


AIM-3: IMDb Movie Classification

Florian Gößler & Xi Yang

Introduction



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

All

Movies, TV & Showtimes

Celebs, Events & Photos

News & Community

Watchlist

Unbegrenzter Film- und Seriengenuss mit Prime Instant Video
Jetzt 30 Tage testen.



Contact the Filmmakers on IMDbPro »

Jurassic World (2015)

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



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](#) »

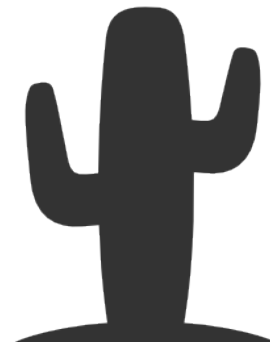
Problem Statement

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?

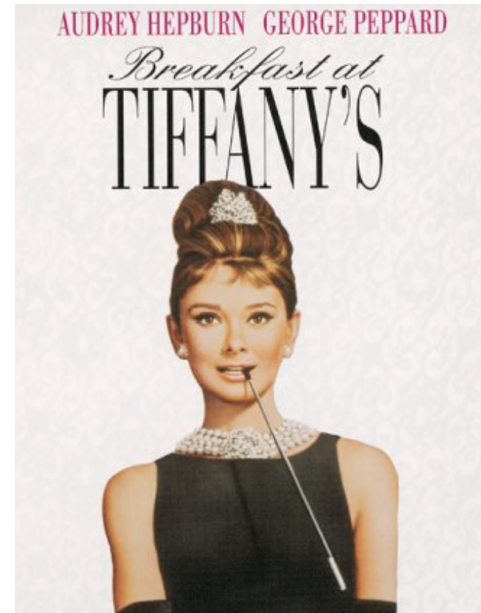


Problem Statement

- What is a movie genre?



Animation
Comedy
Family



Romance
Comedy
Drama

Problem Statement

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

Plot description?

Actors?

Key words?

Dataset

the Internet Movie Database(IMDb):

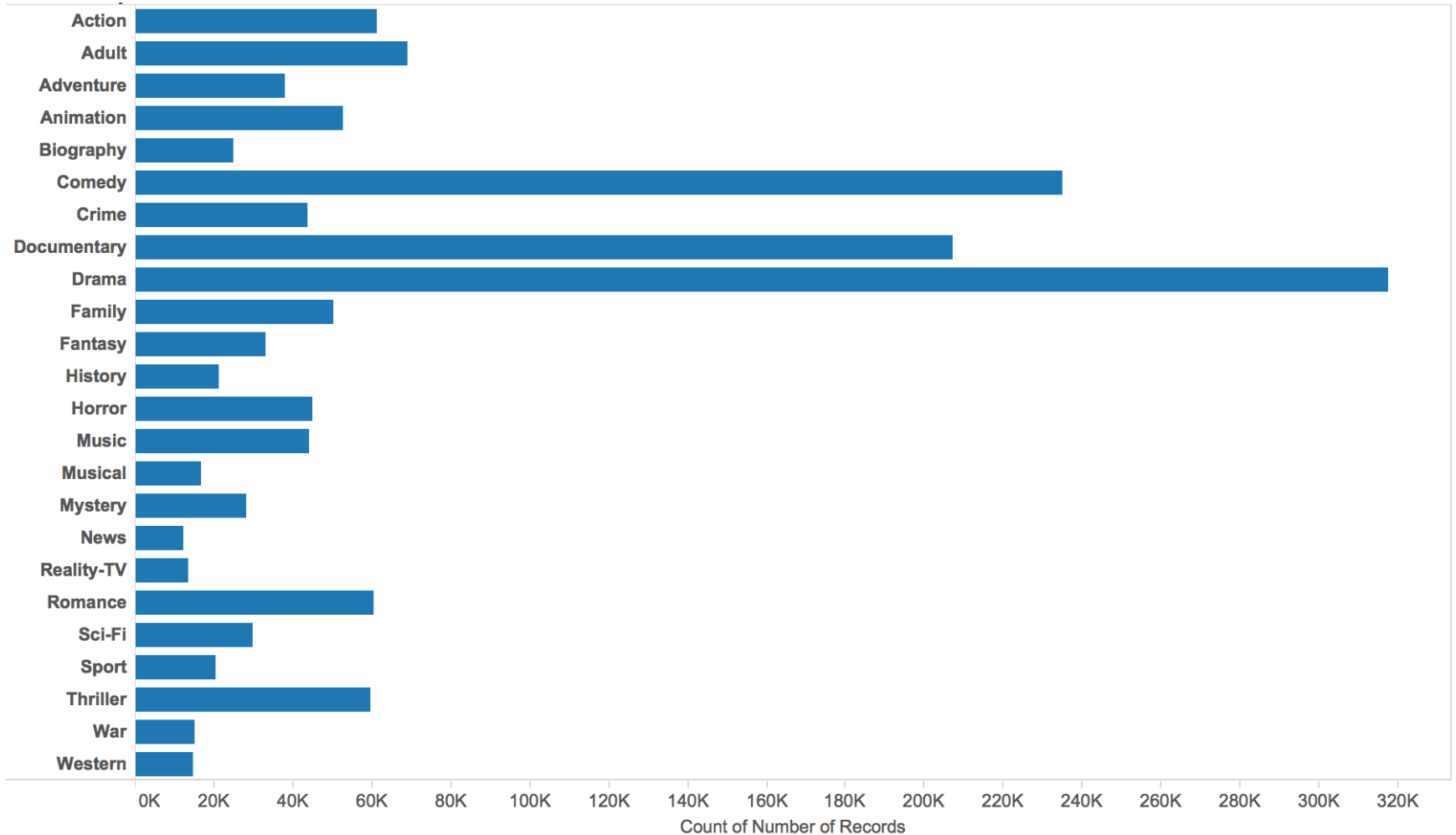
- publicly available data
- provides all kinds of information related to movies
- size: 1.7GB
- formatted difficult to use

