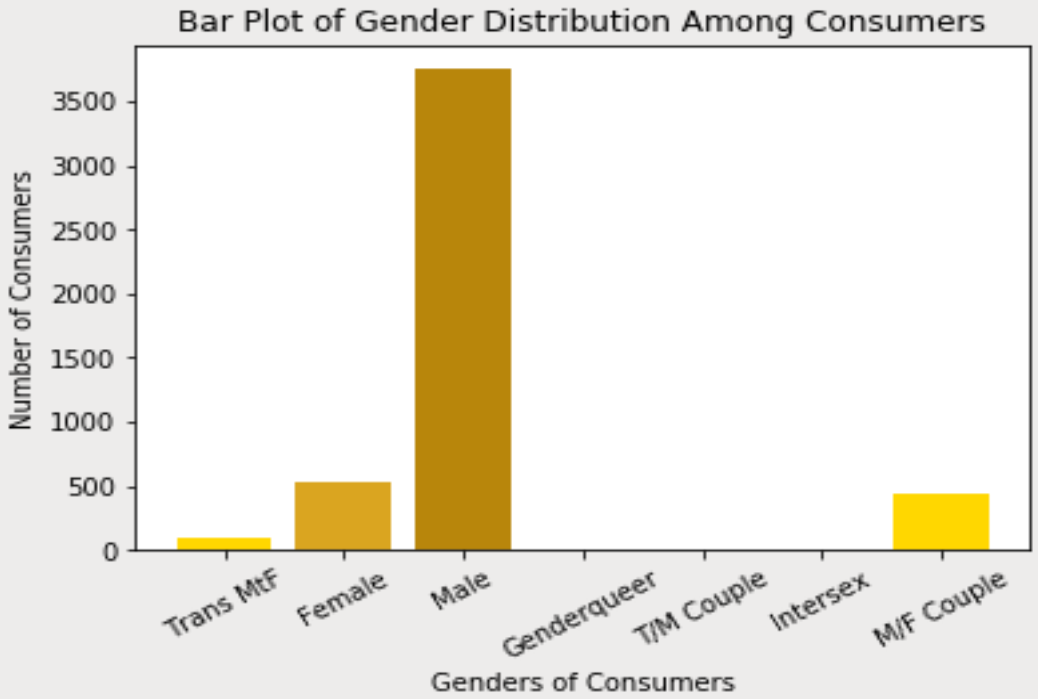
VARIABLES EXTRACTED INTO DATA FRAME

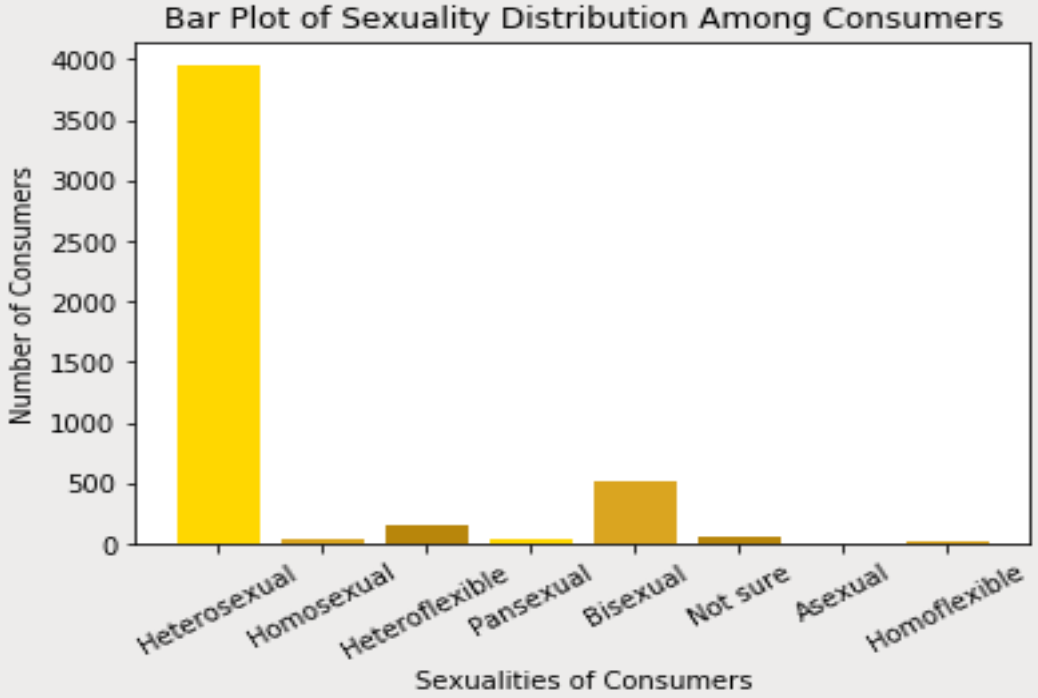
- Query string used to search for videos
- Link to video (36 per query)
- Number of upvotes and number of downvotes per video
- Number of views per video
- Raw comments + cleaned + lemmatized + stemmed
- Gender and sexuality of commenter
- Link to user profile of commenter
- Date of comment posted in UTC

Description of the variables

Graphics were produced using the matplotlib package.







Data Collection:	Selenium (Chrome driver)
Cleaning:	re
Vectorization:	tfidfVectorizer
Tokenization:	NLTK
Stemming/Lemma:	sklearn
Dimensionality Reduction:	PCA
Grid Search:	sklearn
Classifier:	RandomForrestClassifier

Validation in progress!

CONCLUSION

- This was really fun!
- Can potentially use this data for thesis
- Learned to develop class objects and call them from a top file
- Tackled plenty of bugs and glitches
- Played with a new package and some HTML with selenium
- Need to work on developing classification models
- Want to learn more about (un)supervised learning
- Unsure how to label data without expert labeling

ANY QUESTIONS?