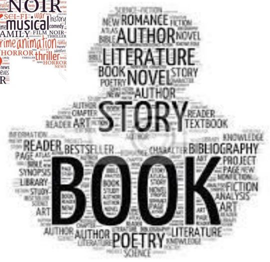
Perception of literature and cinematography according to NLP



Movies move you in their pace.

Books move to yours

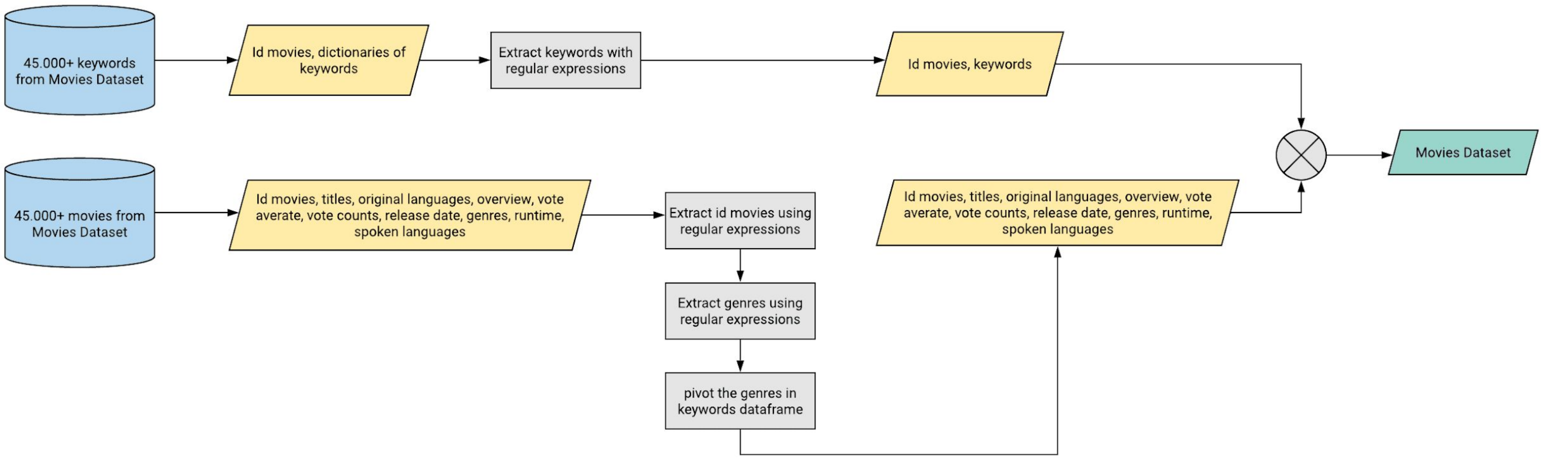
Introduction

- Reading Habits in the **U.S.**
  - American spend **5:42 hours per week** on reading.
  - **37%** of American adults with a high school degree or less, and
  - **7%** of college graduates didn’t read a book **past year**

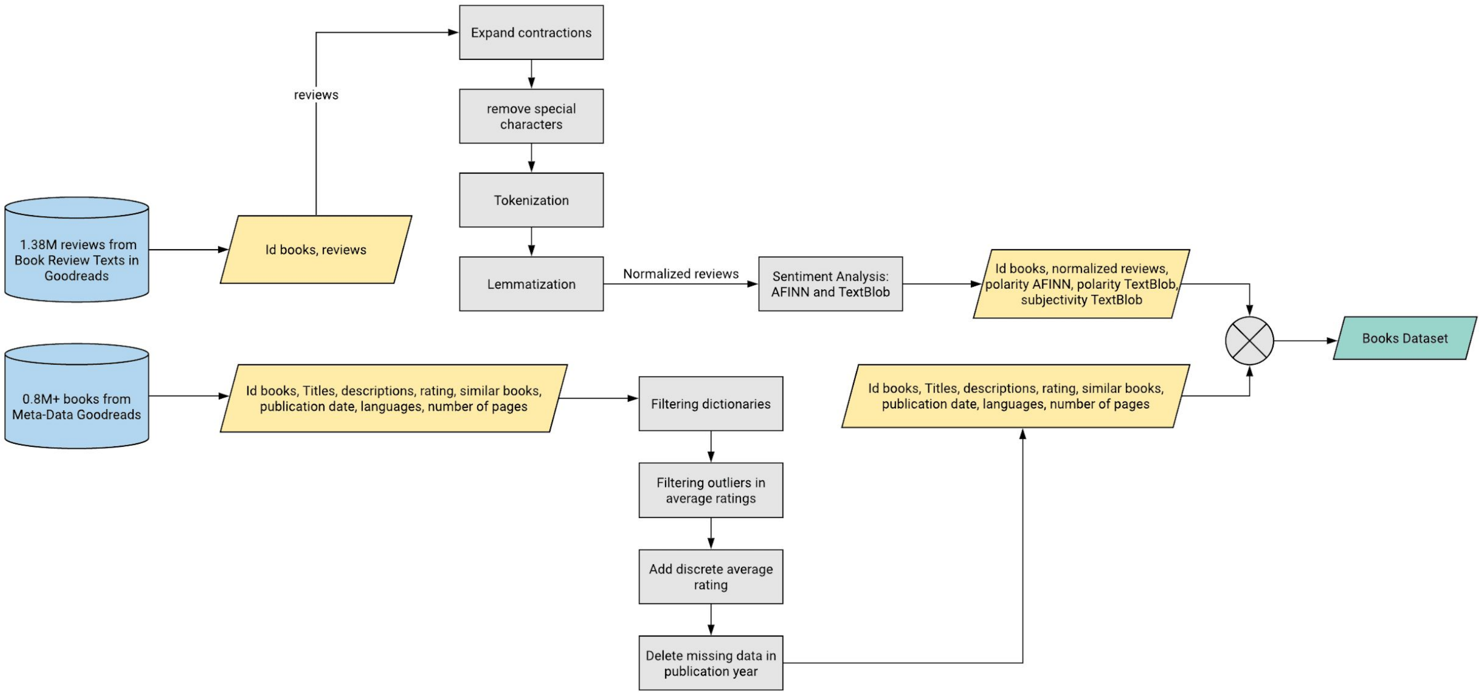
- Cinema and movies in **U.S.**
  - **14%** of U.S adults visite a movie theater **one or more** times per month
  - **46%** go to the cinema **once per year** or less
  - Netflix has **60.1 million** U.S subscribers in **2019**