Dataset

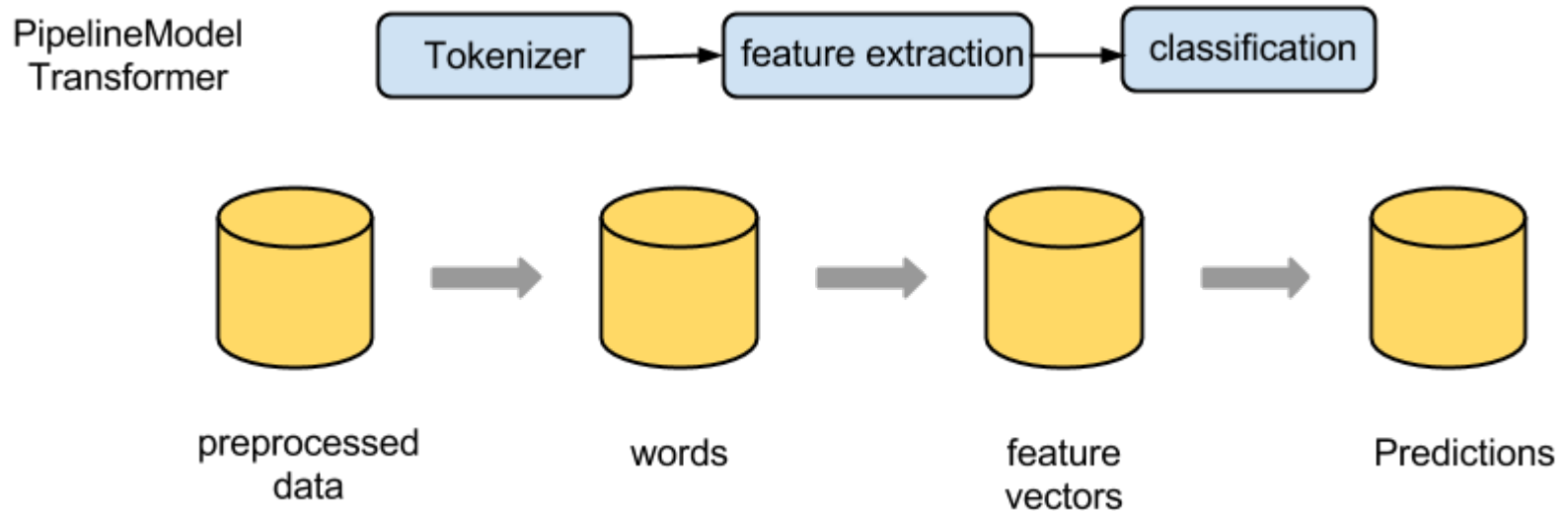
```
13  PLOT SUMMARIES LIST
14  =====
15  -----
16  MV: "#7DaysLater" (2013)
17
18  PL: #7dayslater is an interactive comedy series featuring an ensemble cast of
19  PL: YouTube celebrities. Each week the audience writes the brief via social
20  PL: media for an all-new episode featuring a well-known guest-star. Seven days
21  PL: later that week's episode premieres on TV and across multiple platforms.
22
23  BY: Sum Whan
24
25  -----
```

MovieName | PlotDescription

Dataset



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

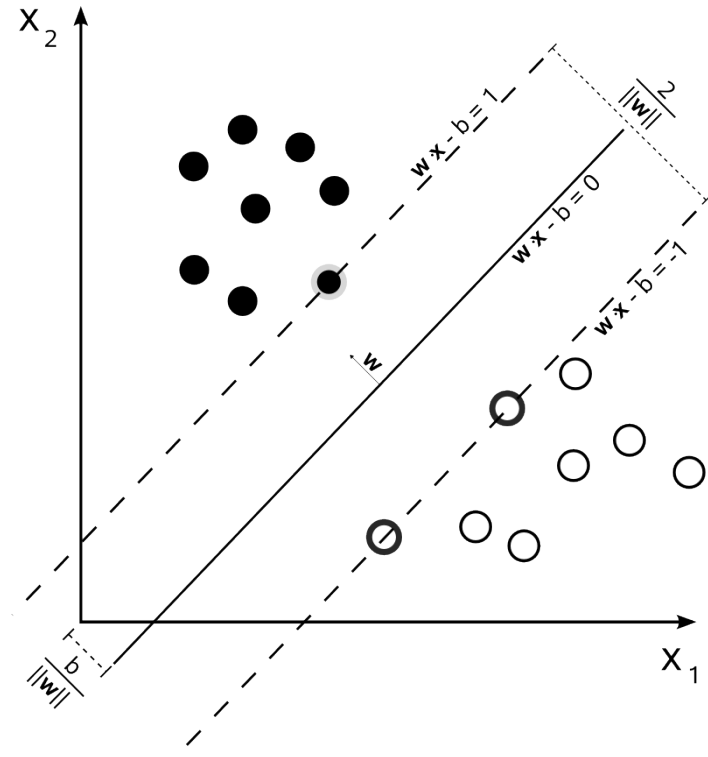
Terms	IDF	Terms	IDF
the	1.405507153	learning	1.405507153
game	2.098726209	unexamined	2.098726209
of	2.098726209	not	2.098726209
life	1.405507153	worth	2.098726209
is	1.405507153	living	2.098726209
a	2.098726209	never	2.098726209
everlasting	2.098726209	stop	2.098726209

