Film Script Analyzer

Luis Castro

‘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

Rationale

- Films are a billion dollar industry.
- Top creatives have to skim over many scripts.
- Writers spend energy without an available tool to guide them.

We can do better.

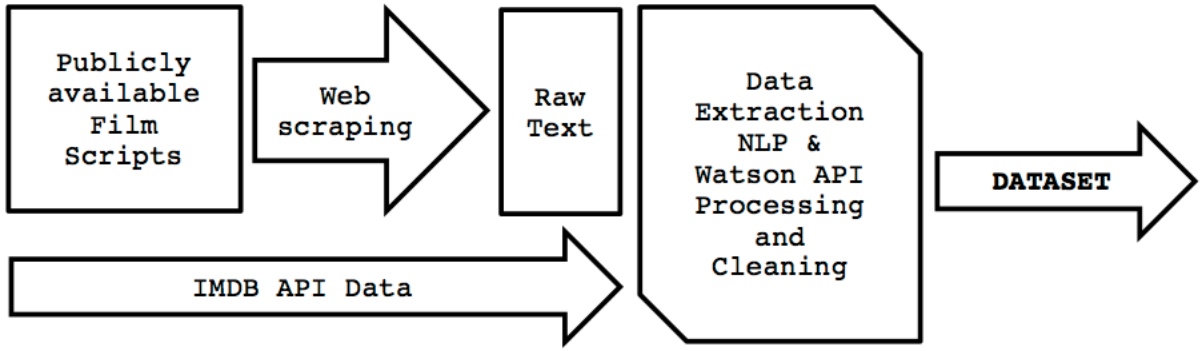
Solution

“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

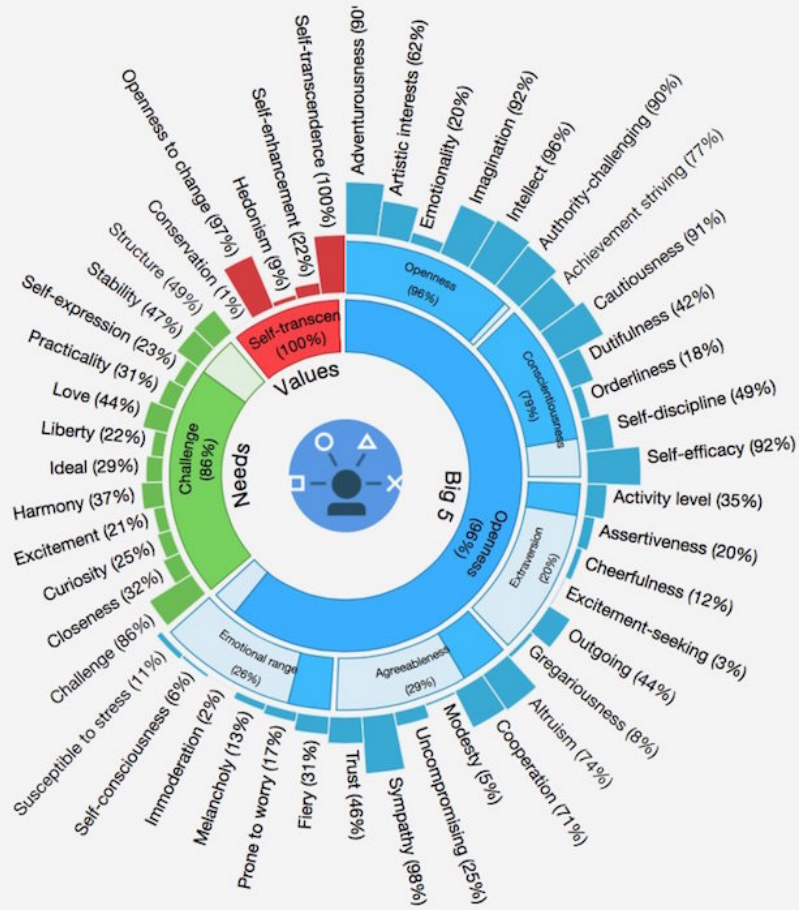
Over 800 scripts scrapped

Over 500 scripts' information contained in the dataset

- 30 personality insights features
- 13 text statistics features
- 19 film characteristics'' features from IMDB

Recommendation matrix for the 500+ 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

Summarize

Eat Pray Love

```
with open('scrapped/Eat Pray Love 2010.txt') as k:
    epl = k.readlines()
```

```
fs.summarize(epl[2],10)
```

```
["I'm going to Italy and then l'm going to David's guru's ashram in lndia... ...and l'm going to e
nd the year in Bali.",
 "It's long, it's tedious, I can't keep up... ...and l get these insane anxieties about everything
 in my life... ...and l've lost my place.",
 'And it was such a foreign concept to me, that l swear l almost began with: "l\'m a big fan of yo
ur work."',
 "-No, I don't even have my-- l-- You don't have your-- You don't-- You're so naked.",
 "And l-- You know, I don't-- I don't know.",
 "Do not tell me what lessons l have and haven't learned in the last year... ...and don't tell me
 how balanced and wise you are... ...and how I can't express myself.",
 "If it wasn't for you, I wouldn't have come back to Bali... ...and l wouldn't have come back to m
yself.",
 "l'm sorry l didn't call sooner.",
 "I don't know why we can't accept... ...we don't wanna live in unhappiness anymore.",
 "You know, it's been a rough day, and if no one takes it personally... ...l'm going to take my la
rge meal someplace else to eat it in silence."]
```

