Film Script Analyzer

Luis Castro

Rationale

- Films are a billion dollar industry.

- Top creatives have to skim over many scripts.

- Boundless data remains to be untapped in written language by machine learning.

We can do better.

Solution

"An application that extracts all relevant information from text so it can describe it, compare it, summarize it, rate it and recommend adequate or similar alternatives."

Tools

- Machine Learning
  - Recommendation Systems
  - Supervised/Unsupervised Learning

- Natural Language Processing
  - Text mining/web scrapping
  - Summarization

- API connections
  - IBM Watson
  - IMDB

Dataset
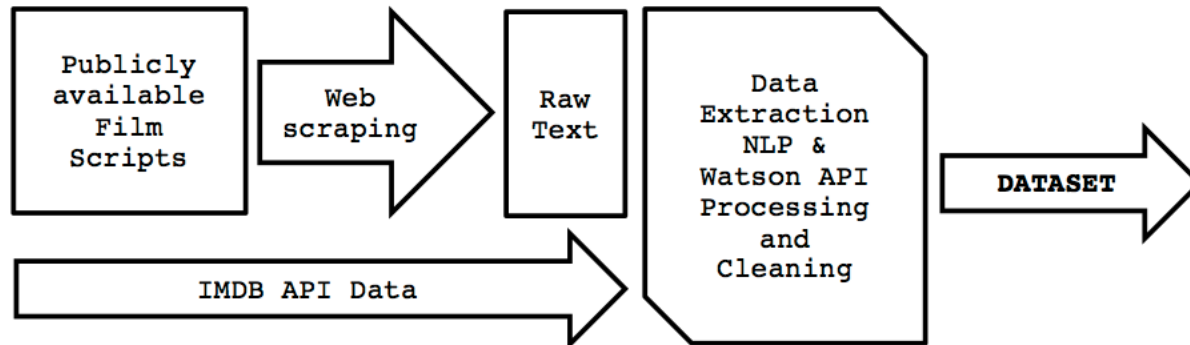
Overview of the process followed to create the dataset.
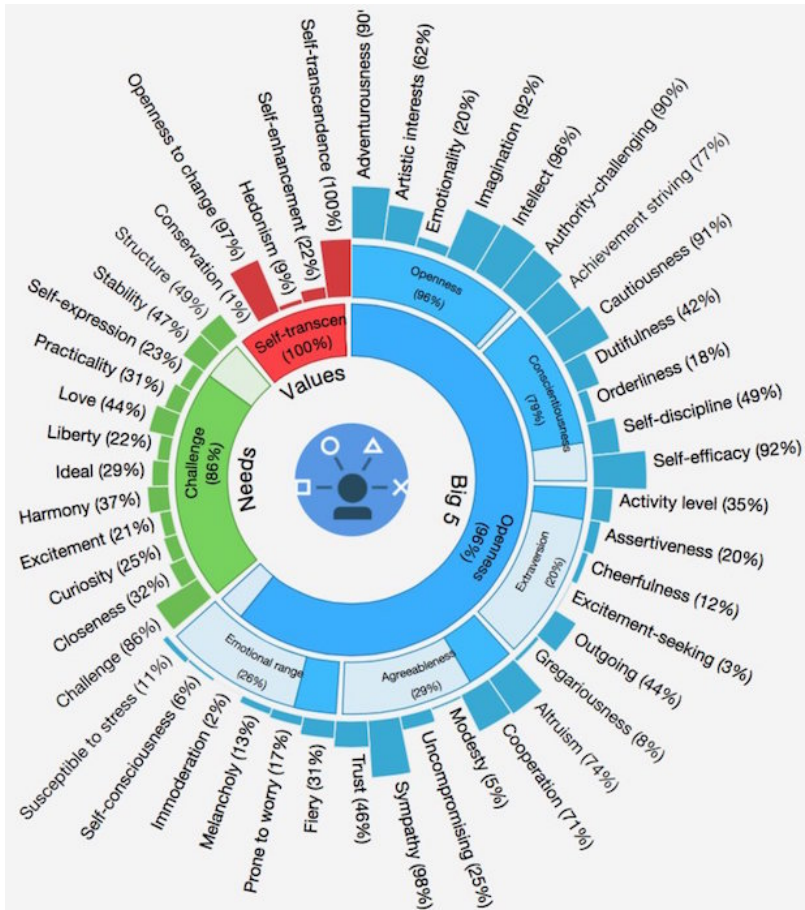
Main dataset