TF-IDF

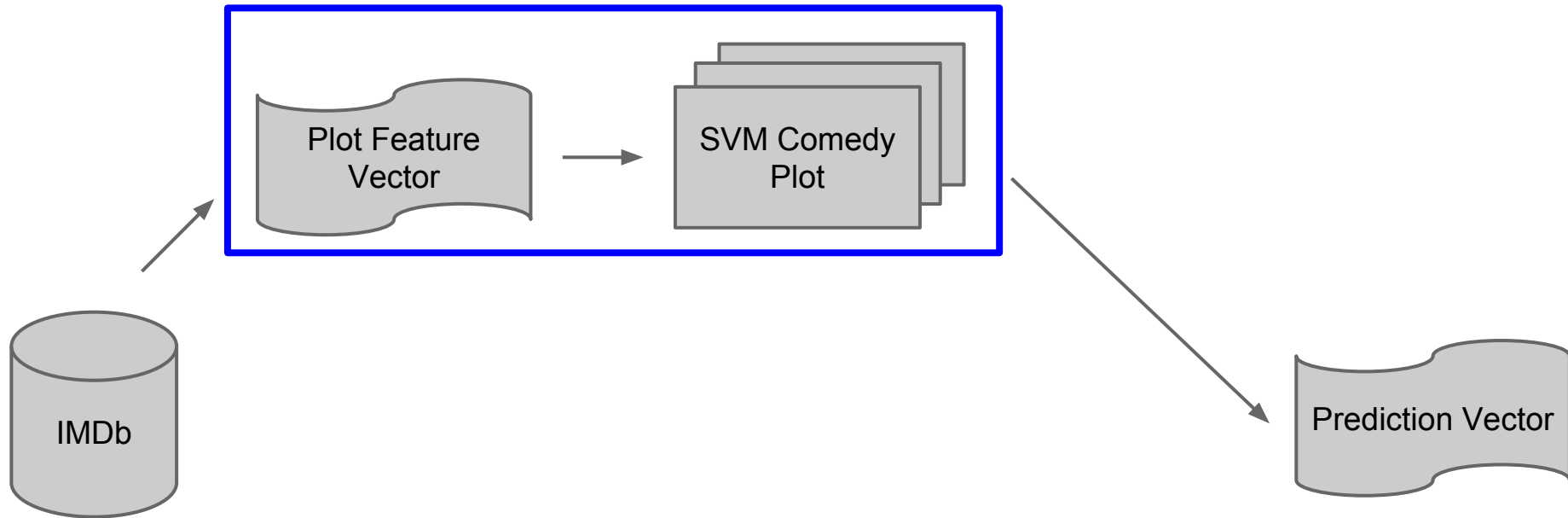
	Document1	Document2	Document3
life	0.140550715	0.200786736	0
learning	0.140550715	0	0.468502384

Support Vector Machine

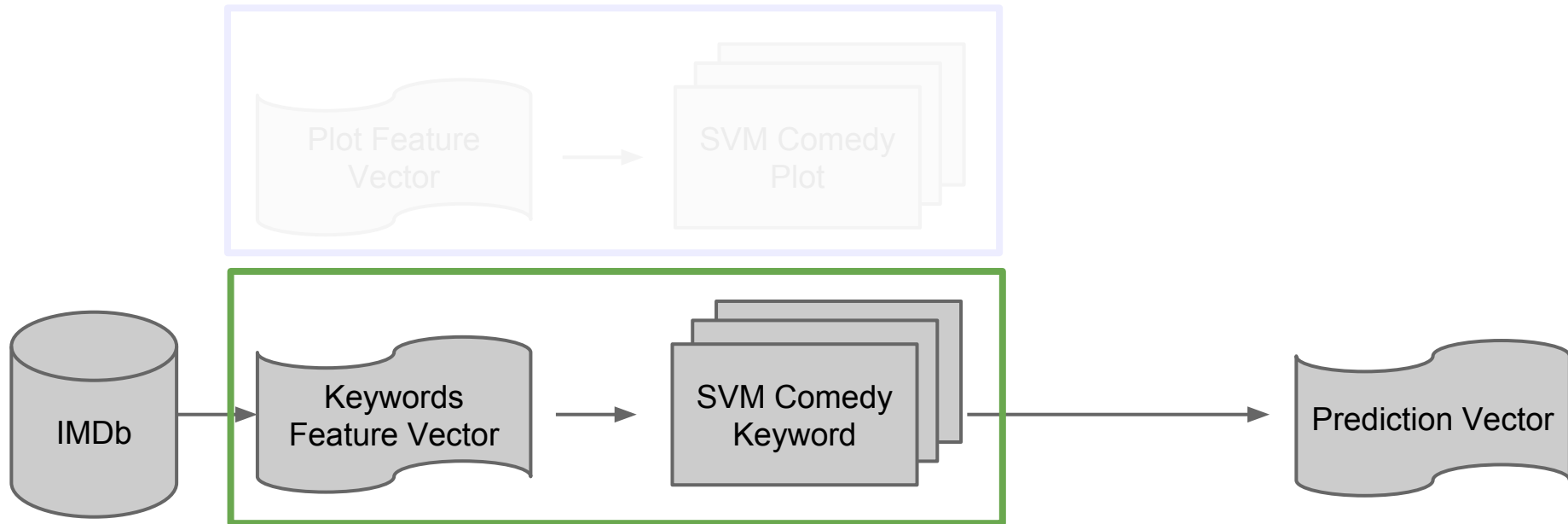
- supervised learning algorithm
- binary classification
- to find the maximum margin hyperplane