- Cinematographic Adaptations
  - In **2018**, Netflix developed around **50** literacy projects.

Movies: Acquisition and Wrangling

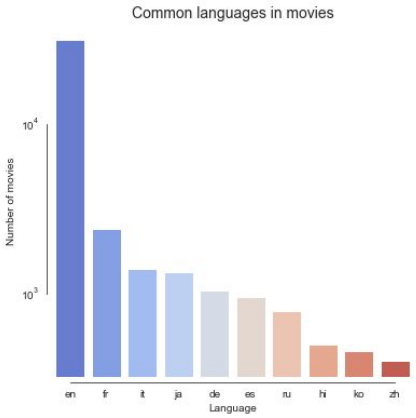


Books: Acquisition and Wrangling



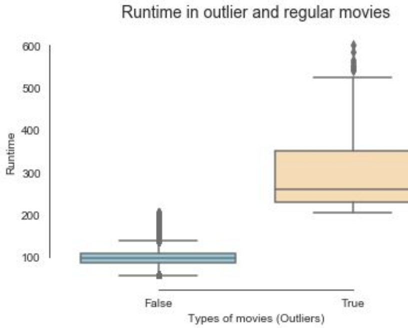
Initial findings for Movies

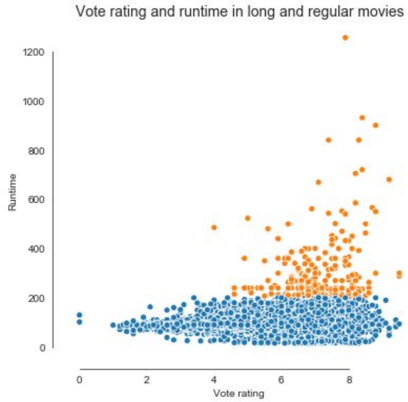
Most popular languages



English is the most popular, followed by **French**, **Italian**, **Japanese**, **Germain**, **Spanish** and **Russian**

How loved or hated are long movies?

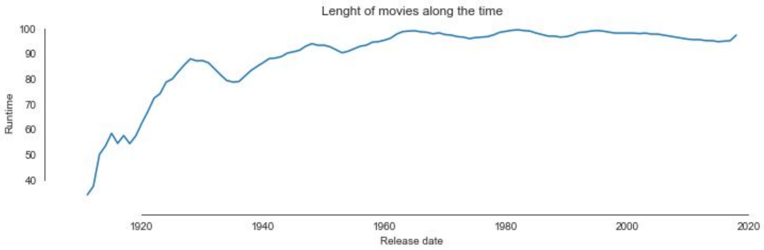




Percentiles 25 and 75 for regular movies are **86 and 107 minutes.** For longest movies, **230 and 350 minutes.**

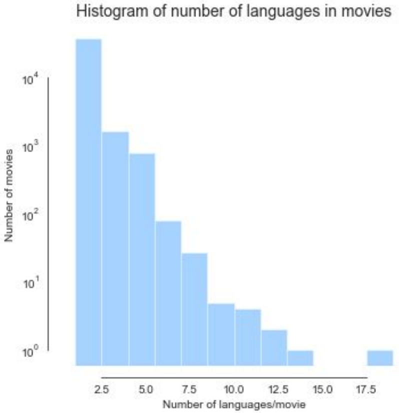
Initial findings for Movies

Are movies getting longer?



