AIM-3: IMDb Movie Classification

Florian Gößler & Xi Yang

Introduction

IMDb

Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes | Celebs, Events & Photos | News & Community | Watchlist

Contact the Filmmakers on IMDbPro »

Jurassic World (2015)

4

PG-13 | 124 min | Action, Adventure, Sci-Fi | 12 June 2015 (USA)

7.4

Your rating: -/10

Ratings: **7.4**/10 from 183,836 users Metascore: 59/100

Reviews: 900 user | 510 critic | 49 from Metacritic.com

A new theme park is built on the original site of Jurassic Park. Everything is going well until the park's newest attraction--a genetically modified giant stealth killing machine--escapes containment and goes on a killing spree.

Director: Colin Trevorrow

Writers: Rick Jaffa (screenplay), Amanda Silver (screenplay), 5 more credits »

Stars: Chris Pratt, Bryce Dallas Howard, Ty Simpkins | See full cast and crew »

+ Watchlist | Watch Trailer | Share...

Cast

Edit

Cast overview, first billed only:

Chris Pratt	...	Owen
Bryce Dallas Howard	...	Claire
Irrfan Khan	...	Masrani
Vincent D'Onofrio	...	Hoskins
Ty Simpkins	...	Gray
Nick Robinson	...	Zach
Jake Johnson	...	Lowery
Omar Sy	...	Barry
BD Wong	...	Dr. Henry Wu
Judy Greer	...	Karen
Lauren Lapkus	...	Vivian
Brian Tee	...	Hamada
Katie McGrath	...	Zara
Andy Buckley	...	Scott
Eric Edelstein	...	Paddock Supervisor

See full cast »

Storyline

Edit

22 years after the original Jurassic Park failed, the new park (also known as Jurassic World) is open for business. After years of studying genetics the scientists on the park genetically engineer a new breed of dinosaur. When everything goes horribly wrong, will our heroes make it off the island?

Plot Summary | Add Synopsis

Plot Keywords: jurassic park | island | dinosaur | sea creature | experiment gone wrong | See All (215) »

Taglines: The park is open.

Genres: Action | Adventure | Sci-Fi | Thriller

Motion Picture Rating (MPAA)

Rated PG-13 for intense sequences of science-fiction violence and peril | See all certifications »

Parents Guide: View content advisory »

Problem Statement

To build a scalable movie genre classification system using Internet Movie Database(IMDb)

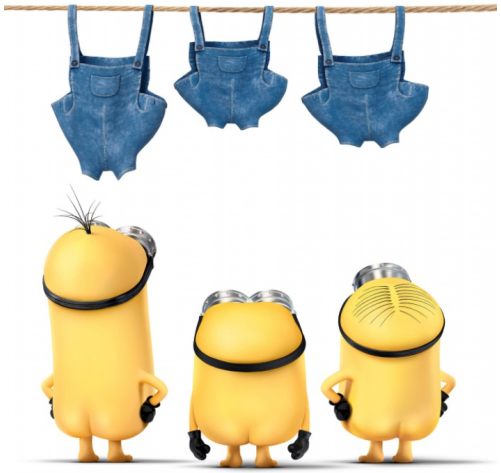
- What is a movie genre?
- What kind of information is sufficient to classify the movies?

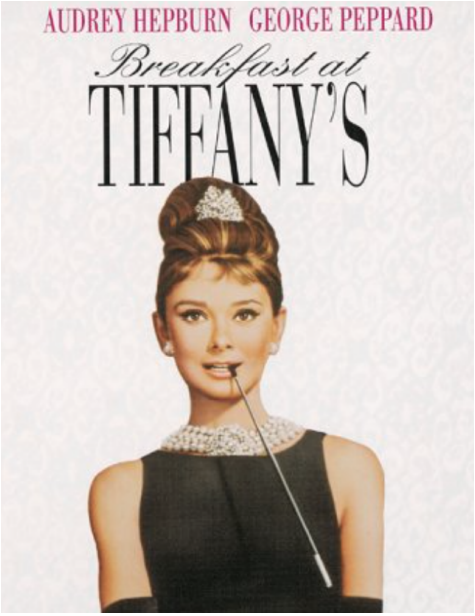
Problem Statement

- What is a movie genre?

Problem Statement

- What is a movie genre?





Animation	Romance
Comedy	Comedy
Family	Drama

Problem Statement

- What kind of information is sufficient to classify the movies?

  Plot description?
  Actors?
  Key words?

Dataset

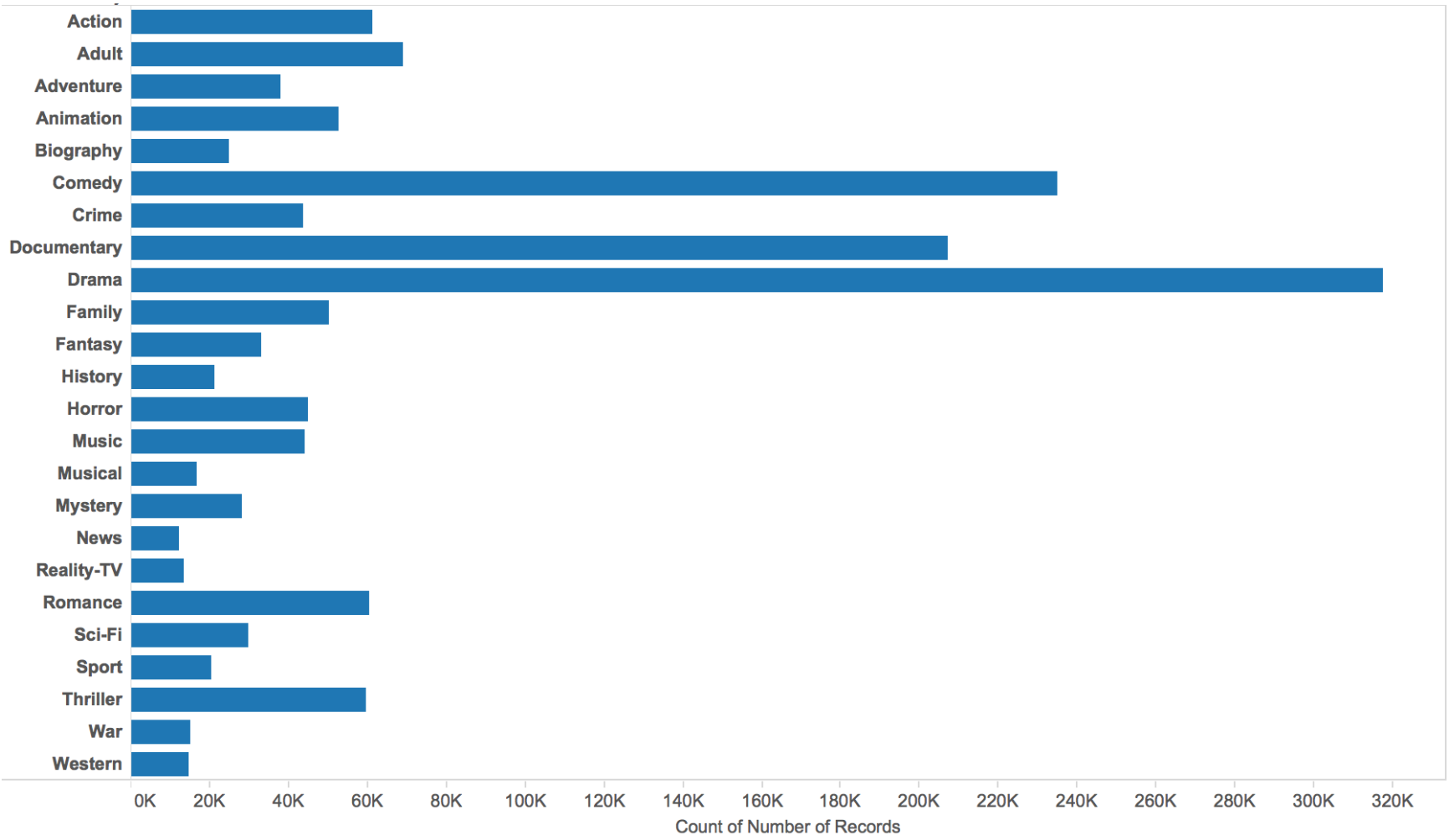
the Internet Movie Database(IMDb):

- publicly available data
- provides all kinds of information related to movies
- size: 1.7GB
- formatted difficult to use

