Topic Modelling of Movie User Reviews

- SUBMITTED BY :
  - TANVI ARORA

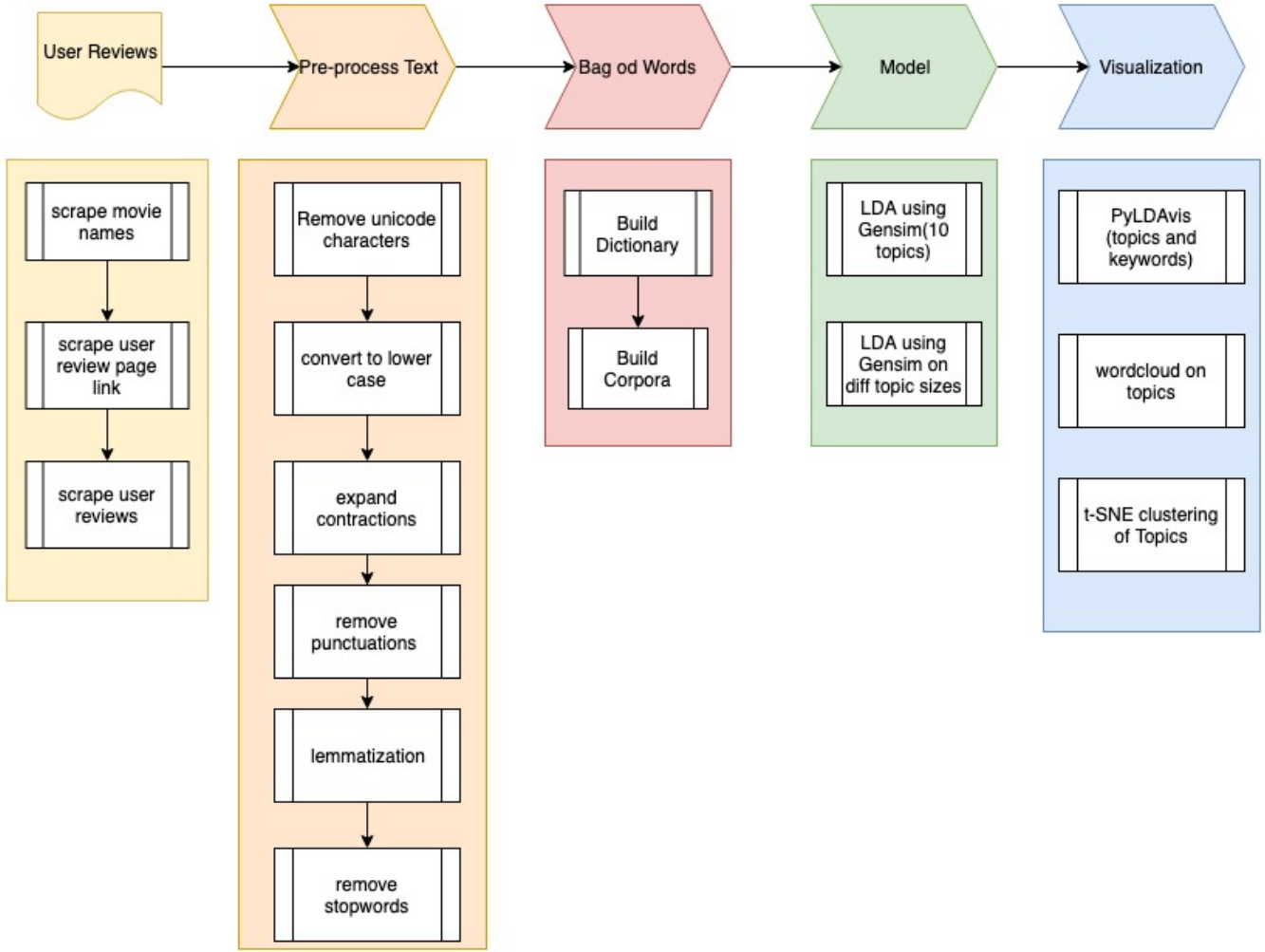
User Reviews for Movies

- Source : imdb.com
- Genre : Biography
  - The Irishman
  - Bohemian Rhapsody
  - When They See Us
  - RocketMan
- What data is available on imdb ?
  - User Rating (scale of 0 to 10) - structured data
    - (eg : positive reviews : >5 , negative reviews <5)
  - User reviews (ranging from 0 to 1000s) - unstructured data
  - Votes

What are they talking about ??