From the beginning of the century until now, **length of films** still being around **100 minutes**

How many languages we can find in one movie?

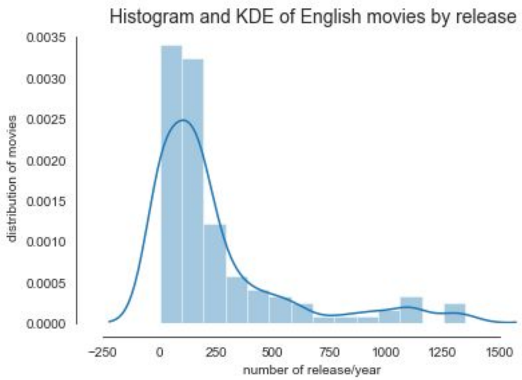


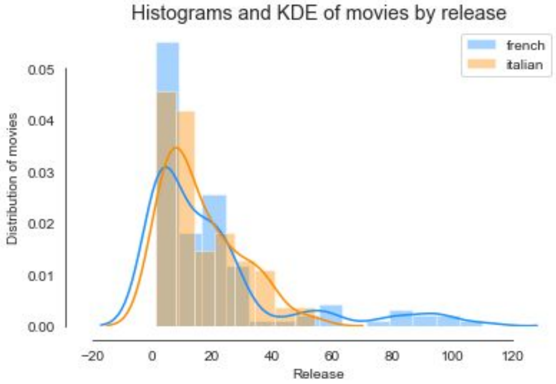
Vision of Europe, anthology film about **25 directors from Europe**

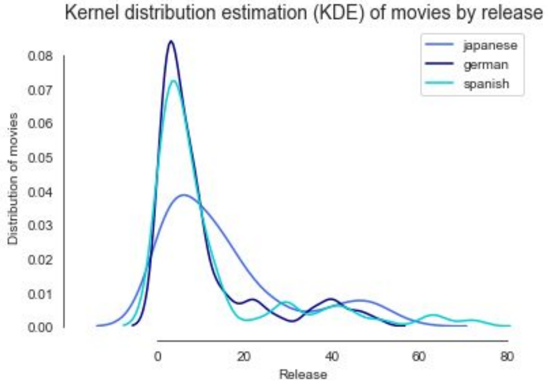
Less than **10 movies** include **more than 9 languages**

Initial findings for Movies

Distribution of movies by language and releases/year







ANOVA statistical test

Do determined distributions have the **same population mean?**

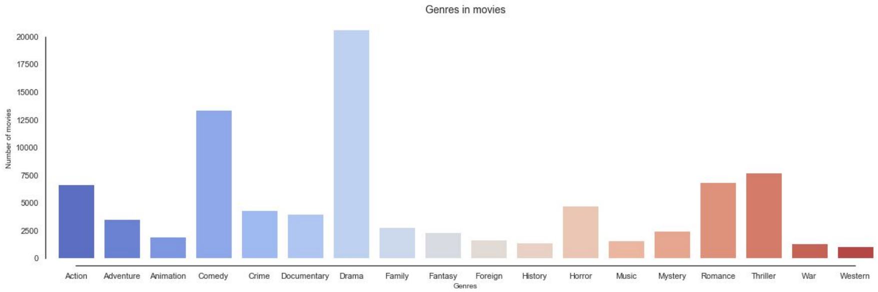
Test 1: English, French and German

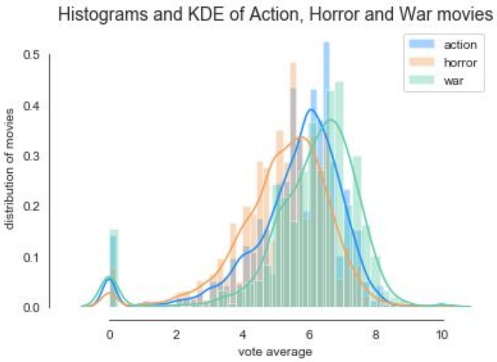
Test 2: French and German

Test 3: Japanese, German and Spanish

Initial findings for Movies

How frequent and likely are movies by genres?