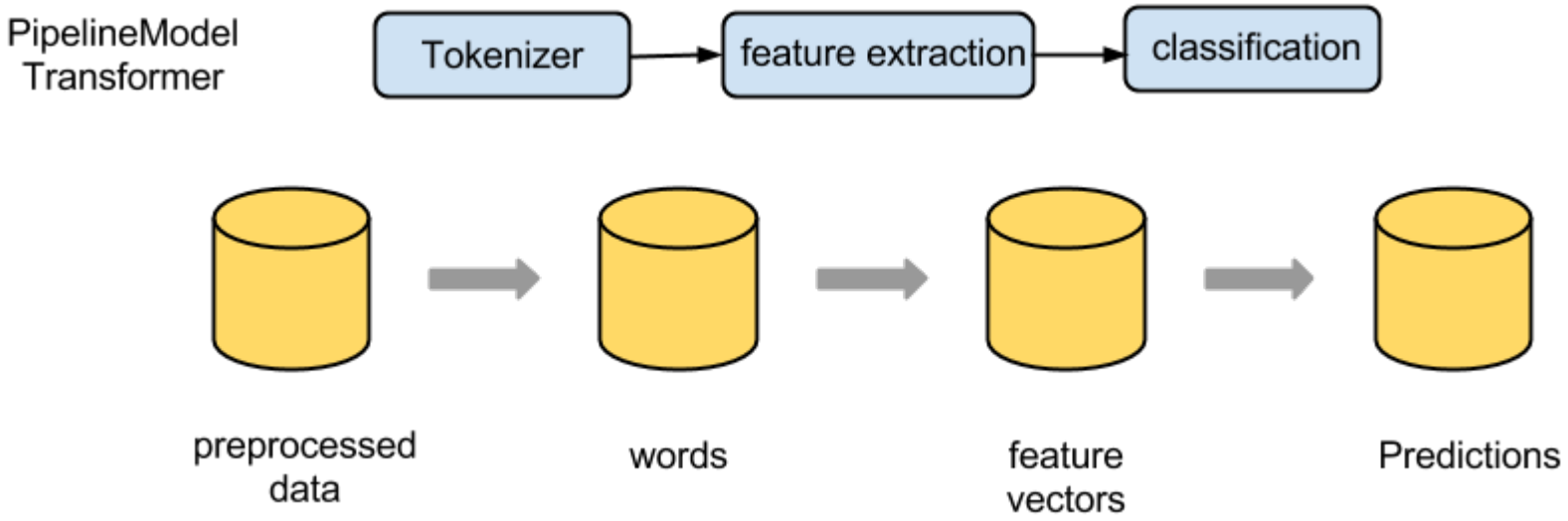
TODO: show data samples before and after preprocessing

Dataset

Dataset


Action
Adult
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
News
Reality-TV
Romance
Sci-Fi
Sport
Thriller
War
Western
0K
20K
40K
60K
80K
100K
120K
140K
160K
180K
200K
220K
240K
260K
280K
300K
320K
Count of Number of Records

Methodology



- feature extraction: TF-IDF, Binary Vectors (TF)
- classification: Support Vector Machines(SVMs)

TF-IDF

- Term Frequency

  counts number of times each term occurs in each document

- Inverse Document Frequency:

  measures how important the term provides

$$IDF(t, D) = \log \frac{|D| + 1}{DF(t, D) + 1}$$

$$TFIDF(t, d, D) = TF(t, d) \cdot IDF(t, D)$$

TF-IDF

document1	the	game	of	life	is	a	overlasting	learning
TF	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.1

document2	the	unexamined	life	is	not	worth	living
TF	0.143	0.143	0.143	0.143	0.143	0.143	0.143

document3	never	stop	learning
TF	0.333	0.333	0.333

TF-IDF

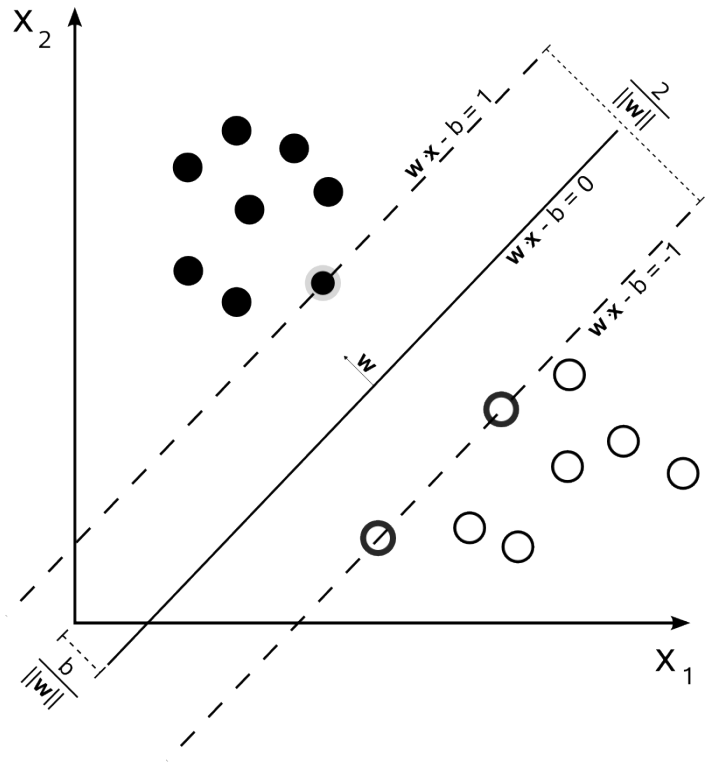
Terms	IDF	Terms	IDF
the	1.405507153	learning	1.405507153
game	2.098726209	unexamined	2.098726209
of	2.098726209	not	2.098726209
life	1.405507153	worth	2.098726209
is	1.405507153	living	2.098726209
a	2.098726209	never	2.098726209
everlasting	2.098726209	stop	2.098726209

TF-IDF

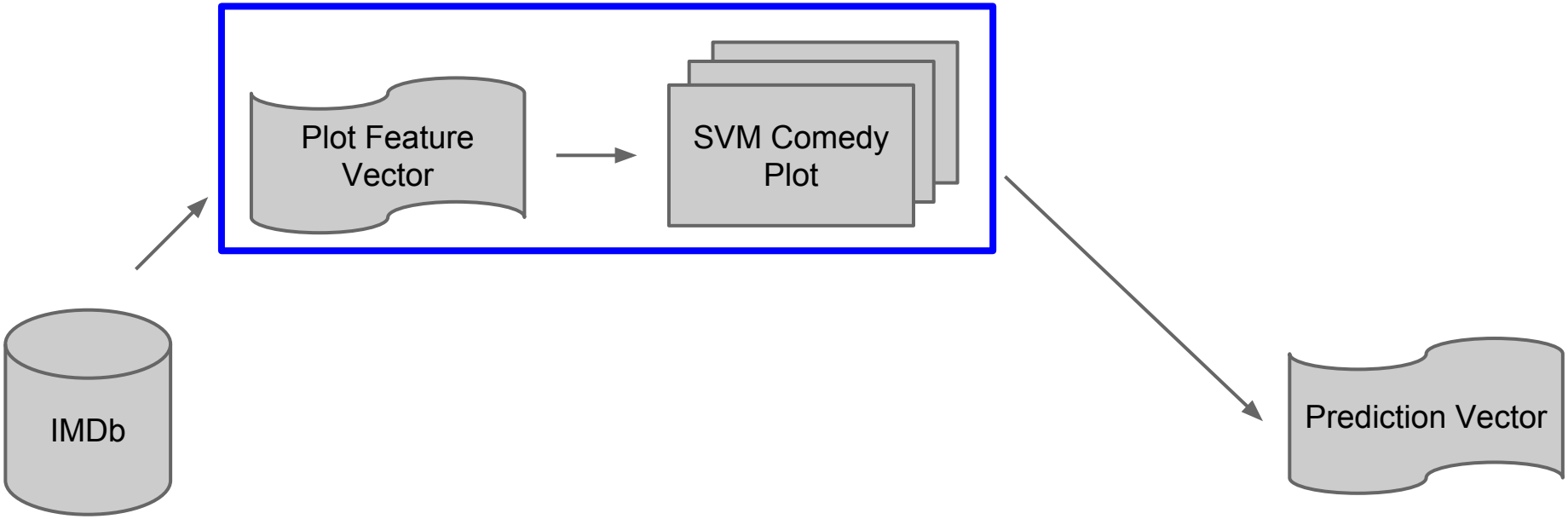
	Document1	Document2	Document3
life	0.140550715	0.200786736	0
learning	0.140550715	0	0.468502384