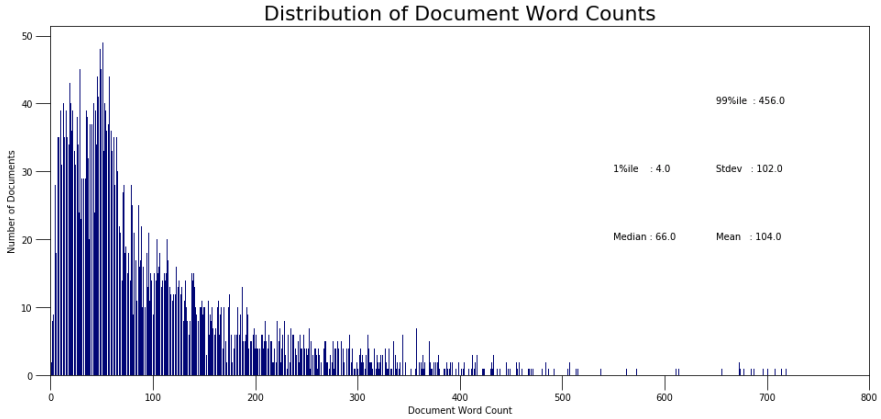
- User reviews can be positive or negative based on the user rating
- But really what are the users talking about ?
  - Movie
  - Actor
  - Performance
  - Type of story
- How do I find that ??
  - Topic Modeling
  - Is a type of statistical modeling for discovering the abstract topics that occur in a collection of documents
  - Help in making sense of the unstructured data

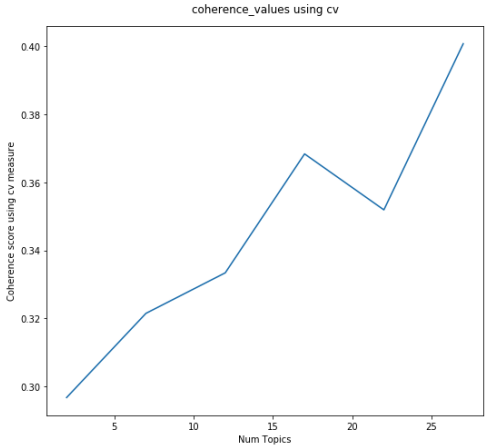
Pipeline


User Reviews
Pre-process Text
Bag od Words
Model
Visualization
scrape movie names
scrape user review page link
scrape user reviews
Remove unicode characters
convert to lower case
expand contractions
remove punctuations
lemmatization
remove stopwords
Build Dictionary
Build Corpora
LDA using Gensim(10 topics)
LDA using Gensim on diff topic sizes
PyLDAvis (topics and keywords)
wordcloud on topics
t-SNE clustering of Topics