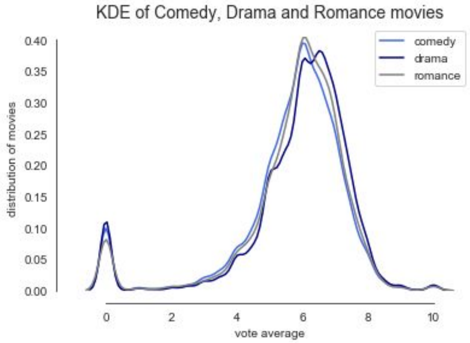


Levene statistical test

Do samples come from **populations with equal variances?**

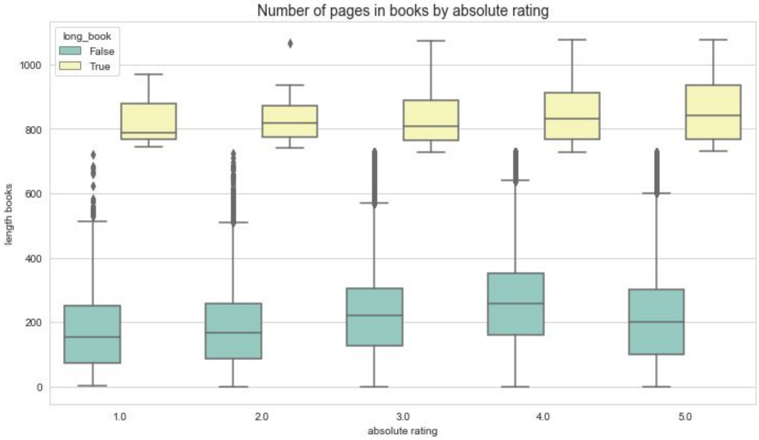
Test 1: Action, Horror and War

Test 2: Comedy, Drama, Romance



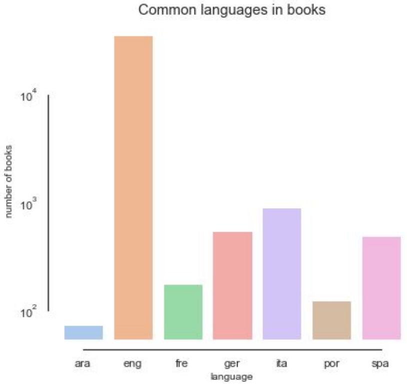
Initial findings for Books

Do readers prefer the shortest or longest books?



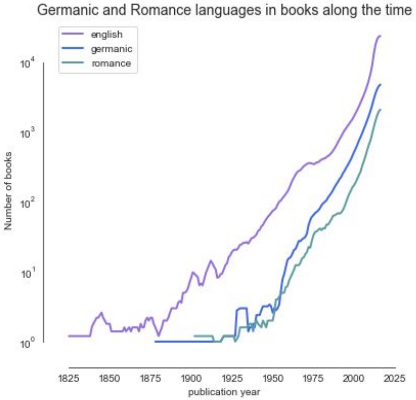
The mean length for regular books is **200 pages vs 800** for long books.

Popular languages in books



English is the most popular language, covering the **60%** of the whole dataset.

Romance languages include French, Spanish, Portuguese, Italian, Romanian, Catalan, Aragon
Germanic language family includes German, Swedish, Danish, Dutch, Norwegian, Afrikaans and Icelandic

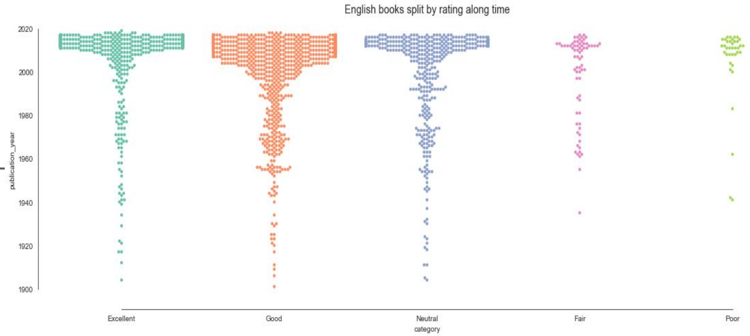


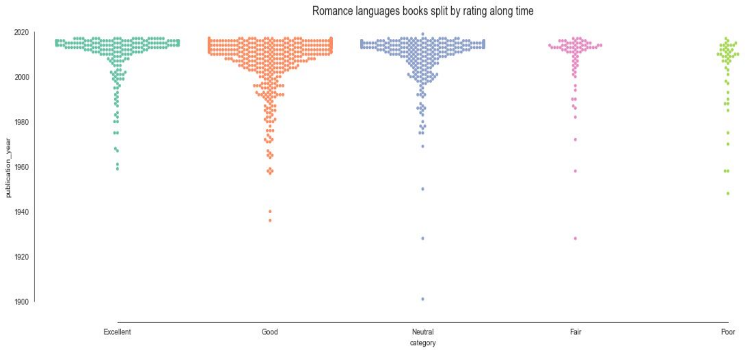
Initial findings for Books

Books by language and average score through the time

How we validate this sample data?