Support Vector Machine

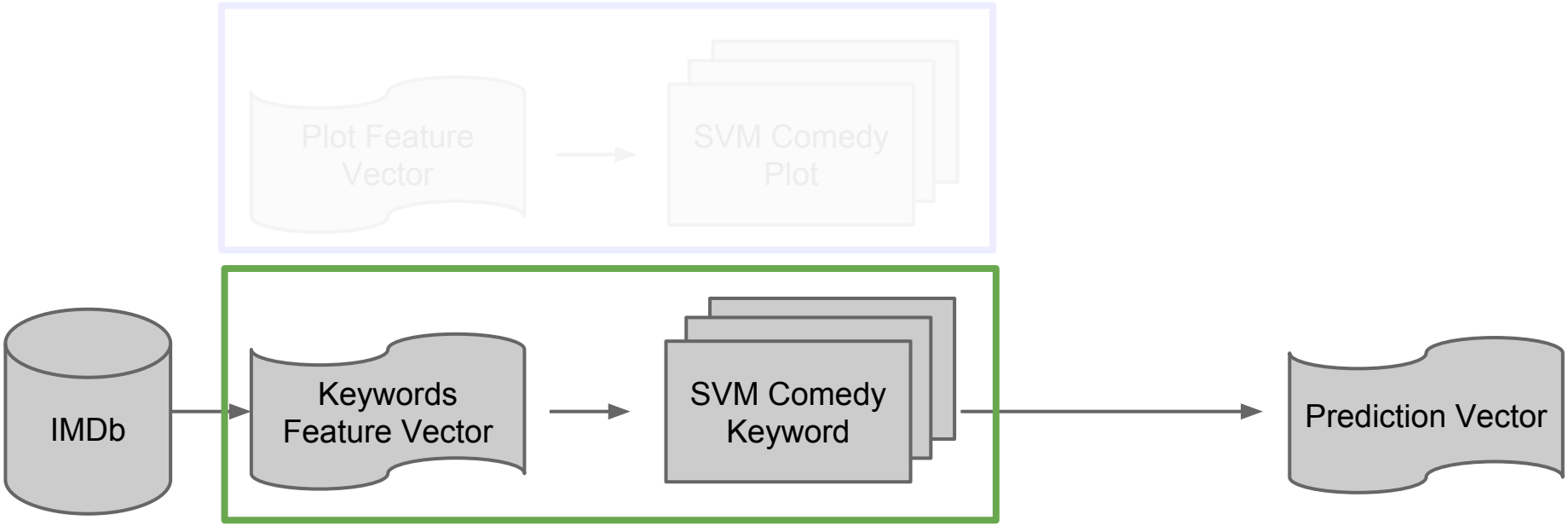
- supervised learning algorithm
- binary classification
- to find the maximum margin hyperplane