Over 2300 scripts scrapped

Over 1300 scripts' data within the dataset
- 30 personality insights features
- 13 text statistics features
- 19 film features from IMDB

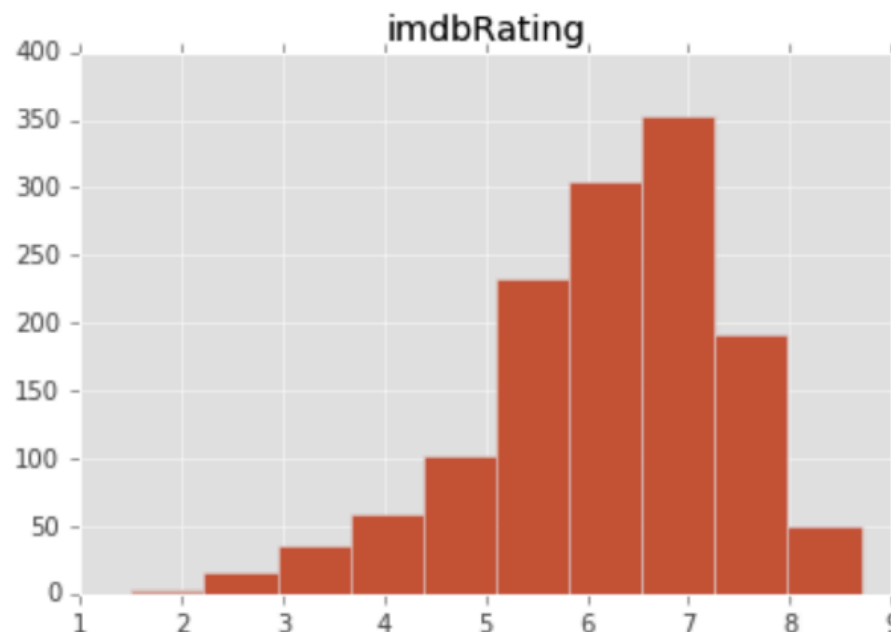**Recommendation matrix** for the 1300+ scripts.
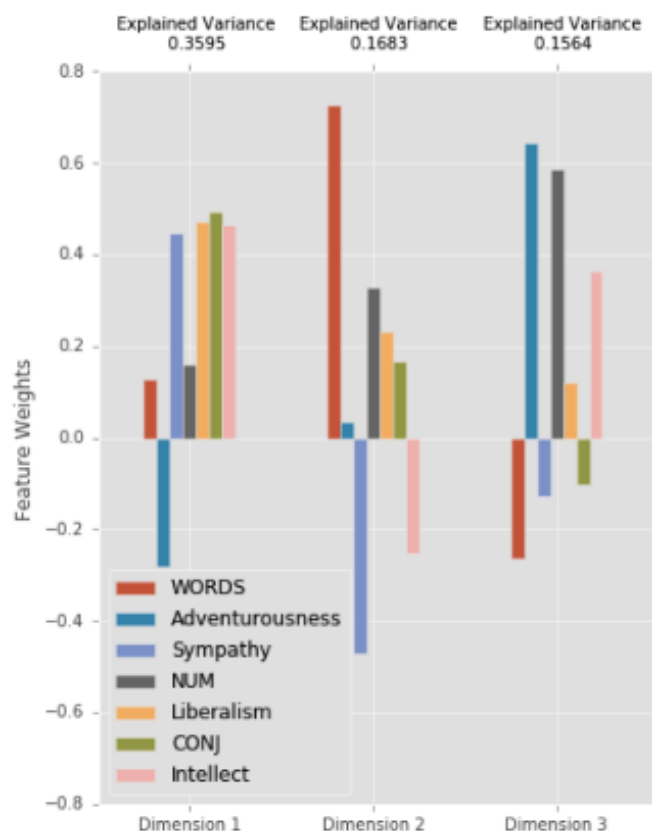
IBM Watson

"IBM Watson is a technology platform that uses natural language processing and machine learning to reveal insights from large amounts of unstructured data"

Main functions

- Description
- Recommendation
- Classification
- Summarization
- Evaluation

Describe

	WORDS	Adventurousness	Sympathy	NUM	Liberalism
importance	0.086854	0.062778	0.040557	0.038585	0.032985

Importance as evaluated vs. 'imdbRating'.