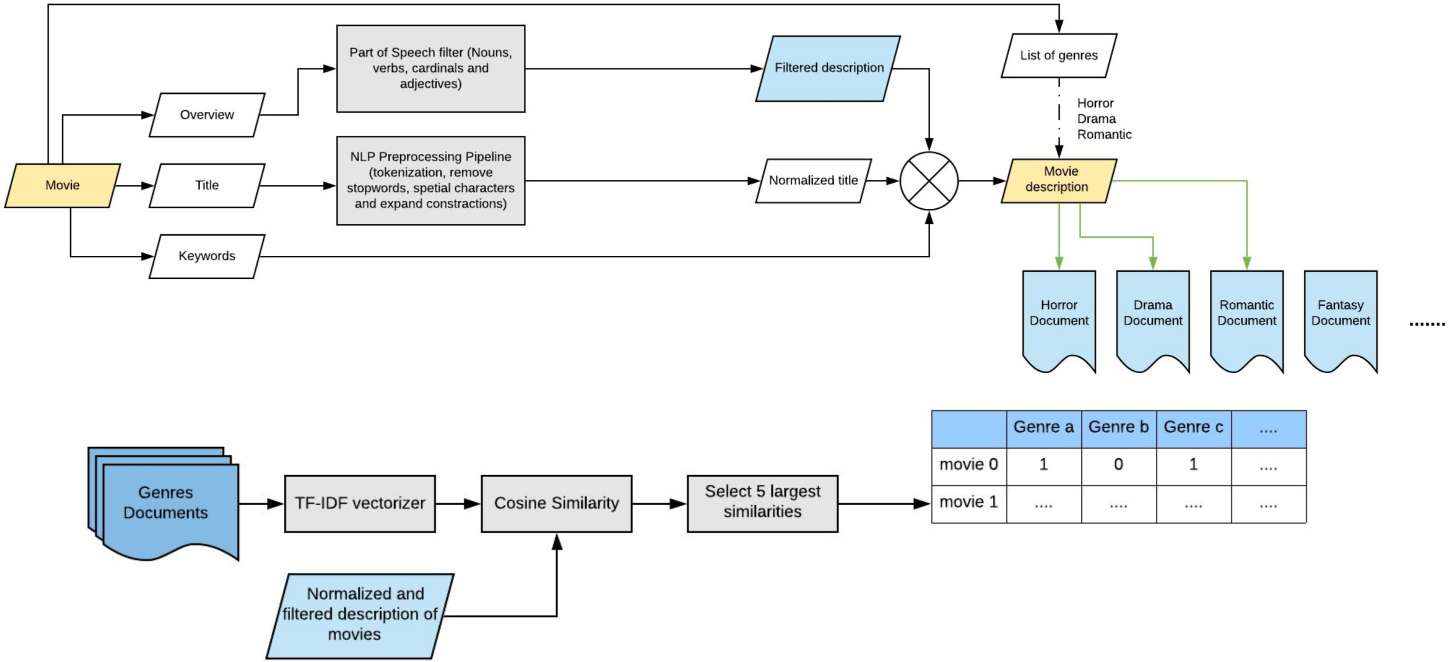
Chi-square bootstrapping test: how many times we fail to reject the null hypothesis of chi-square test between the ten most popular languages in population and sample data





In-depth Analysis

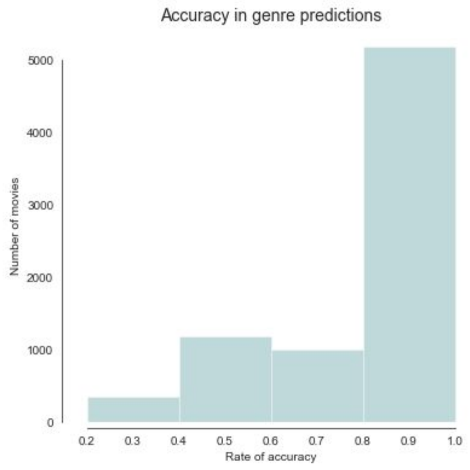
Predicting genres in movies using the description of movies



	Genre a	Genre b	Genre c	....
movie 0	1	0	1	....
movie 1	....	....	....	....

In-depth Analysis

Predicting genres in movies using the description of movies



Description	Genres	Predicted genres
*american **president** widower widowed u.s. president andrew shepherd one world powerful men anything sydney ellen wade **washington lobbyist** shepherd attempts spark wild rumors approval ratings (movie: The American President)*	Comedy Drama Romance	History Adventure Action Documentary War
*goldeneye cuba kgb satellite cossack **james bond mysterious** head janus syndicate leader goldeneye **weapons** system revenge britain (movie: GoldenEye)*	Adventure Action Thriller	Action Thriller Adventure Western