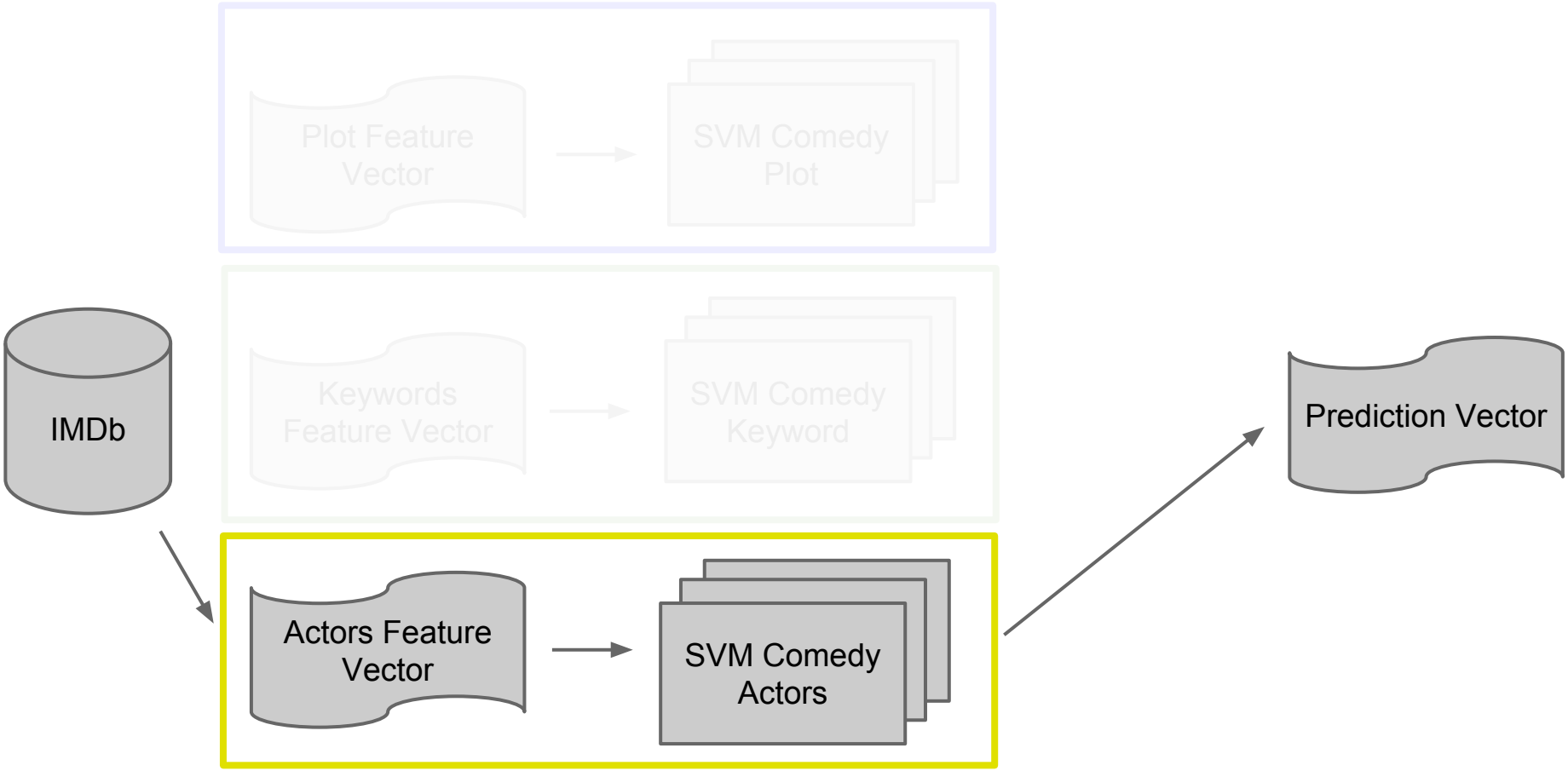
Experiments



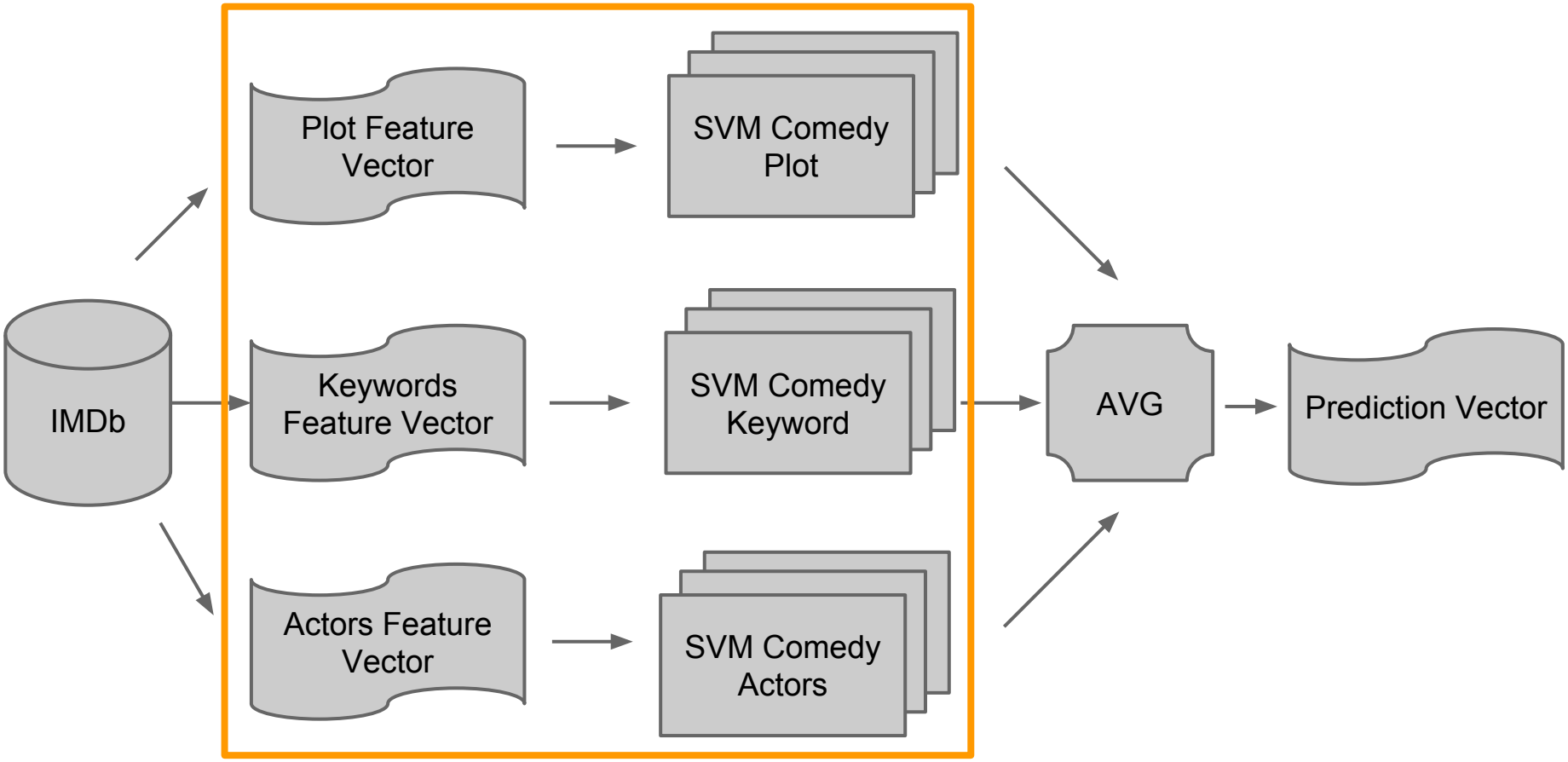
Experiments



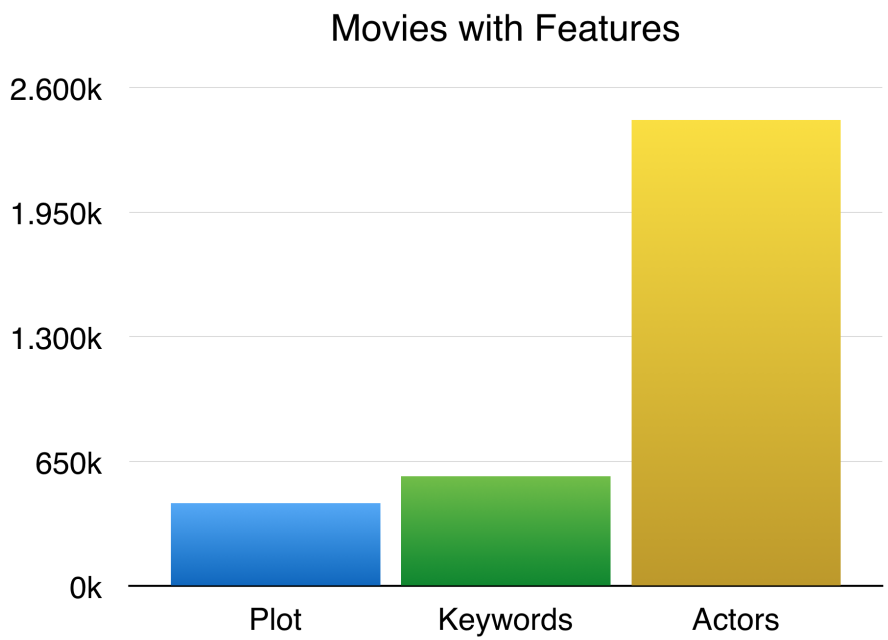
Experiments



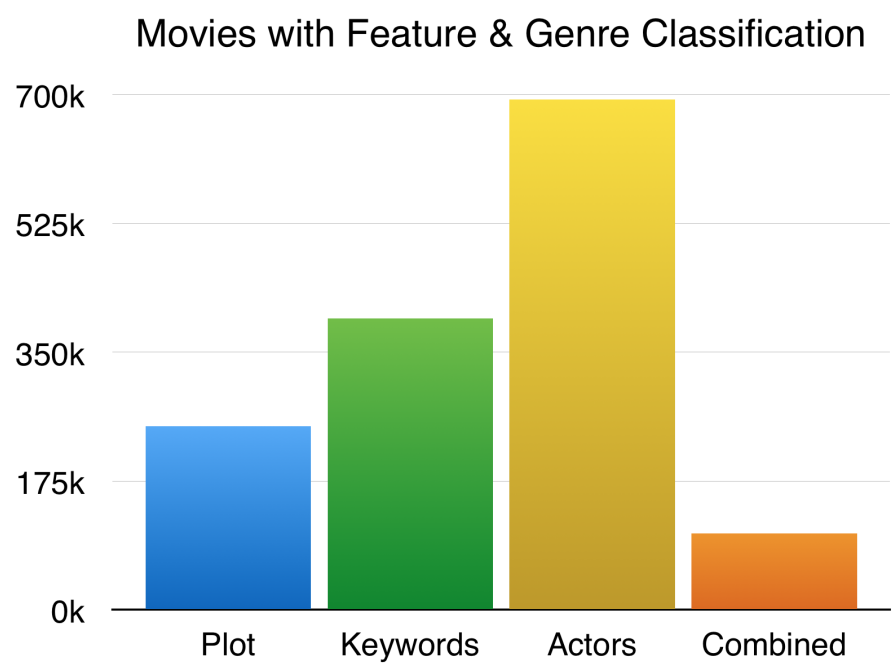
Experiments