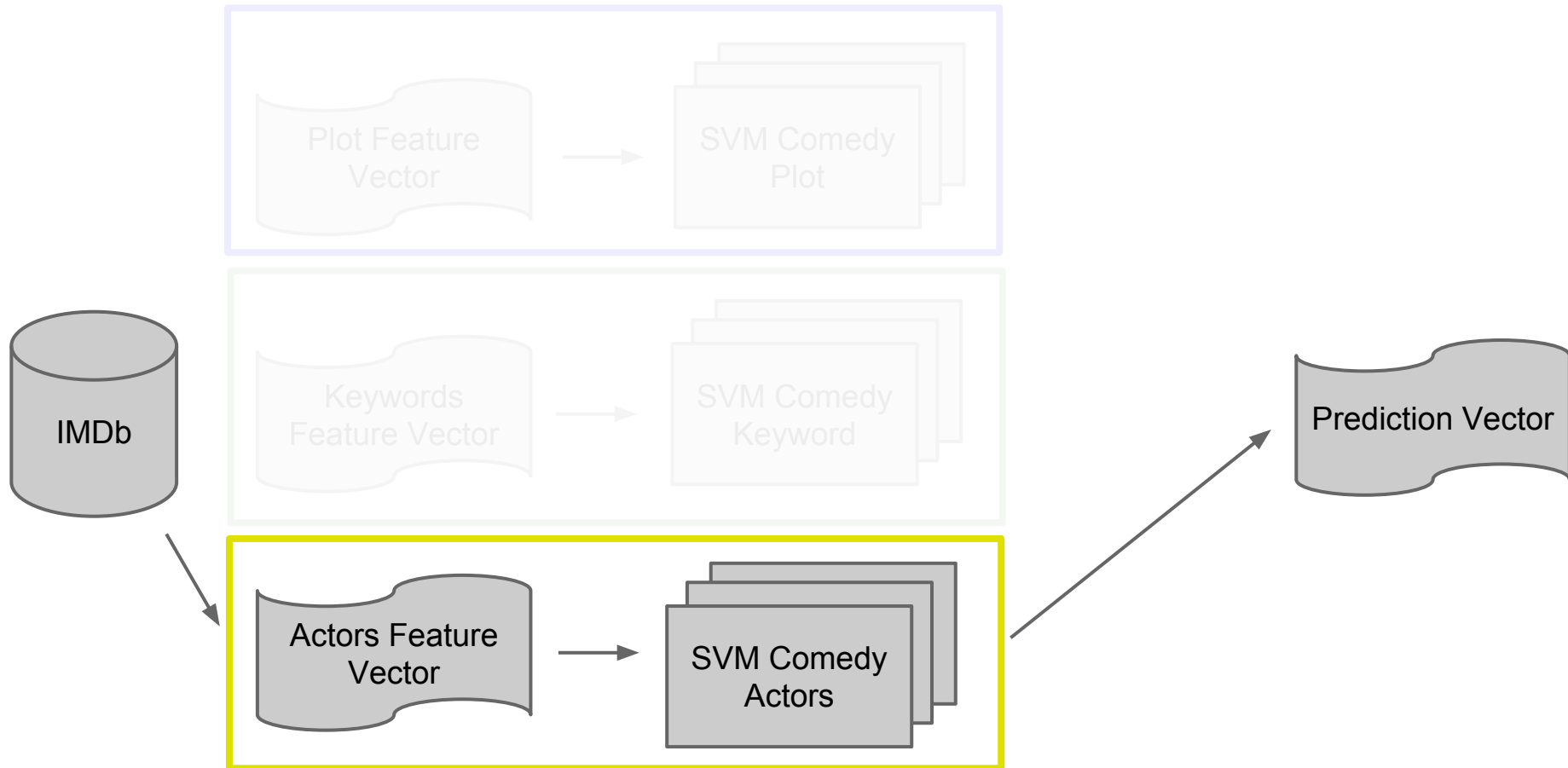
Experiments



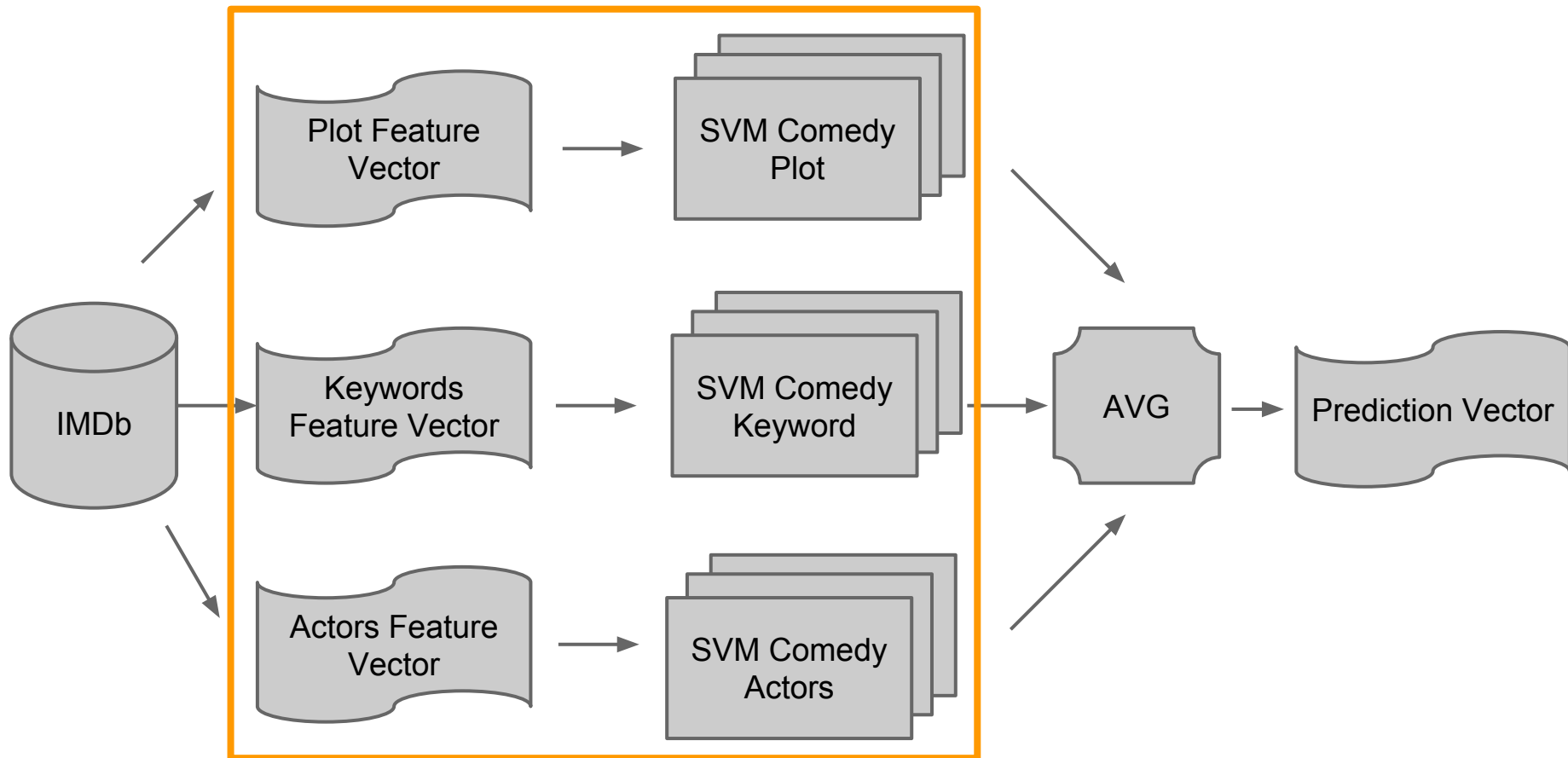
Experiments



Experiments