In-depth Analysis

Determining genres in books using the movie dictionary

Book title	Words from description (title normalized and text filtered by PoS)	Genres
Dark Matter	**dark matter**, **thriller**, **unconscious**, **reality, mask**, **gunpoint**	Thriller, Drama, Mystery, Horror
Sofia's Magic Lesson	**magic**, **magical**, **amulet, friendship**, **tricky**	Fantasy, Family, Adventure, Drama
Portugal's Guerrilla Wars in Africa	**wars**, **army**, **conflicts**, **colonies**, **liberation**	War, History, Action, Documentary
Harry Potter and the Chamber of Secrets	**wizarding**, **friends**, **hogwarts**, **legendary** ...	Family, Adventure, Fantasy, Action
Batman: Detective Comics, Vol. 3	**detective**, **vigilantes**, **shadows mysterious**, **deadly**, **dark, knight**, **crime-fighters**, **league plan** ...	Mystery, Crime, Action, Thriller
The Secret Life of a Dream Girl	**dream, girl, teen**, **drinking, drugs, kiss**, **superstar**	Romance, Drama, Family, Fantasy

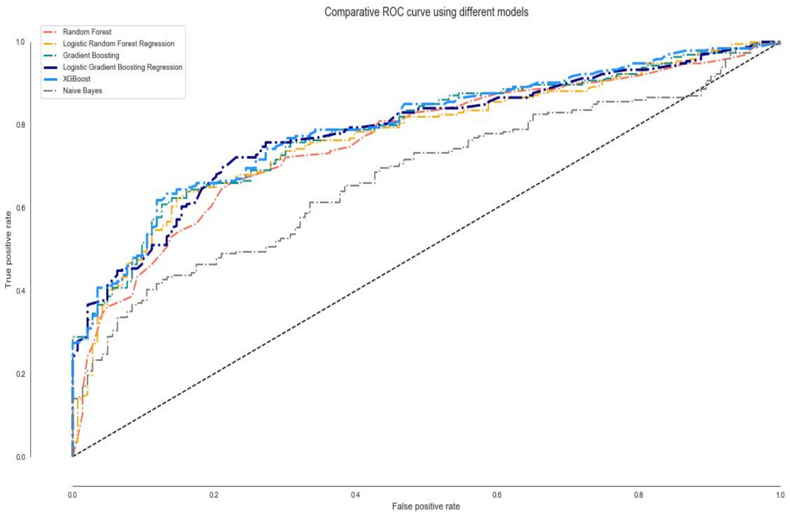
In-depth Analysis

Selecting features to predict reception of books (chi-square test)

Feature	Score	Feature	Score
length	12.73	Action	2.09
rating count	9.26	Adventure	1.90
review counts	6.49	Drama	1.67
History	3.45	Horror	1.00
Romance	3.36	**Mystery**	**0.47**
Music	3.05	**War**	**0.34**
Fantasy	2.41	**Thriller**	**0.13**
Family	2.40	**Crime**	**0.12**
Documentary	2.15		

Using the new features of books, we discover that **History, Romance, Music, Documentary** and **Action** are some genres related to the reception by readers

ROC Curves of Ensemble trees and Naive Bayes models



In-depth Analysis

Accuracy and AUC by model

Model	AUC	Accuracy
RANDOM FOREST	0.76	71.81%
LOGISTIC RANDOM FOREST	0.77	70.62%
GRADIENT BOOSTING	0.79	71.81%
LOGISTIC GRADIENT BOOSTING	0.79	74.48%
XG BOOST	0.80	73.29%
NAIVE BAYES	0.67	52.52%

XGBoost get the second best performance

Logistic Gradient Boosting get better performance than the best version of the original Gradient Boosting

Confusion Matrix XGBoost and Logistic Gradient Boosting

False negative rate of 32.1%

False positive rate of 22.7%

XGBoost	Predicted Positives	Predicted Negatives
Actual Positives	97	46
Actual Negatives	44	150

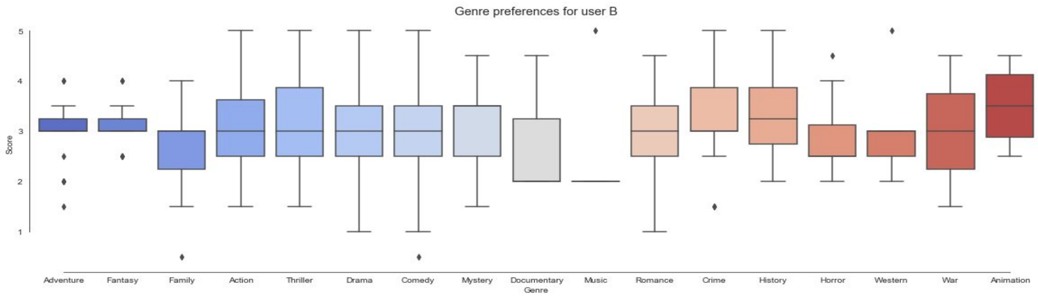
Log Gradient Boosting	Predicted Positives	Predicted Negatives
Actual Positives	104	39
Actual Negatives	47	147