Recommend

Eat Pray Love

```
In [94]: import webbrowser
         from IPython.display import HTML, Image

         poster = df[df.index=='In the French Style']['Poster'][0]
         HTML('<iframe src='+poster+' width=800 height=400></iframe>')
```

Out[94]:

```
nrec['Eat Pray Love'].sort_values(ascending=False).head(3)
```

```
Title
In the French Style         0.990937
Under the Greenwood Tree    0.990831   ←————————————
Far from the Madding Crowd  0.990372
Name: Eat Pray Love, dtype: float64
```

For context:

```
print df[df.index=='In the French Style']['Plot'][0]
```

A young American art student must decide whether to stay in Paris with her boyfriend or go back to
   the U.S. when her wealthy father arrives to bring her back.

Summarize

Eat Pray Love

```
fs = FrequencySummarizer()
fs.summarize('beat    spay    blove',5)  ←
```

Title: Eat Pray Love
Rated: PG-13
Runtime: 133.0
Director: Ryan Murphy
Actors: Julia Roberts, I. Gusti Ayu Puspawati
Genres: Drama, Romance
Rating: 5.7

Poster: https://images-na.ssl-images-amazon.com/images/M/MV5BMTY5NDkyNzkyM15BM15BanBnXkFtZTcwNDQyNDk0Mw@@._V1_SX300.jpg

Summary:
I'm going to Italy and then I'm going to David's guru's ashram in India... ...and I'm going to end the year in Bali.
It's long, it's tedious, I can't keep up... ...and I get these insane anxieties about everything in my life... ...and
 I've lost my place.
And it was such a foreign concept to me, that I swear I almost began with: "I'm a big fan of your work."
-No, I don't even have my-- I-- You don't have your-- You don't-- You're so naked.   ←
And I-- You know, I don't-- I don't know.