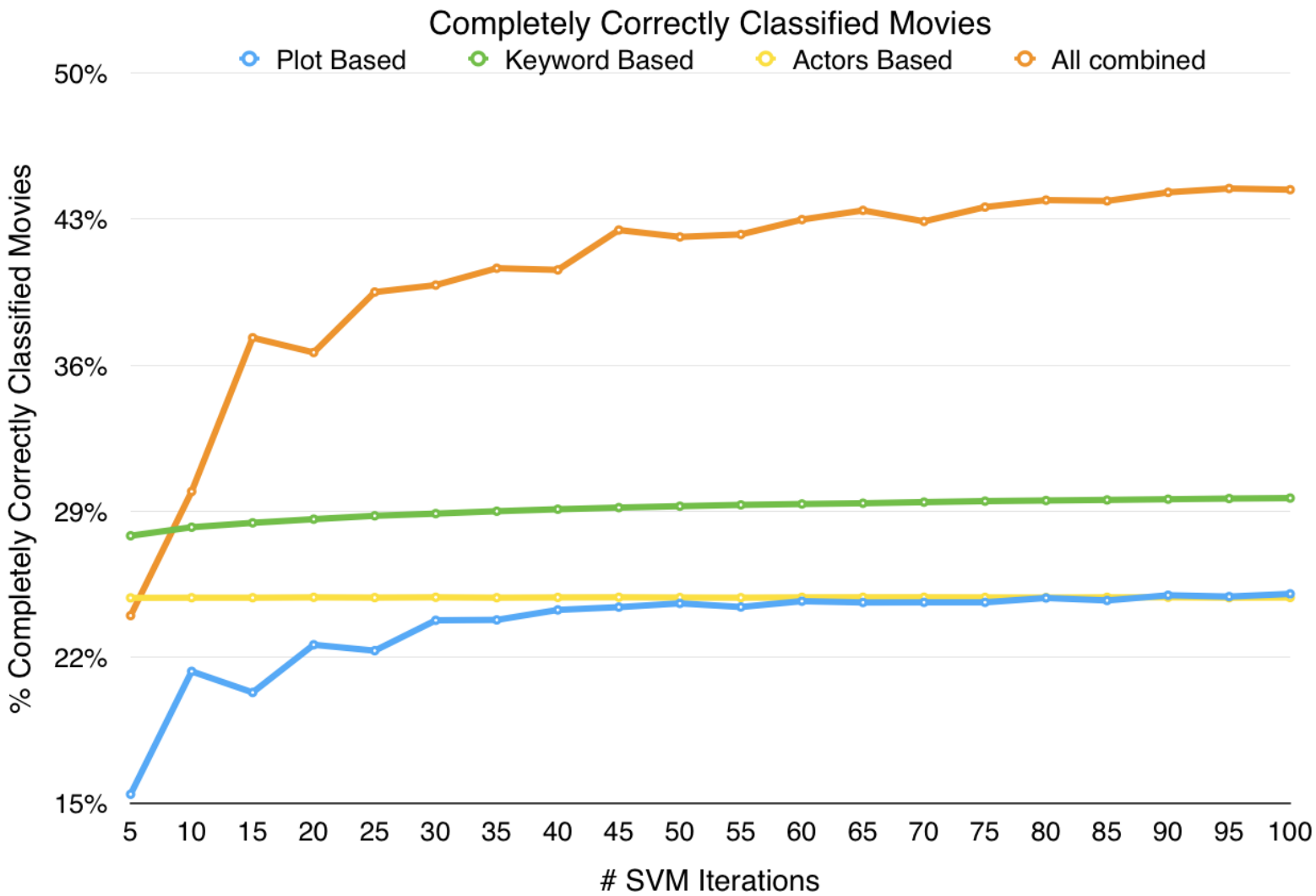
Experiments - Data



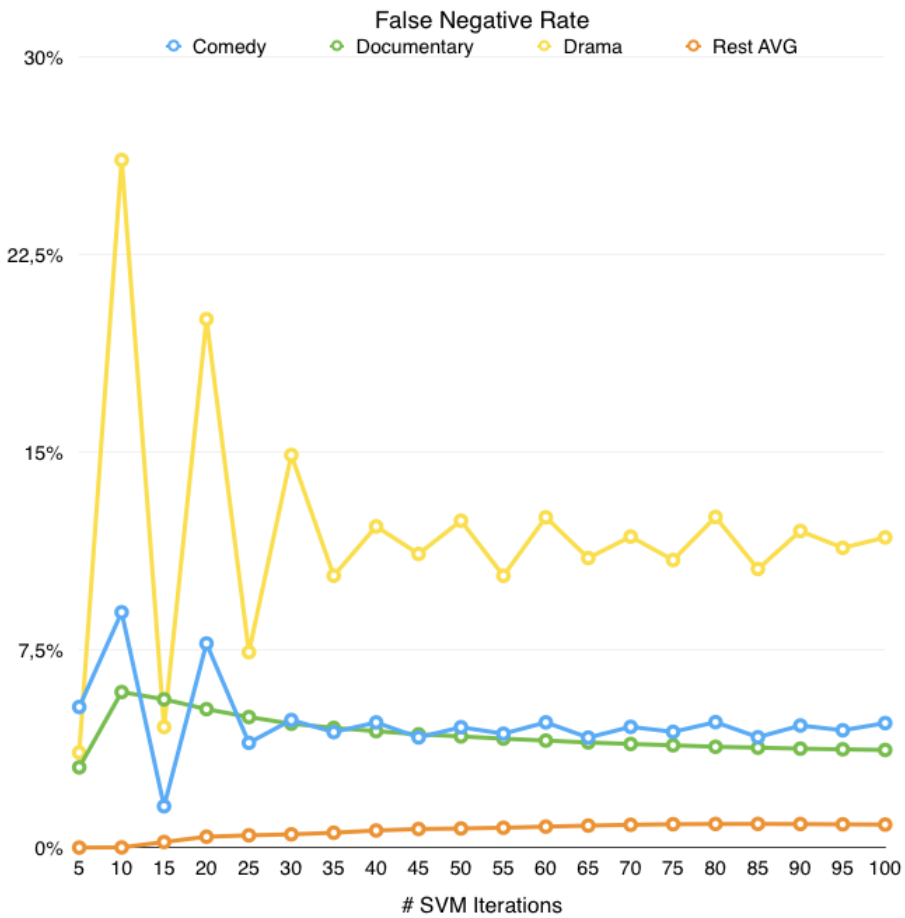
1033k with genre data

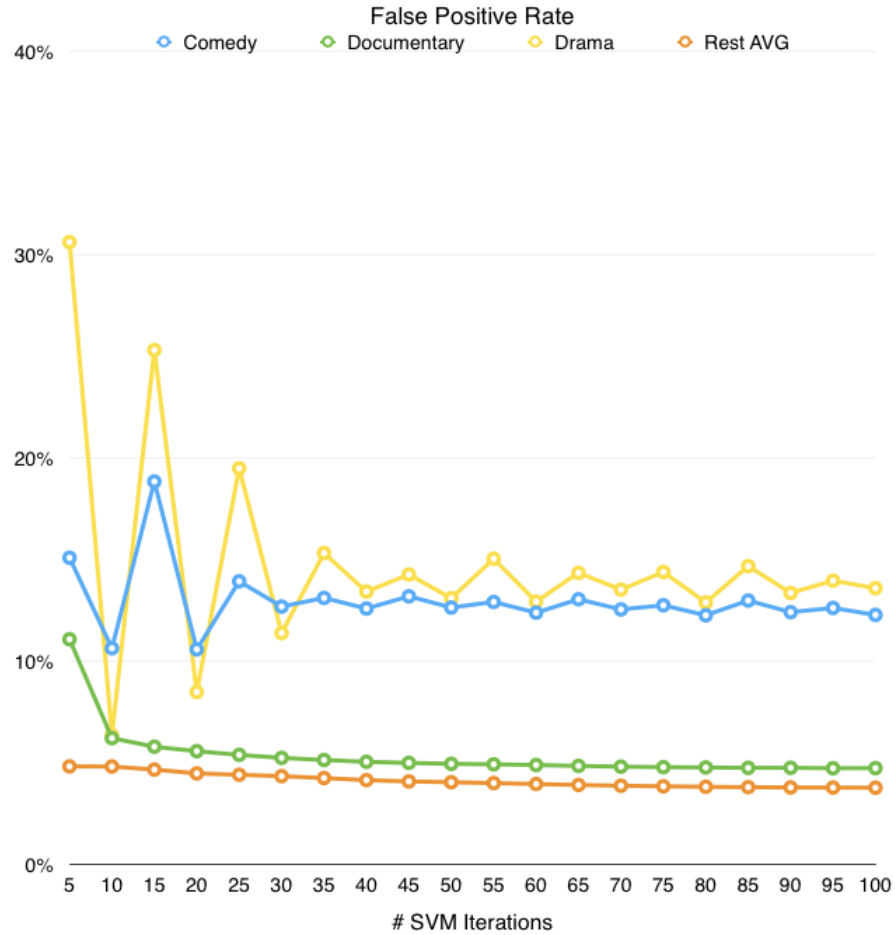


Results

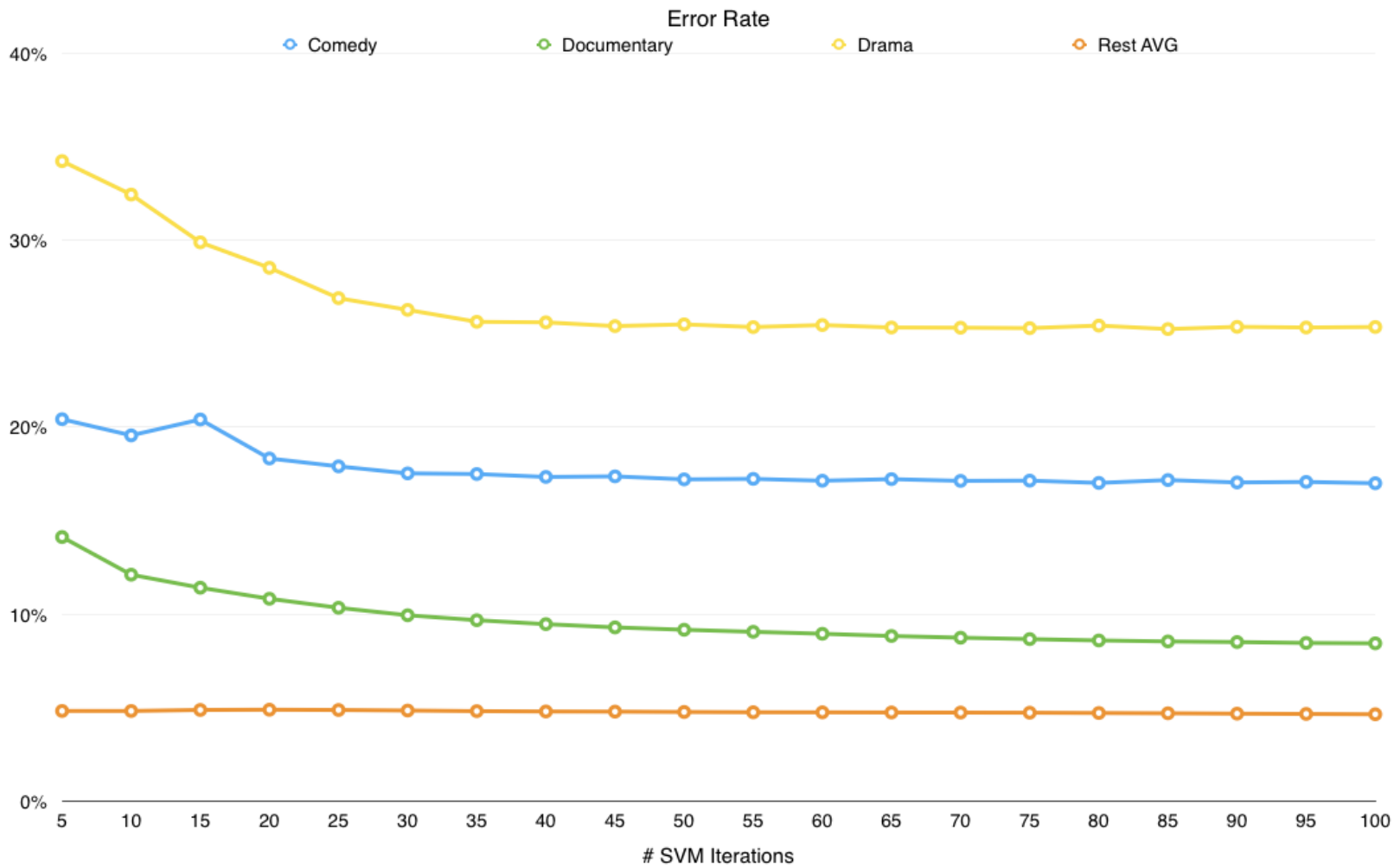

Completely Correctly Classified Movies
Plot Based
Keyword Based
Actors Based
All combined
50%
43%
36%
29%
22%
15%
% Completely Correctly Classified Movies
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
SVM Iterations