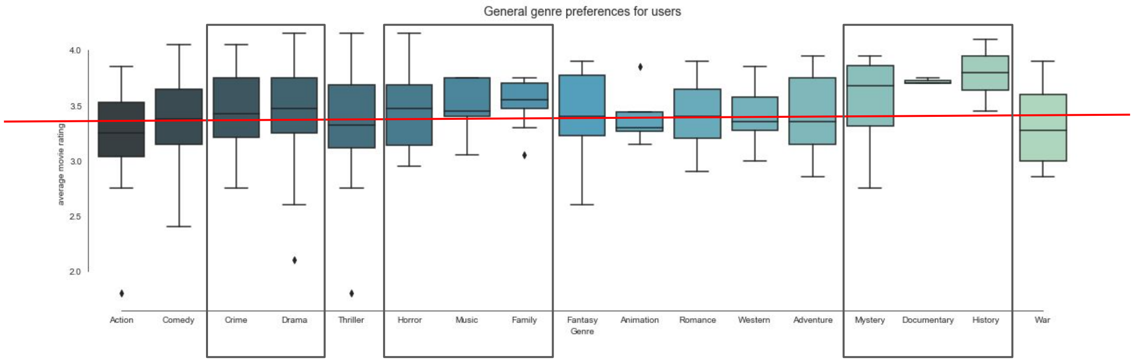
False negatives rate of 27.2%

False positive rate of 24.2%

In-depth Analysis

User *cinephiles* profiles





Cinephiles: people with 120-400 movies voted

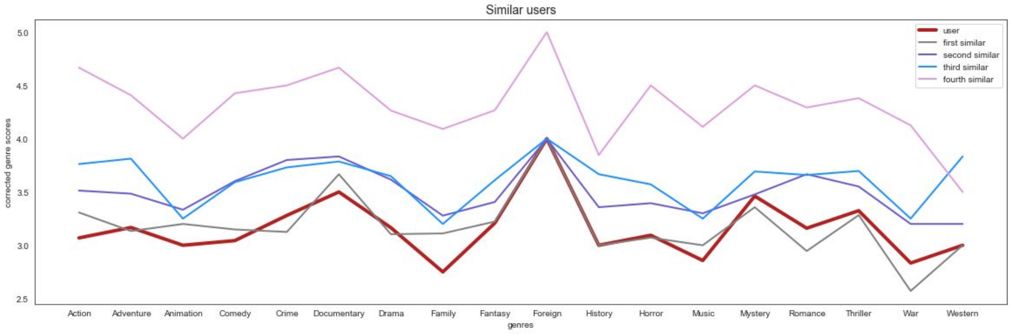
User scores require a **correction** to consider ratings based on **average scores** and **quantity** of movies belonging to the genres

$$s = 5 \cdot \frac{p}{10} + 5 \cdot \left(1 - e^{-\frac{q}{Q}}\right)$$

genre	average score	count	corrected score
Drama	3.16	100	3.93
Foreign	4.00	2	2.06
Comedy	3.04	57	3.04

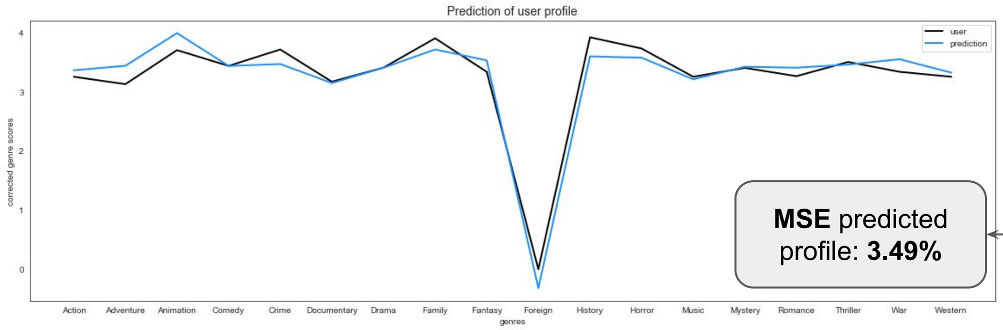
In-depth Analysis

User Based Collaborative Filtering (UBCF)



Similar users: K-Nearest Neighbors with cosine similarity as distance metric

UBCF algorithm to predict score for pair (user, genre): mean rating of user, plus a weighted average of deviations from neighbor's mean

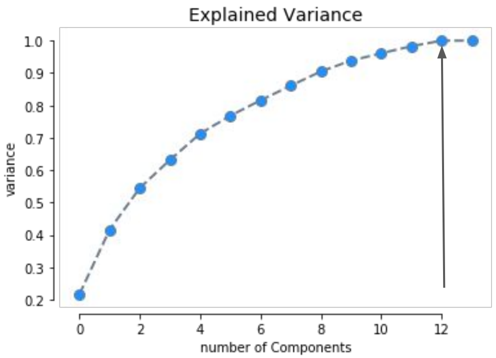


MSE predicted profile: **3.49%**

Iteration over the **UBCF algorithm** asking for the different genres scores for the same target user.