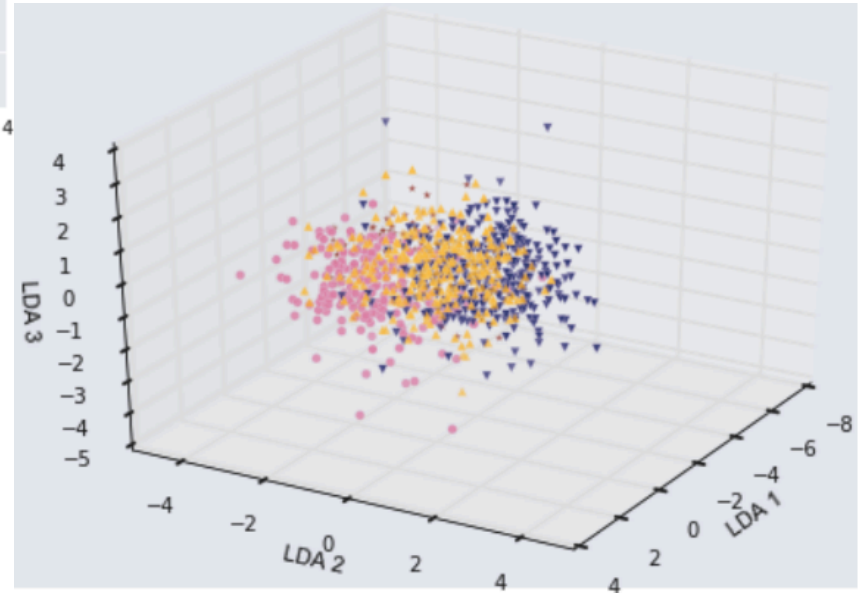
Similar Titles:
Me Again
In The French Style
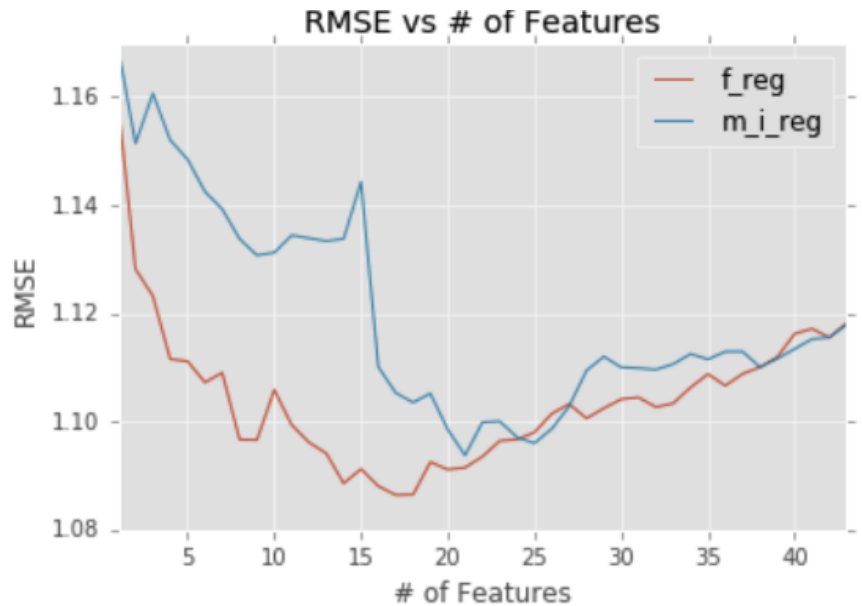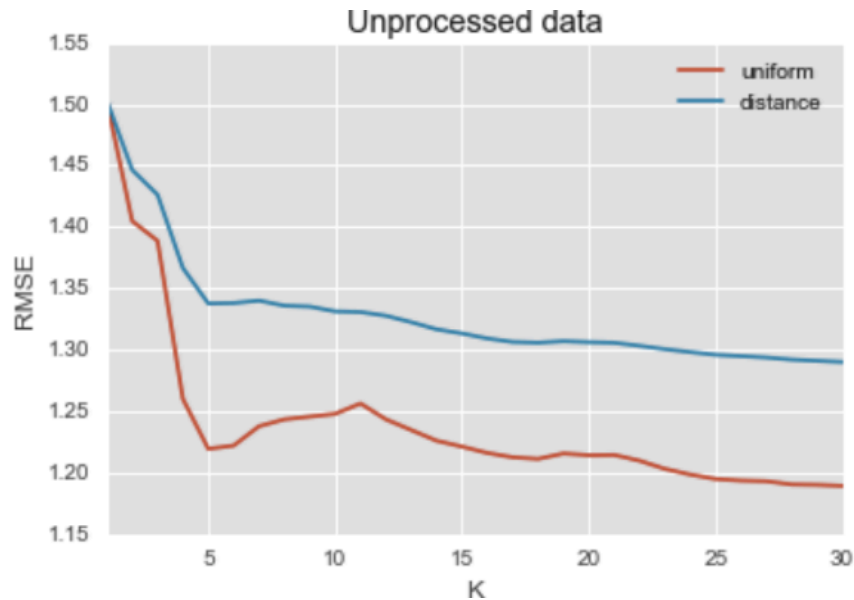Under the Greenwood Tree
Far from the Madding Crowd   ←
Dames

Classify

LDA of Scripts dataset

Current classification accuracy ~74%

Evaluate

Best 5 fold cross-validated RMSE: 1.064
Best 5 fold cross-validated MAE: 0.79

Why this?

- I have helped proofreading various scripts, one even sent to Cannes.

- It may seem arrogant to qualify art, but we do it all the time, so lets be precise about it.
  - *'That is not art, that is a piece of…'* **classification**.
  - *'I guess it's good, but doesn't move me'* which is like 6 or 7, **regression**.

- It can be expanded to literature "books", speech to text, plagiarism detection, etc.

Next steps

So far the results are promising but more data, more analysis and experimentation remain to be done.

- **Gather more data:** Scripts analyzed now are in the order of thousands; make it be in the order of hundreds of thousands.
- **Semantic features / word count:** Use word2vector to have another view or perspective of the scripts. Create word count of the top non stop words used by genre.
- **Ensemble models and NN:** Create ensemble classification and regression models, this way can help reduce the over fitting of a certain algorithm and bring closer to the real values or classes.
- **Develop web interface:** Create an interface to display the results and statistics of the scripts, and receive new ones to be analyzed and compared with the others.
- **Expansion:** Use the application infrastructure to provide similar uses for literature, books, poems, essays, etc.

'The limits of my language mean the limits of my world.' L. Wittgenstein

'The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents.' H.P. Lovecraft

Thank you.