Predicting a Movie's Success through character dialogue

Research & Application Track

Avijeet, Pedro, Sean

Motivation:

- Every year movie studios and producers manually validate countless of scripts and select ones that are likely to yield a high ROI, if produced as a movie.
- Even so, some of these movies fail!
- Minimize manual effort using NLP system while maximizing profitability
- At the very least, we aimed to lower the amount of false positives, to "weed out" the bad scripts and let the humans focus on the potential good ones.
- We also attempted to identify 'character types' through clustering

Data:

- We scraped scripts in .txt format from
    - https://www.imsdb.com/
- Movie Budget data was taken from
    - https://www.the-numbers.com/movie/budgets/all

Initial Preprocessing:

- Most of our scripts followed the standard format
- Identify characters names:
    - Count the frequency of each unique sentence
    - Get a list of potential characters: frequency > 5
    - Select the top 5 characters with the most dialogue
      (need next bullet point for this final step)
- Extract dialogue for each potential character:
    - Loop over each sentence.
    - If the whole sentence is uppercase (filtering some stuff),
        - we get all the sentences below it with the same level of indentation.
        - This level is identified as the indentation of the first sentence that does not contain “(<action description>)”.

11.

JOHN
Well, one can’t have everything.

CUT TO:

EXT. JOHN AND MARY’S HOUSE - CONTINUOUS

An old car pulls up to the curb and a few KNOCKS as the engine shuts down.

MIKE steps out of the car and walks up to the front door. He rings the doorbell.

BACK TO:

INT. KITCHEN - CONTINUOUS

JOHN
Who on Earth could that be?

MARY
I’ll go and see.

Mary gets up and walks out.

The front door lock CLICKS and door CREAKS a little as it’s opened.

Improving this approach & Saving the dialogues:

- We compared our list of characters with the ones listed on IMDb, and obtained 0.82 accuracy. This is under evaluating as some names might not match.
- While performing clustering and extracting different features, we identified edge cases that were accounted for in the code
- We also identified most of the movies that did not follow the standard format, and removed them from our dataset (about 40/700)
- We saved the dialogues for each movie with .json format: https://raw.githubusercontent.com/PedroUria/NLP-Movie_Scripts/master/dialog_jsons/Big-Lebowski%2C-The_script.json
- We also distinguished lines by keeping a count [i]

Extracting Features:

1. Iterate through each character dialogue (for each movie)
    a. Subset dialogues to top 5 characters per film in terms of dialogue length
2. Calculating features per character (examples below)
    a. Each character´s overall polarity
    b. Cosine similarity between each of the top 5 characters in a film
    c. Number of times characters mention each other (and polarity of sentences where they are mentioned)
    d. Use of certain types of words and POS
3. Aggregating features to the movie level for the success model

Character Clustering: Topic Modeling (LSA)

Goal: find topics corresponding to genres or types of characters

- Preprocessing/ Methodology: Filtered out stopwords, numbers, character names, and all POS except Noun, Verb, Adverb and Adjectives.
- Tried different combinations of stemming/lemmatization and n-grams (up to n = 4)
- Results: Bad. Many of the topics actually looked very similar to each other, and a handful included statements that looked more like camera directions than dialogue.

Character Clustering: Topic Modeling (DBSCAN)

Methodology/ Preprocessing: Extracted textual features at the character level (Relative Character Vocabulary, Frequency of verbal pauses like ‘um’, Frequency of syntactic hedges, Polarity, etc.) Did a Grid Search of the model parameters (minpts and epsilon/radius size) and searched for clusterings that provided interesting results.

Results: Promising. One model produced 8 clusters (18% of the characters were considered noise points) 4 of the clusters had a small number of characters who had similar characteristics in their dialogues.

Cluster 1 3 characters, two of which were Gibbs From pirates of the Caribbean and Charlie Frost from 2012, who both provided exposition.

Cluster 2: 3 characters, two of which were Malcolm X from Ali and Bevilacqua from The Box (both boxing movies).

Cluster 3: 3 Characters, Mitchell from The Damned United, Van Houten from the Fault in Our Stars, and Captain Idaho from Postman, all of whom were all very direct, rude characters

Cluster 4: 5 Characters: Jigsaw From Saw, Dr. Waldman From Frankenstein, Razor from Matrix Reloaded, TV reporter from Signs and Mirror on the Wall from Shrek.

Clusters 5-8: Too many characters to analyze, no similarities

Success Prediction:

- Budget + Box Office Data -> ROI (%) -> Success if ROI (%) > 0
- Our dataset is very imbalanced: about 82% of successful movies
- We computed the point biserial correlation coeff between all the features and our target. All were very low, only the ones above abs(0.05) are shown
- We also tried different kinds of aggregation but for all cases the correlation never went above 0.1
- We built models using various combinations of all the features we had extracted.
- We also tried different definitions of success (ROI (%) > threshold)

```
[('feels_per_sent_char_4', -0.08888232022410765),
 ('n_unique_words_char_5', -0.07048985653489775),
 ('hw_per_sent_char_5', -0.06577199616838104),
 ('num_pass_sents_char_2', -0.0649465684944353),
 ('neg_polarity_of_mentions_char_2', -0.06100731867019248),
 ('hw_per_sent_char_4', -0.06046108725323846),
 ('neg_polarity_of_mentions_char_5', -0.05208250003937597),
 ('stdvs_n_mentions_others_above_mean', 0.05387715079884093б),
 ('compound_polarity_of_mentions_char_1', 0.05740745993397291),
 ('compound_polarity_of_mentions_char_2', 0.05749812972244838),
 ('overall_polarity_char_5', 0.05887453635649594),
 ('overall_polarity_char_1', 0.0641859885522022),
 ('overall_polarity_char_2', 0.06591015813071763),
 ('wav_polarity', 0.06725406939228437),
 ('n_coref_sents_char_5', 0.06896815473624167),
 ('compound_polarity_of_mentions_char_5', 0.07238797113615202),
 ('pos_polarity_of_mentions_char_1', 0.07387336719748638)]
```

Success Prediction:

- We run a grid search but our models were predicting everything as successful, i.e, they were not learning
- We tried oversampling and undersampling, but the result was the same
- We decided to focus on precision, in order to lower the amount of false positives

Predicted	1
True	
0	23
1	108

Predicted	0	1
True		
0	9	14
1	37	71

Predicted	0	1
True		
0	6	17
1	10	98

Conclusion:

- Profitability Prediction
    - Through this project we learned that there is probably no underlying function between language features and profitability
    - Features like polarity, similarity and so on, represent the quality of dialogues but do not provide success predictive power
    - A much more complex model for movie scripts is needed in order to predict success, using features such as character interactions, progression of the story, etc.
- Character Clustering:
    - We were able to find clusterings of characters that were similar, or at least, spoke similarly using **very basic** features...
    - If more sophisticated linguistic features to cluster on, we could theoretically find clusterings that match hollywood character stereotypes.
    - We weren't able to get perfect clusterings because our features weren't capturing relationships between the characters, or the characters personalities, just their speech patterns.

Future Scope:

- Find ways to capture relationships between character
- Predict topic progressions in the story across genres
- Create methods map between characters types and actors, and thereby reducing the effort and bias while Casting

Thank you

"Catch you on the Flippety-Flip."

- Michael Scott