Recommend

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)
```

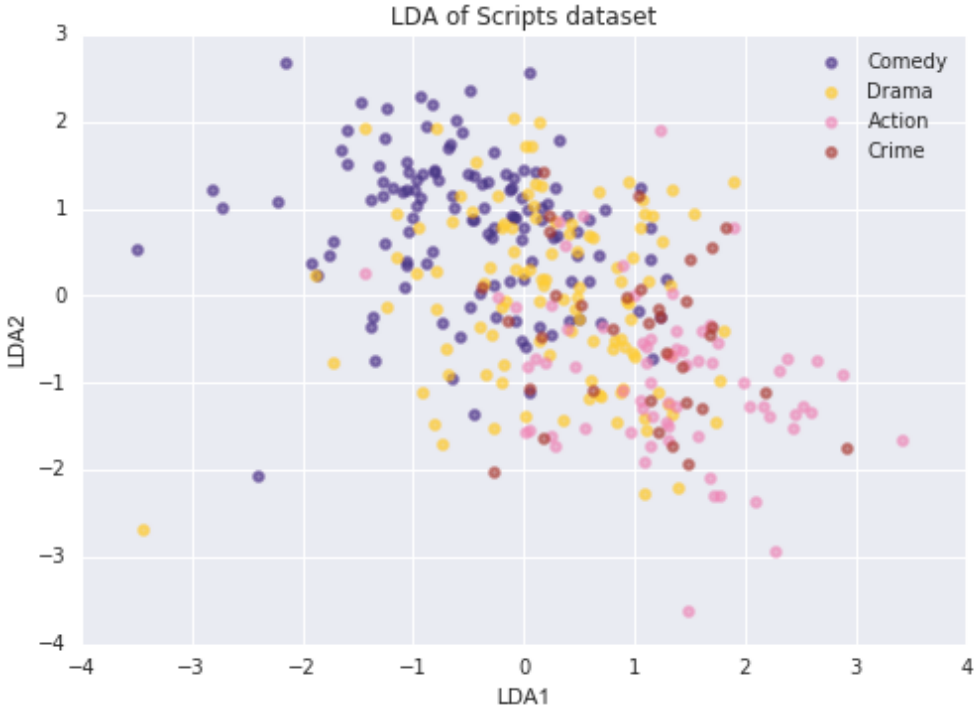
```
Title
In the French Style           0.990937
Under the Greenwood Tree      0.990831
Far from the Madding Crowd    0.990372
Name: Eat Pray Love, dtype: float64
```

For context:

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

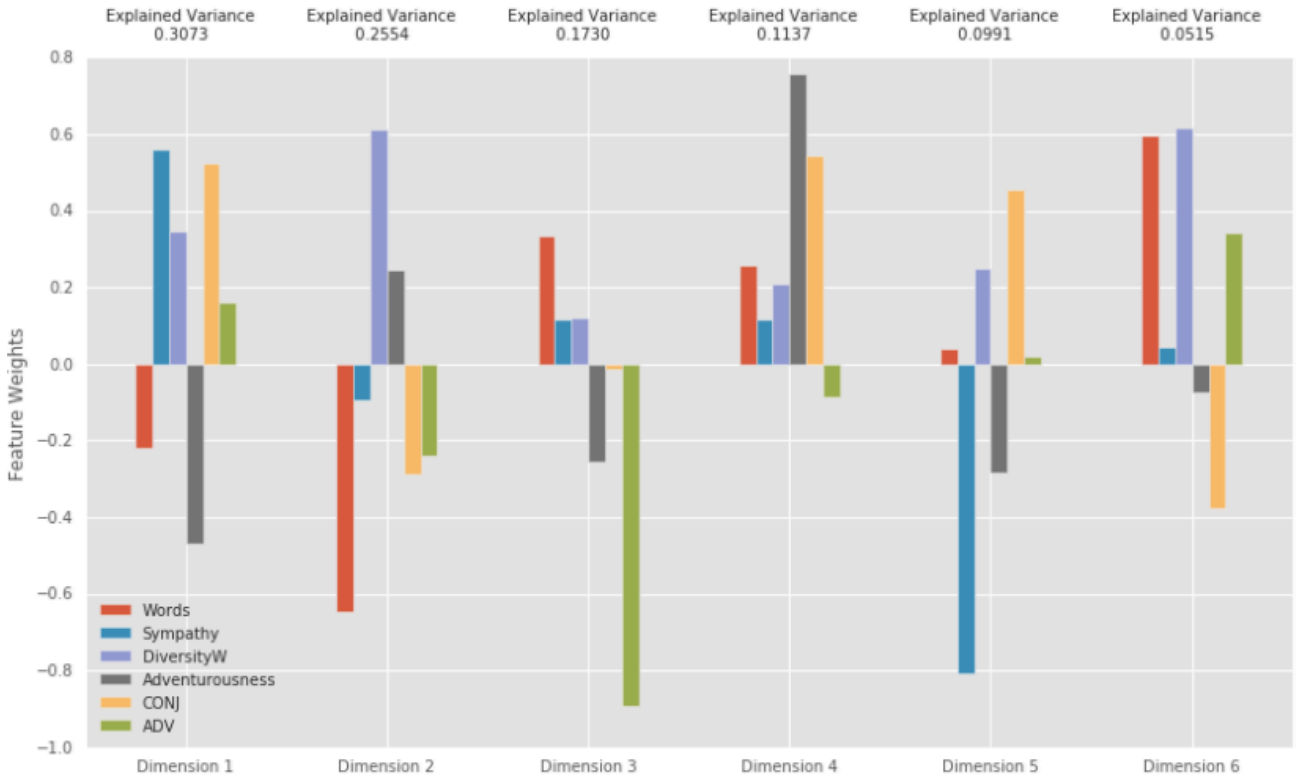
```
A young American art student must decide whether to stay in Paris with her boyfriend or go back to
 the U.S. when her wealthy father arrives to bring her back.
```

Classify



Current classification accuracy ~60%

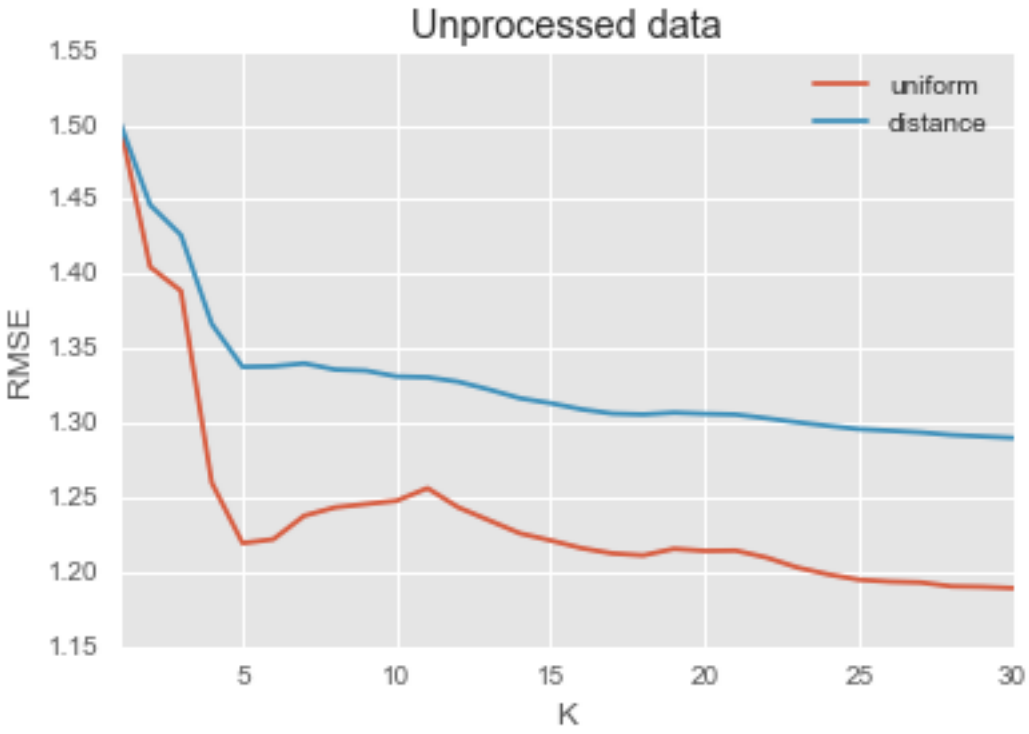
Describe



	Words	Sympathy	DiversityW	Adventurousness	Excitement-seeking	Immoderation	WordL
importance	0.087441	0.043237	0.040828	0.038594	0.027215	0.025561	0.025129

Importance as evaluated vs. ‘imdbRating’.

Evaluate




```
RMSE
RF: 0.962762, LR: 1.034057, SVM: 1.081481, ABR: 0.782190, KNR: 0.610792, Mean: 0.822118
```

Why?

- 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

- Data mining (Keep increasing dataset):
  - Web crawling
  - API's
  - Additional Books dataset
- Machine Learning (classification/regression):
  - Tune KNN, SVM, RF, AdaBoost
  - Implement Keras NN, xgBoost
- Develop web application:
  - User interface
    - Allow to submit text files
    - Allow to request for recommendation
    - Data visualization

Thank you.