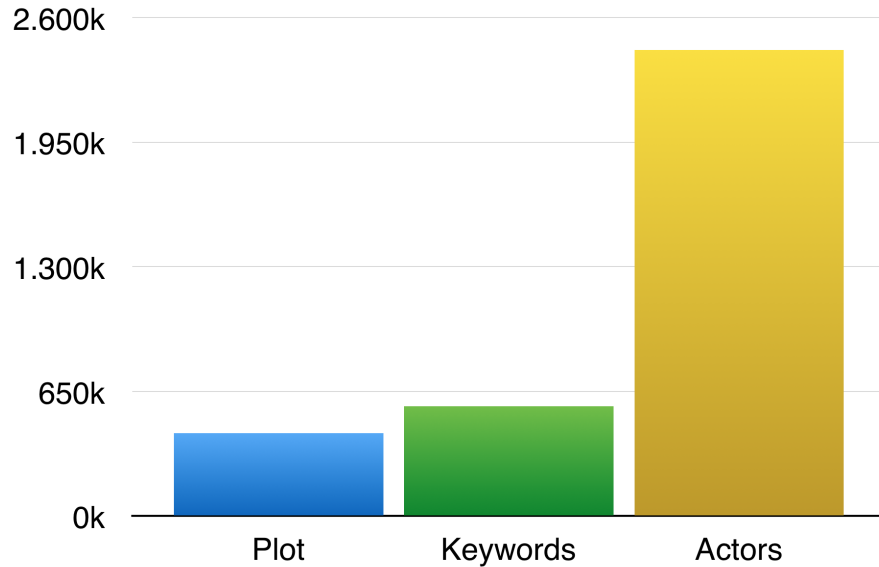


Experiments

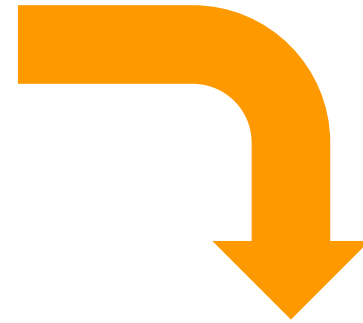


Experiments - Data