Results - Plot Based

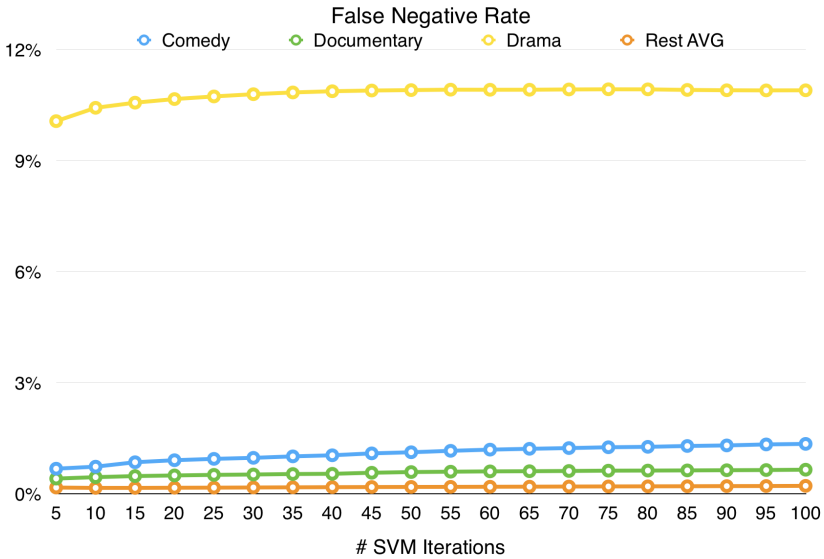


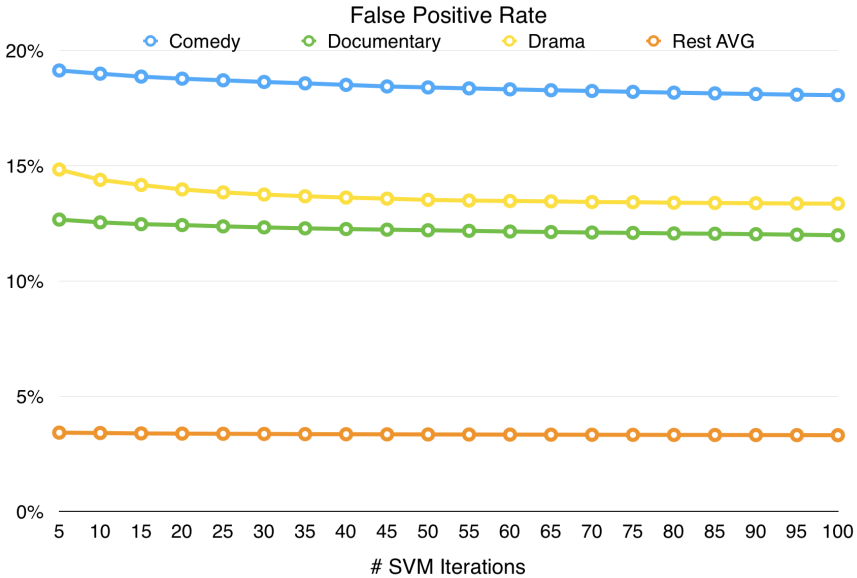


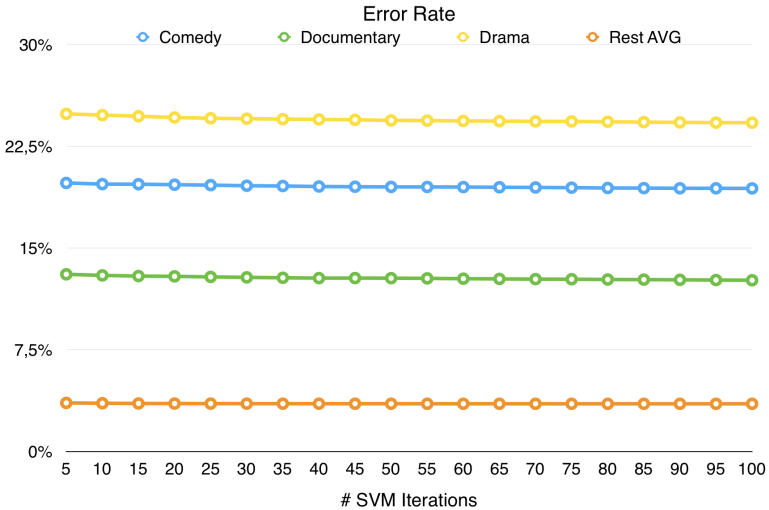
Results - Plot Based



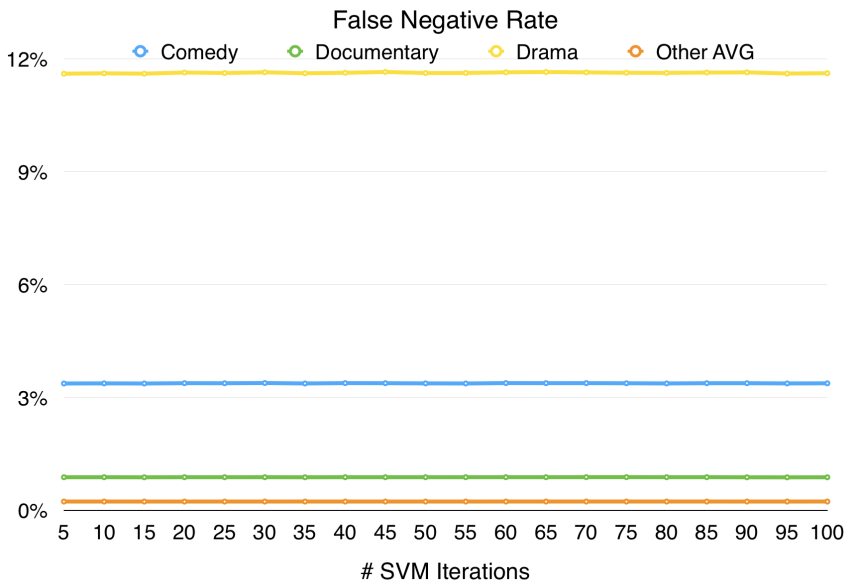
Results - Keyword Based

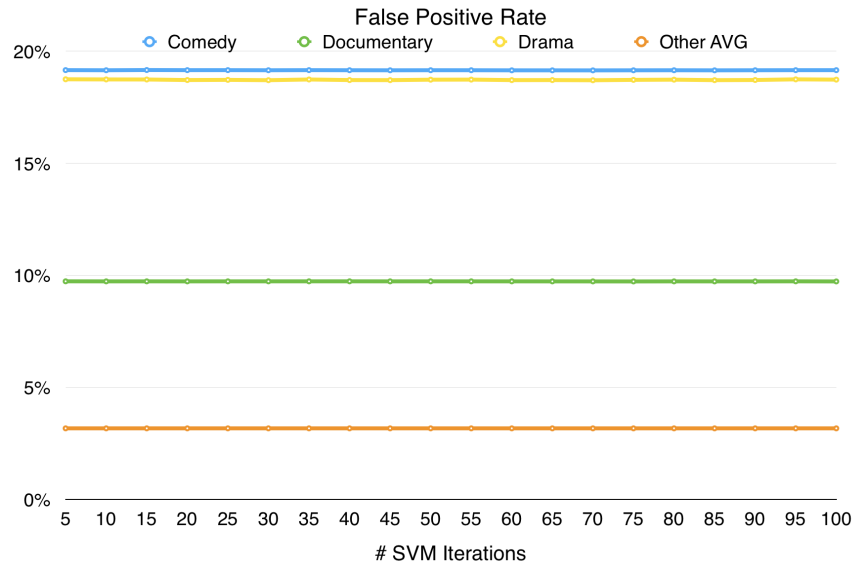


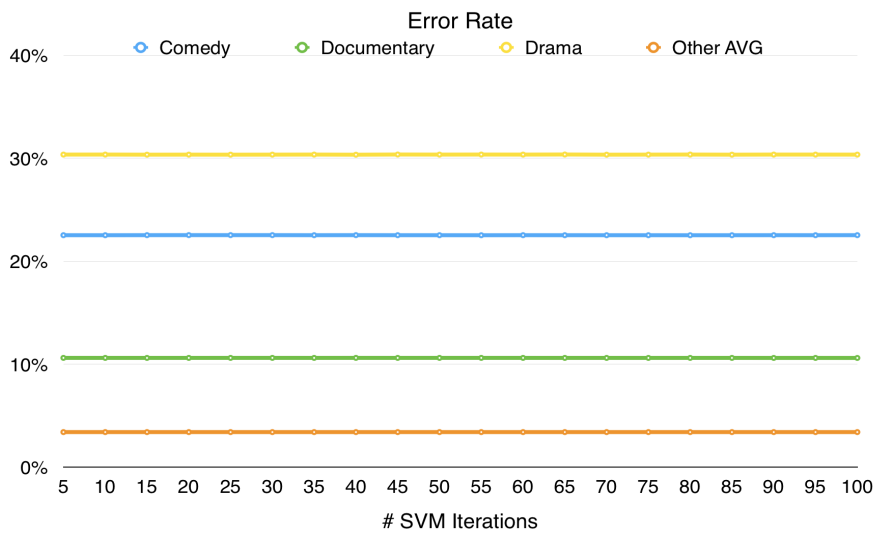




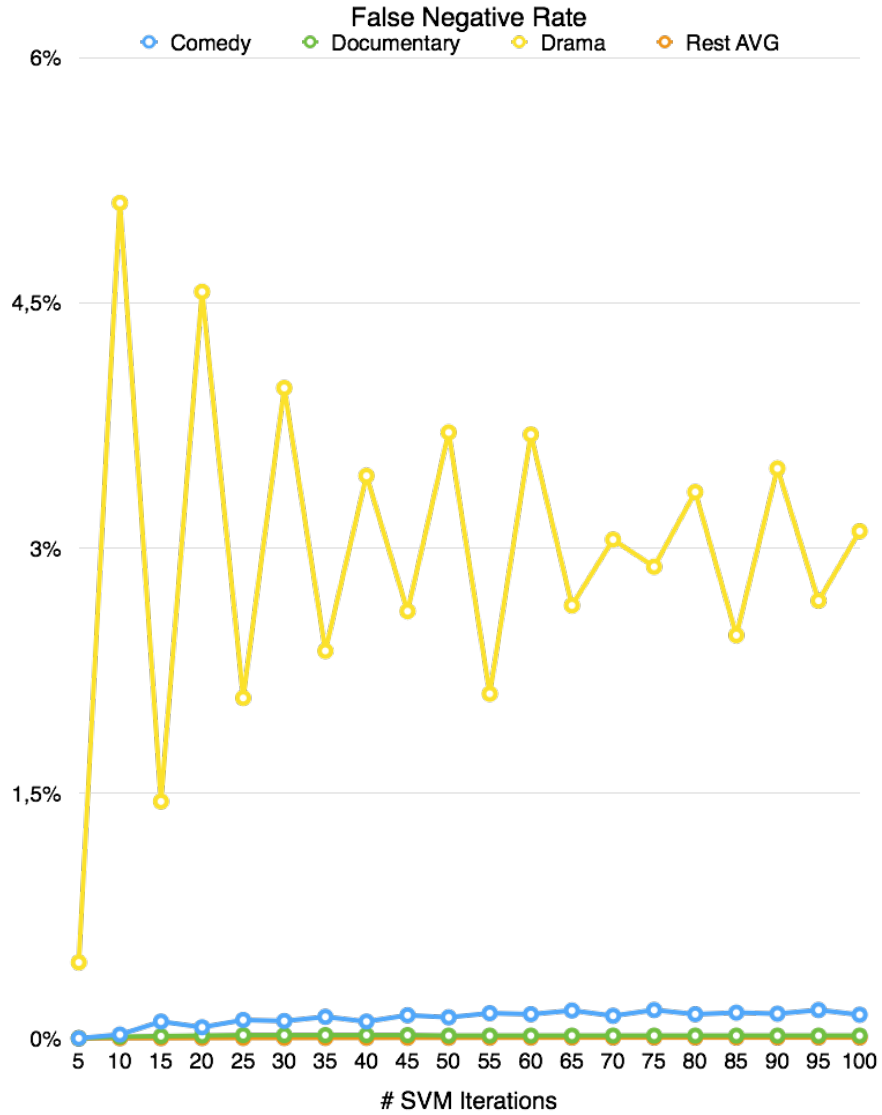
Results - Actors Based

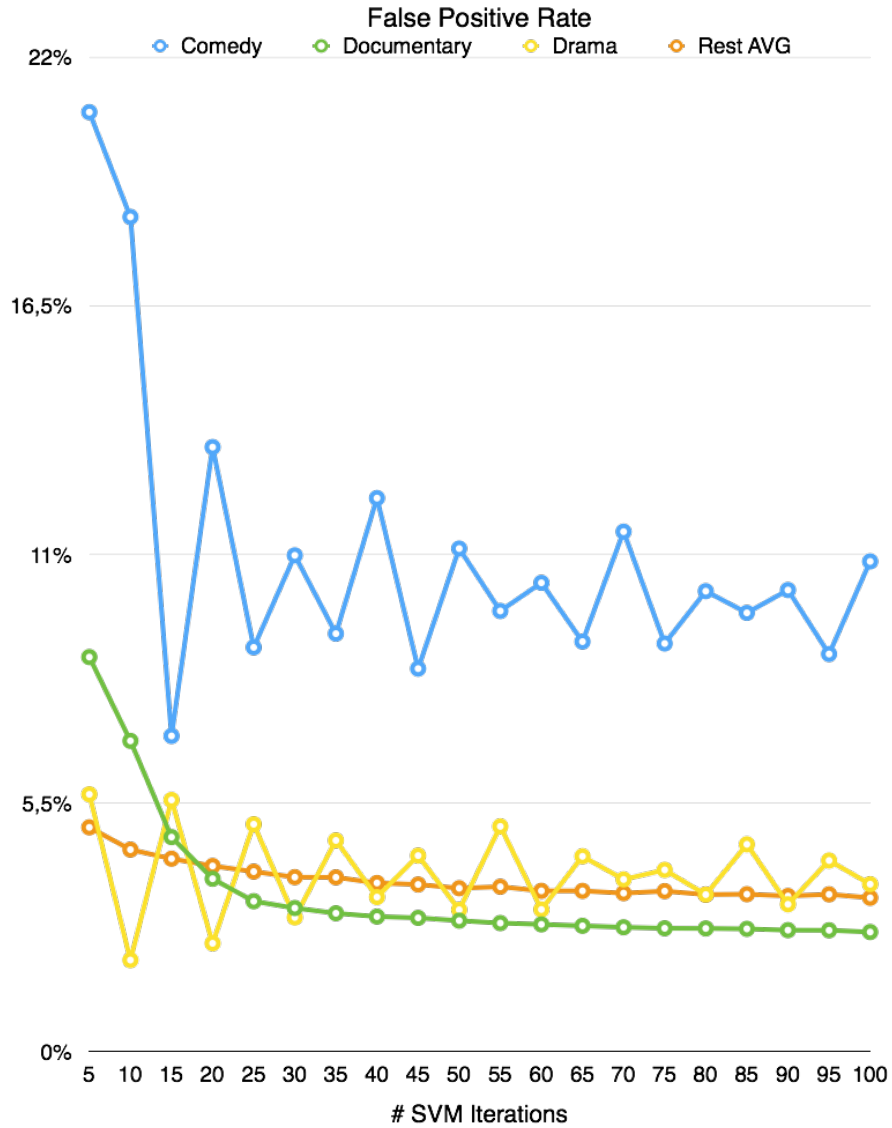




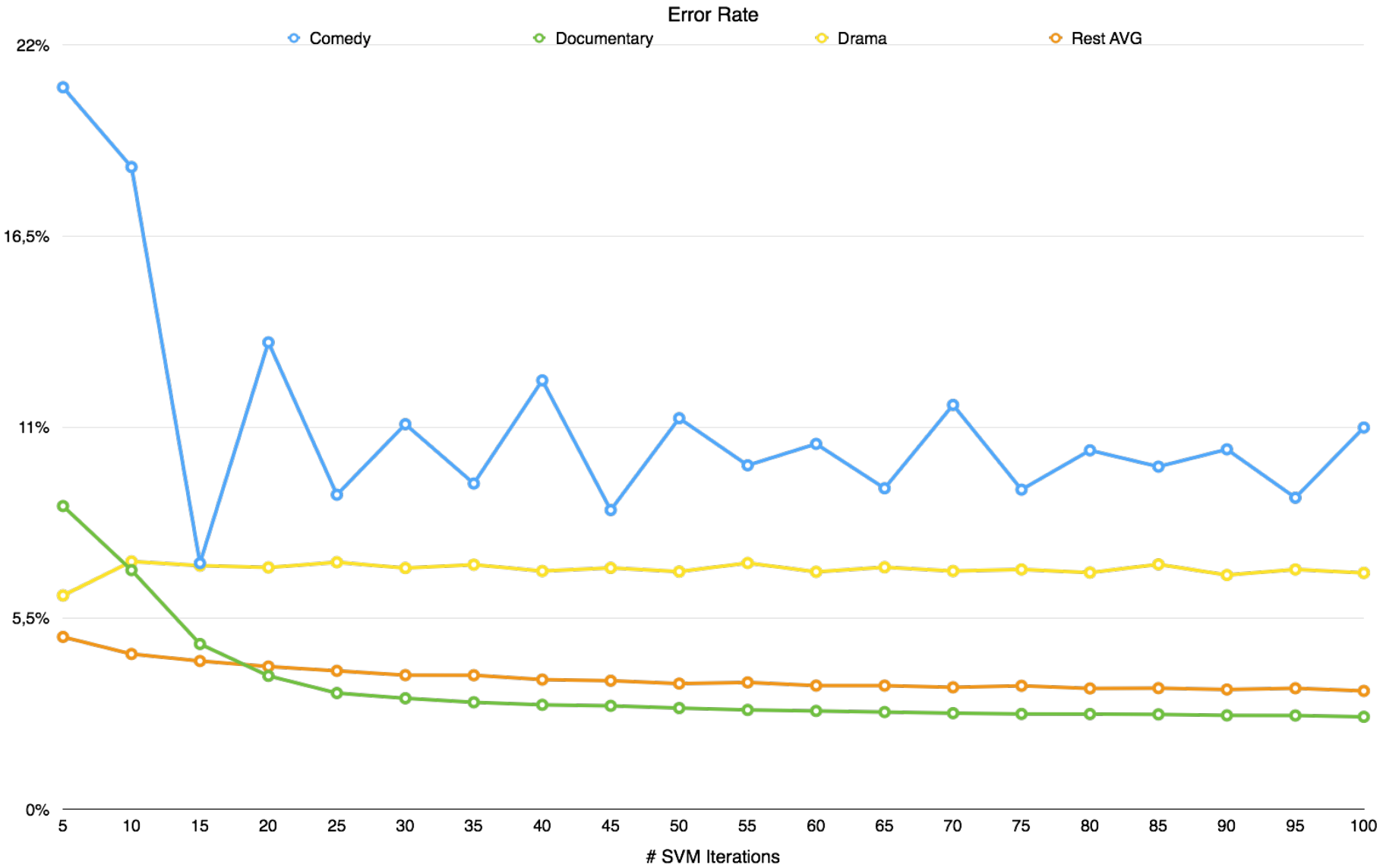


Results - Combined Features

