Understanding Movie Quality from Plot Summaries using Natural Language Processing

Aakash Tandel • 09.04.2017

Overview

Objective

Data

Natural Language Processing Basics

Algorithms and Modeling

Additional Analysis

Summary

Objective

Build a model that takes plot summaries in as inputs and determines whether that plot would result in a good movie.

Acknowledgements

PyData Lecturers

Patrick Harrison
Bhargav Desikan

Github Users

Mark Riedl

General Assembly Lecturers

Matt Brems
Matt Speck
Mark Mummert
Evann Smith

Data

Wikipedia Plot Summaries
Metacritic Metascores

Wikipedia Plot Summaries

Synopsis

- 112,000+ plots
- Included movies, novels, and television shows
- Utilized Natural Language Toolkit, spaCy, and Gensim

spaCy was used for general exploratory data analysis but was eventually abandoned for NLTK in the analysis section.

Preprocessing Text Data

Stop Words

Removing common words in the English language like “the” and “a”

Punctuation

Stripping plots of punctuation prevents characters like “&” and “,” as tokens

Stemming

Reducing words to their root, such as “information” becomes “info”

Lemmatization

Reduces inflections and plural words to their lemma, such as “geese” becomes “goose”

'A Game of Thrones follows three principal storylines simultaneously. At the beginning of the story, Lord Eddard "Ned" Stark executes a deserter from the Night\'s Watch, who has betrayed his vows and fled from the Wall. On the way back, his children adopt six direwolf pups, the animal of his sigil. There are three male and two female direwolf pups, as well as an albino runt, which aligns with his three trueborn sons, two trueborn daughters, and one bastard son. That night, Ned receives word of the death of his mentor, Lord Jon Arryn, the principal advisor to Ned\'s childhood friend, King Robert Baratheon. During his own visit to Ned\'s castle of Winterfell, Robert recruits Ned to replace Arryn as the King\'s Hand.

“

A Game of Thrones follows three principal storylines simultaneously. At the beginning of the story, Lord Eddard "Ned" Stark executes a deserter from the Night's Watch, who has betrayed his vows and fled from the Wall. On the way back, his children adopt six direwolf pups, the animal of his sigil. There are three male and two female direwolf pups, as well as an albino runt, which aligns with his three trueborn sons, two trueborn daughters, and one bastard son. That night, Ned receives word of the death of his mentor, Lord Jon Arryn, the principal advisor to Ned's childhood friend, King Robert Baratheon. During his own visit to Ned's castle of Winterfell, Robert recruits Ned to replace Arryn as the King's Hand.

Sentence 1:
A Game of Thrones follows three principal storylines simultaneously.

Sentence 2:
At the beginning of the story, Lord Eddard "Ned" Stark executes a deserter from the Night's Watch, who has betrayed his vows and fled from the Wall.

Sentence 3:
On the way back, his children adopt six direwolf pups, the animal of his sigil.

Entity 12: Lord Jon Arryn - PERSON

Entity 13: Ned - PERSON

Entity 14: Robert Baratheon - PERSON

Entity 15: Ned - PERSON

Entity 16: Winterfell - ORG

Entity 17: Robert - PERSON

Token	Part of Speech
follows	verb
three	number
Eddard	noun

Web Scraping Metacritic

Summary

- Beautiful Soup
- Metacritic HTML accessible
- Batched by Genre

The scraper pulled movie title by genre. Additionally, it pulled cast member names, director names, release year, plot summaries, and the compiled Metacritic score.

“Metascore” is an average critic review weighted by the critic’s publication and quality (as determined by Metacritic).

Above 0.5 became a “Good” movie

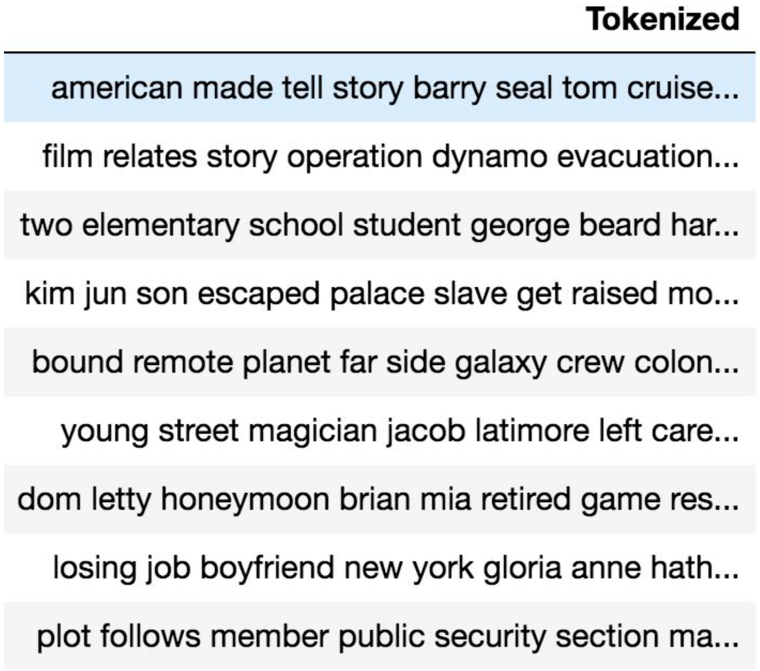
NLP Algorithms

Bag of Words

- Counts the number of times a word (or token) is present in a coprus

TF-IDF

- Word frequency for the plot discounted by the popularity of the word in the entire corpus of plots
- Stands for Term Frequency - Inverse Document Frequency
- Rewards frequent, distinct words



Tokenized
american made tell story barry seal tom cruise...
film relates story operation dynamo evacuation...
two elementary school student george beard har...
kim jun son escaped palace slave get raised mo...
bound remote planet far side galaxy crew colon...
young street magician jacob latimore left care...
dom letty honeymoon brian mia retired game res...
losing job boyfriend new york gloria anne hath...
plot follows member public security section ma...

Ready for Modeling

Model

- TF-IDF Vectorizer
- Random Forest Classifier
- Support Vector Machine
- XGBoost
  - 0.61 AUC-ROC Score
- Model predicts better than a weighted coin flip

Additional Analysis

Topic Modeling

Latent Dirichlet Allocation

Hierarchical Dirichlet Process

Word2Vec

Turn words into feature vectors

Keeps similar words together

Jump to Jupyter Notebook for pyLDAvis

model.most_similar('war')

```
[(‘wage', 0.5109046697616577),
('conquer', 0.5086956024169922),
(‘invas', 0.500656008720398),
('faction', 0.4899093806743622),
('iraq', 0.4895903468132019),
(‘vietnam', 0.48379379510879517),
('civil', 0.47669172286987305),
('grecopersian', 0.47665053606033325),
(‘phantomattack', 0.4651501476764679),
('union', 0.46387824416160583)]
```

Summary

1) Data

Web scraped Metacritic.com for Metacritic Metascores and utilized previously scraped Wikipedia plot summaries

2) Natural Language Processing

Preprocessing text data with lemmatization, removing stop words, and removing punctuation.

3) Algorithms and Modeling

Term Frequency-Inverse Document Frequency algorithm used to vectorizer text data.

Support Vector Machines, Random Forest Classifiers, and XGBoost used to model

4) Additional Analysis

Topic modeling with Latent Dirichlet Allocation and Hierarchical Dirichlet Process

Word embeddings with Word2Vec

THANKS!

Any questions?

You can find my Jupyter Notebook at https://github.com/aakashtandel