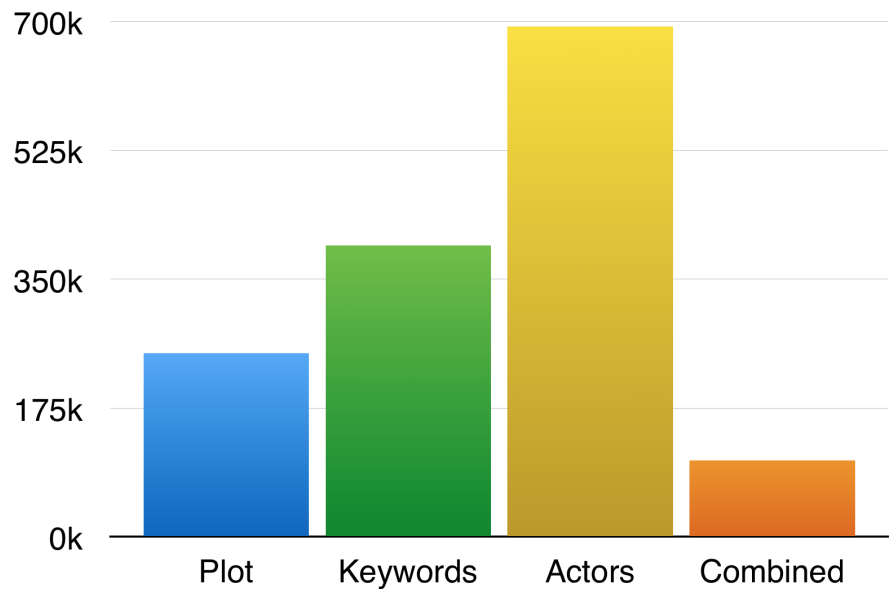
Movies with Features



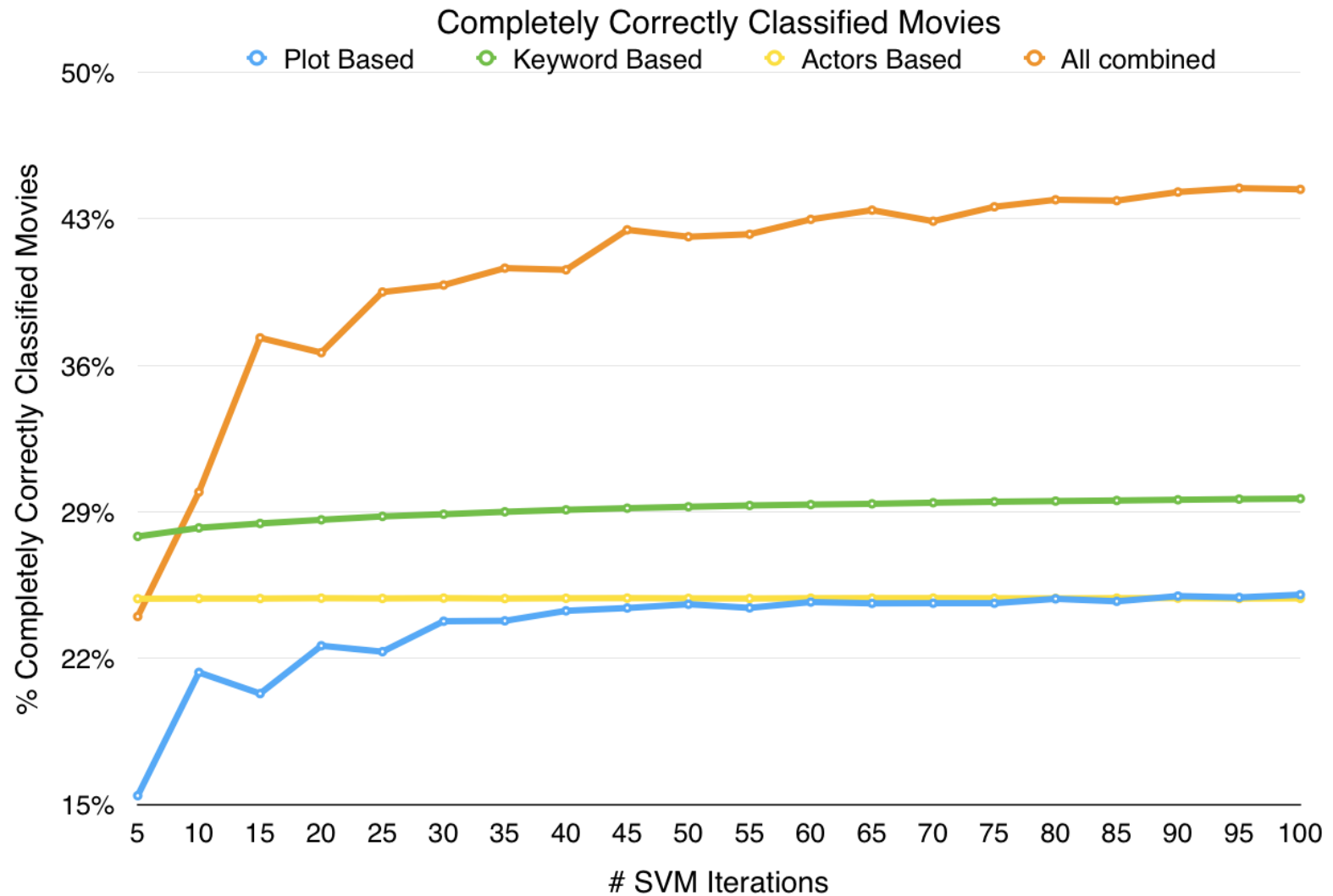
1033k with
genre data