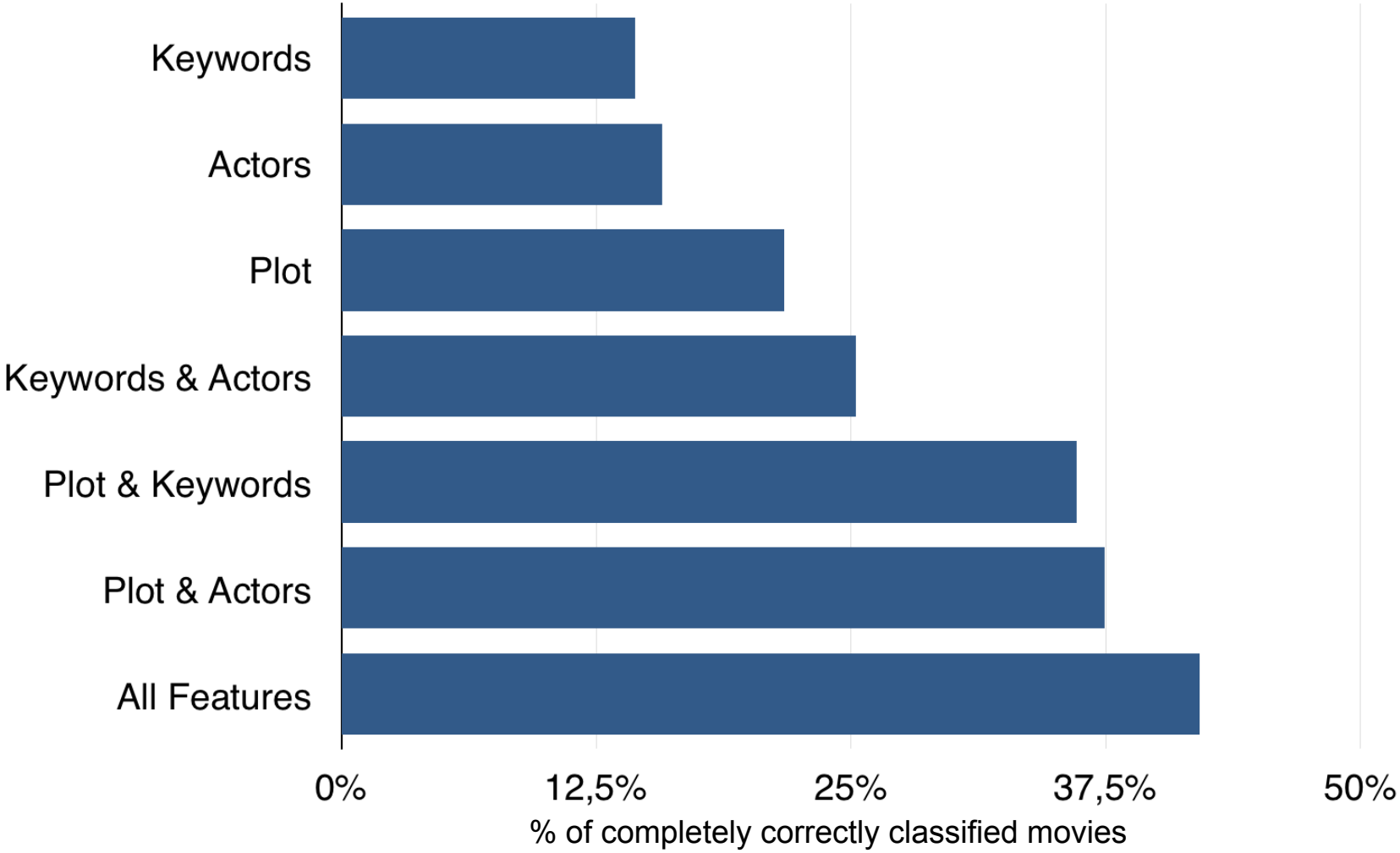




Results - Combined Features



Results - Combined Features



trained with 50 iterations on the subset of ~100k movies which provide all 3 features

Conclusion

- Classification not good enough
- Feature combination is better
- Data Quality is important

- SVMs are powerful and comparatively easy to use
- SVM training time vs precision trade-off

- zick-zack curves are mysterious

- Spark has a nice & stable API with good basic ML tools
- other tools for preprocessing
- IMDb is no “real Big Data”