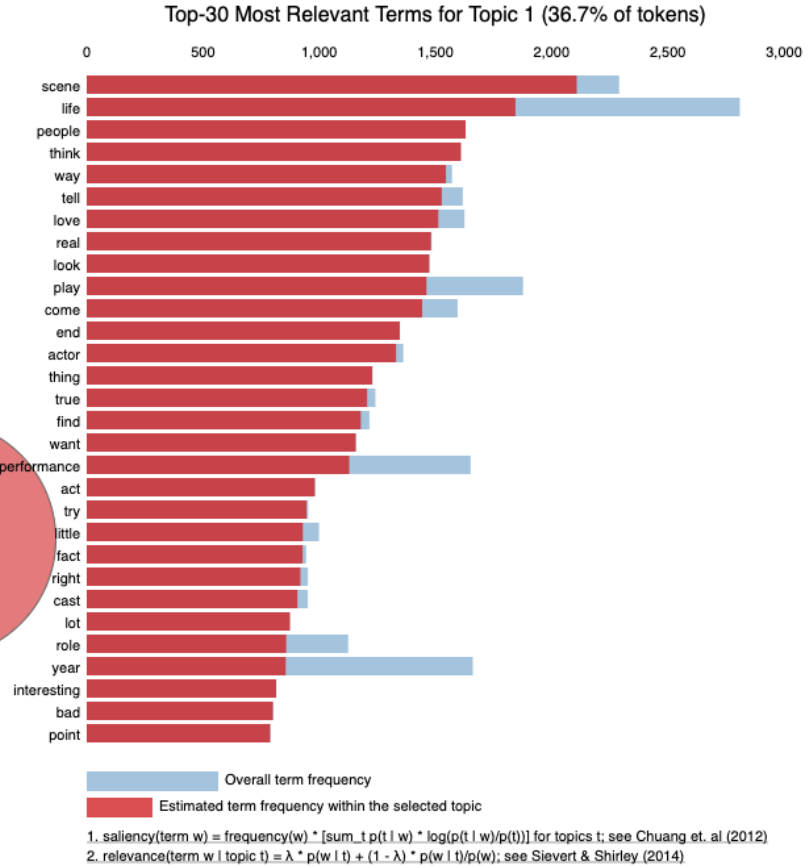
Model

- How much data are we talking about ~ 4500 user reviews
- Latent Dirichlet Allocation (LDA)
- A few open libraries exist but main contender is Gensim
- Cannot automatically determine number of topics
- Tested with a range of 2 to 30 with an interval of 5 for the number of topics
- Topic Models give no guarantee on the interpretability of the output
- Coherence Score : measures the relative distance between words within a topic
    - We want coherence score to be as high as possible

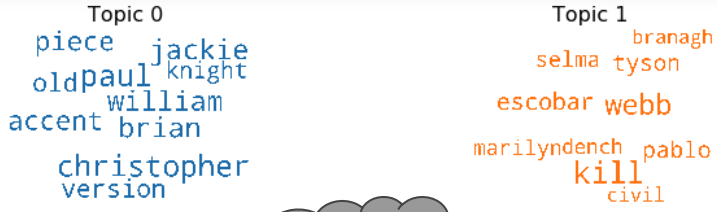


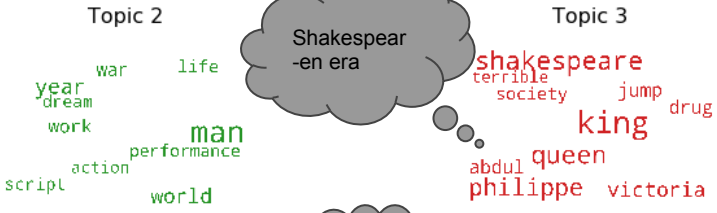


Topic and Keywords



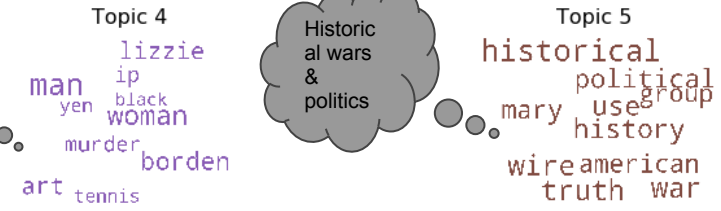
TOPICS





Shakespear -en era

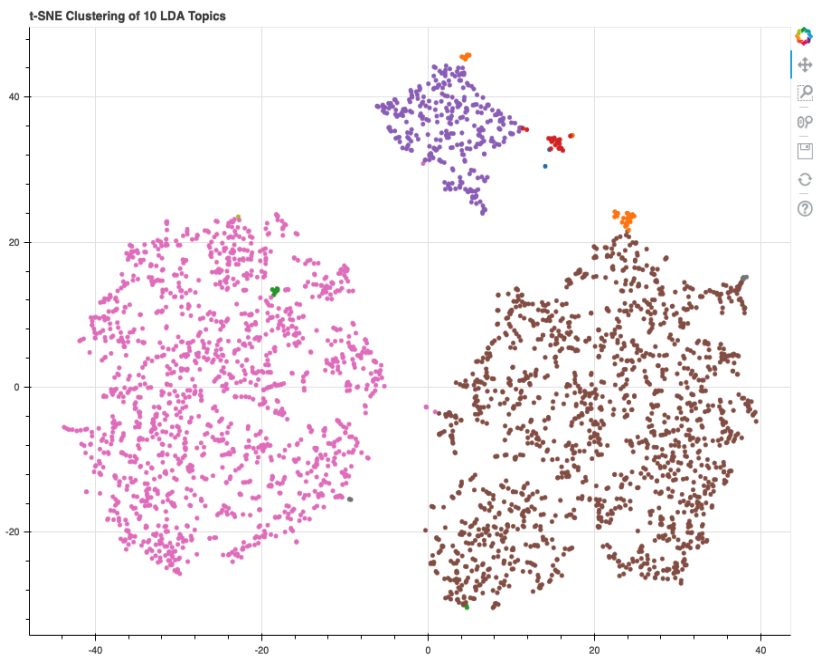
Black man & woman



Historic al wars & politics

About the movie

International movies



Thank you