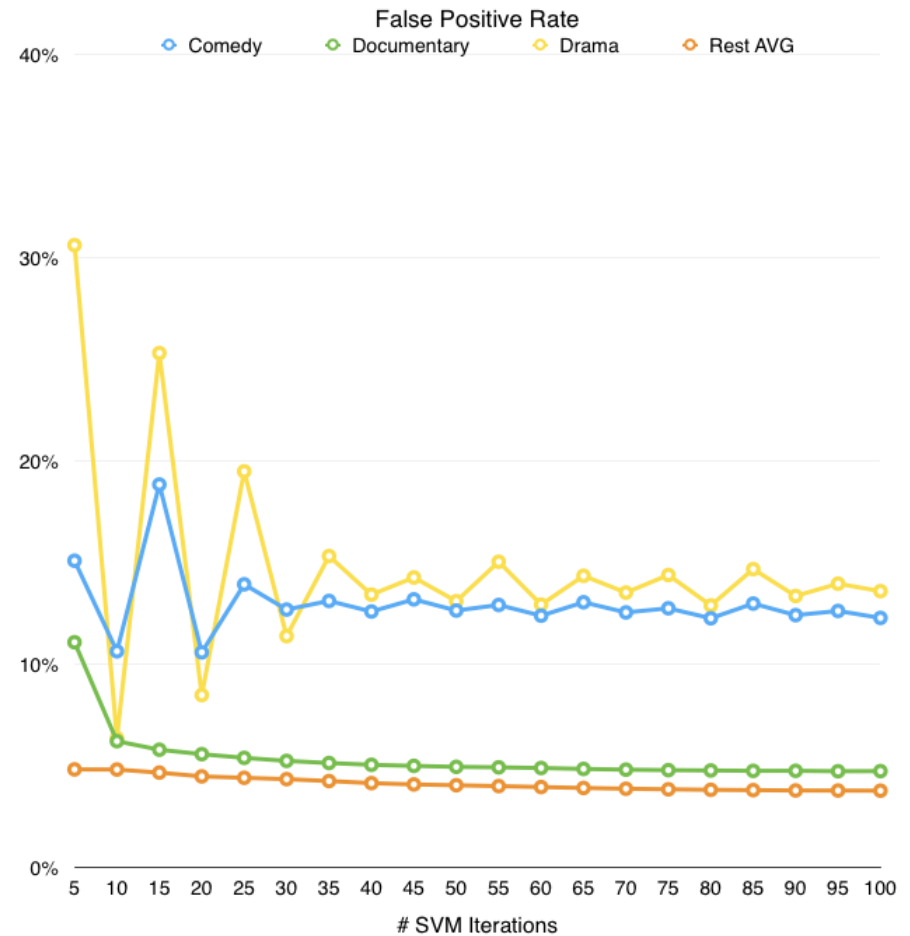
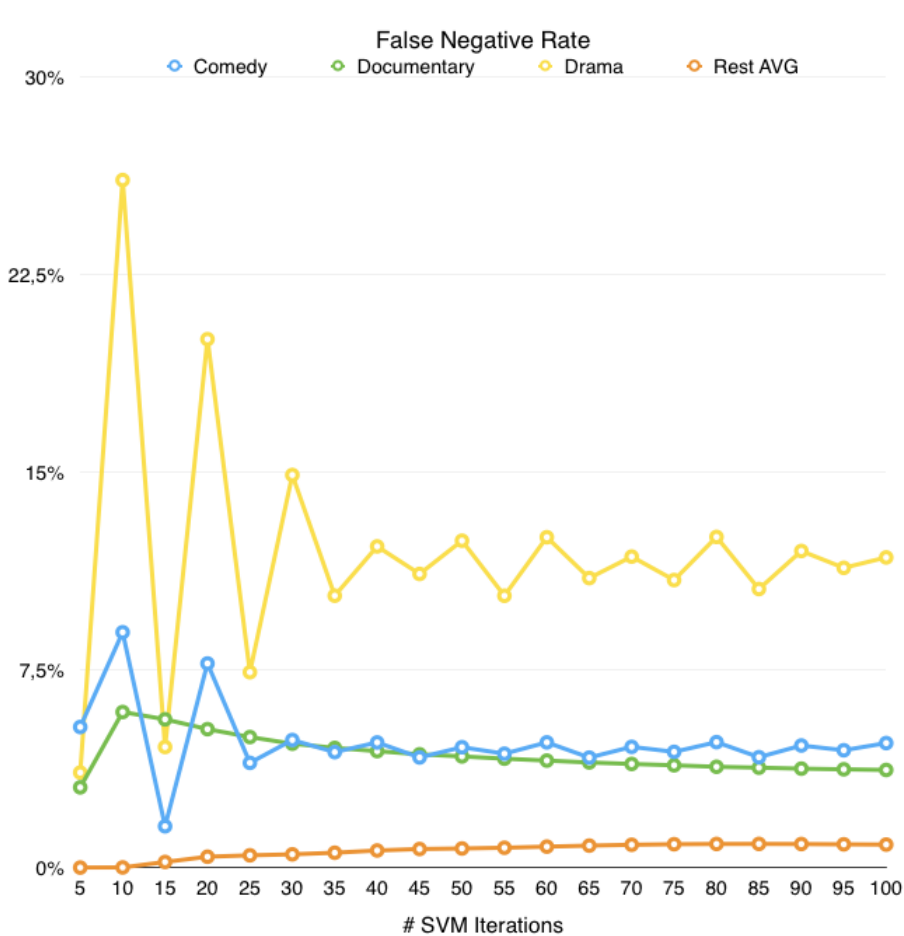
Movies with Feature & Genre Classification