In-depth Analysis

Clustering of books based on genres



According to the **Cumulative Summation of the Explained Variance,** using 12 components, we get a variance of **0.98**

Agglomerative Clustering for 9 clusters

The **Adventures** Of Sniffy Doo Da: The Giant Lizard
Adventure In A Glorious Hole
The **Adventures** of Pyjama Boy

Family, Adventure

Australia and the **Second World War**, 1939-45
The **Evolution** of **Russia**
The **Auschwitz** Volunteer
Cold War

War, History

Galactic Storm
The **Supernatural** Goes High Tech
Fear the **Darkness**
Stranger Abduction

Fantasy, Horror

Murder With A View
The **Ghost** of Robert Brown
X-Men **Homicide** Squad
Her Majesty's Historical **Detective**
The **Case** of the **Missing** Twin 4

Crime, Mystery, Thriller

Love or Kill Them All
Five Gold Rings: an Elizabethan **Love Story**
The **Seduction** of Anita Sarkeesian
First **Kiss**

Drama, Romance

Spanish Agent, An Erotic Spy **Thriller**
Pirates of Nirado River
When **Darkness** Finds You

Action, Adventure, Thriller

An **Unexpected** Visit
The Dream **Killer**
The Amateur's Guide To **Death** and **Dying**

Drama, Thriller

The **Drama** of Masculinity and **Medieval English**
The Threshold of **Christianity**
The **Biography** of Prophets

Documentary, History

Superman
Monsters Galore
DC **Superhero** Girls Sampler
Harry Potter and the Deathly Hallows

Action, Adventure, Fantasy

Conclusion

- The **average runtime** of movies has not suffered considerable changes over the last decades. **Length of books** increase according to the rating.

- ANOVA tests are used to compare the distribution of movies by release/year for the most popular languages. **French-German** and **Japanese-German-Spanish** distributions **have the same population mean**.

- The genres more recurrent in movies are *Drama* and *Comedy* followed by *Action*, *Horror, Thriller*, and *Romance.* Levene statistic tests concluded that distributions of *Drama*, *Comedy* and *Romance* films **come from populations with equal variances.**

- For sampling data in books, **bootstrapping** tests asserts that the samples are representative of the population.

Conclusion

- **Overviews**, **titles**, and **keywords** of movies are used to **predict genres**, through **cosine similarity** between documents by genres.

- Documents by genre included normalized titles (NLP preprocessing pipeline for deleting **stop-words**, **expanding constractions**, removing **special characters**, **tokenization** and **lemmatization**), overviews filtered by **Part of Speech** and keywords.

- The model is validated using a testing data (20% of movies). 91.81% of trials at least one genre was successfully predicted.

- Documents by genre create new features for books. Using this new features and number of pages, number of ratings and reviews, a binary classification problem of quality of books (good scores and bad scores) is resolved.

Conclusion

- *History*, *Romance* and *Music* played more relevant roles to predict the ratings. **Logistic Gradient Boosting and xgBoost** achieved the highest scores in both metrics AUC and Accuracy.

- **User-Based Collaborative Filtering** was used to resolve the problem of predict the vote of one user for an unexplored genre. We compared the right and predicted scores from one user, getting a **mean squared error** of 3.49%.

- Book Clustering: feature vectors (containing genres) were reduced by **PCA** inspecting the **Cumulative Summation of the Explained Variance,** selecting 12 components and creating new representations of the features in other dimensional spaces. **Agglomerative Clustering** for 9 clusters using **Euclidean distance.**

References

- Movies Dataset: Metadata on over 45,000 movies. 26 million ratings from over 270,000 users.Available in: movies_dataset
- Goodreads Metadata of books. Available in: Goodreads
- Understanding Gradient Boosting Machines. Available here
- Named Entity Recognition with NLTK and Spacy. Available here
- Collaborative Filtering Based Recommendation Systems Exemplified. Available here
- Text Similarities : Estimate the degree of similarity between two texts. Available here
- Visualize dependencies and entities in your browser or in a notebook. Available here
- How to use Data Scaling Improve Deep Learning Model Stability and Performance here
- Traditional Methods for Text Data (Dipanjan Sarkar). Towards Data Science. Available here
- Practical Statistics for Data Scientist (Peter Bruce and Andrew Bruce). O'REILLY, 2017.
- A Practitioner's Guide to Natural Language Processing (Part I) - Processing and Understanding Text (Dipanjan Sarkar). Towards Data Science. Available here