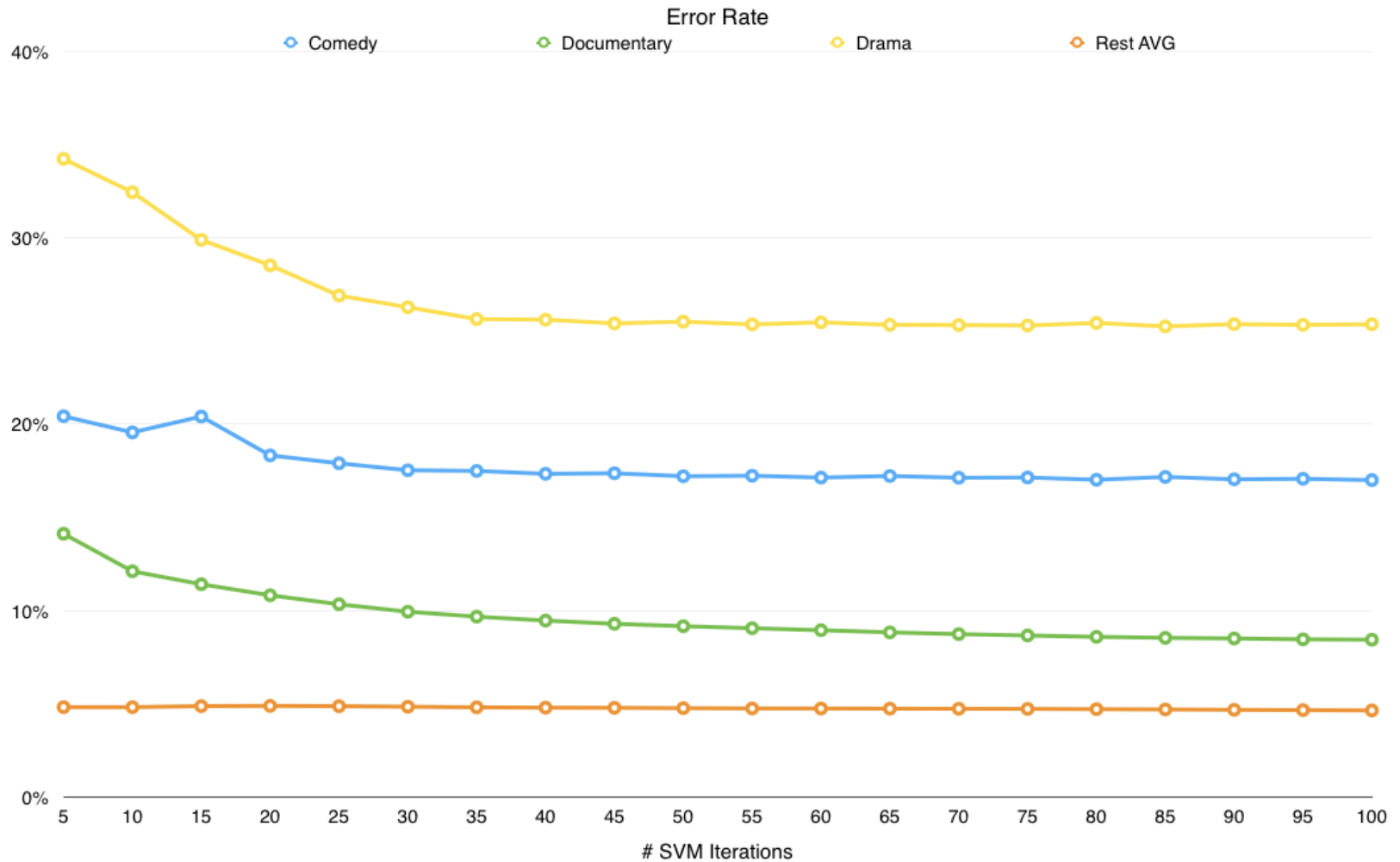
Results



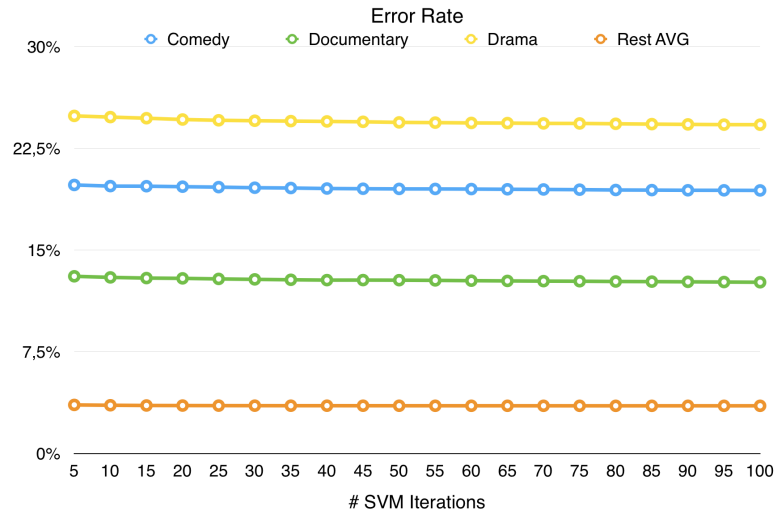
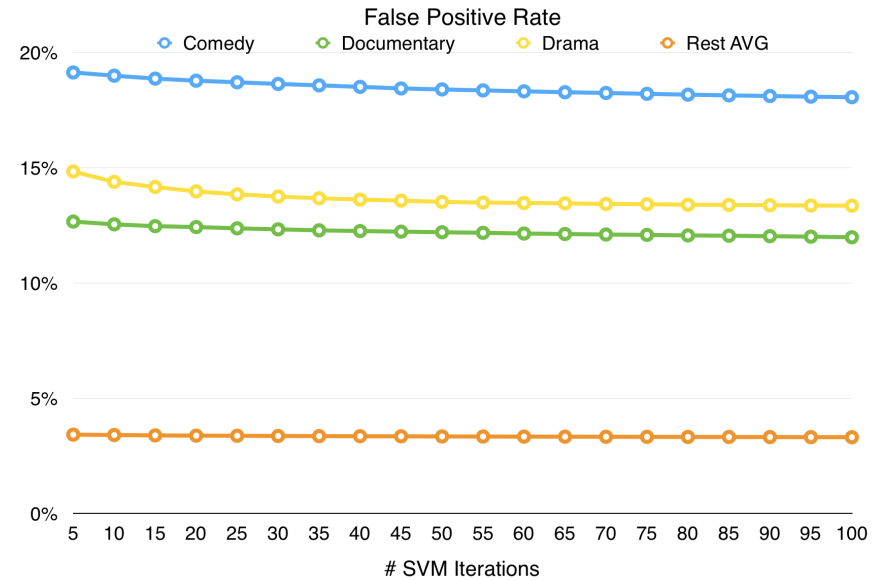
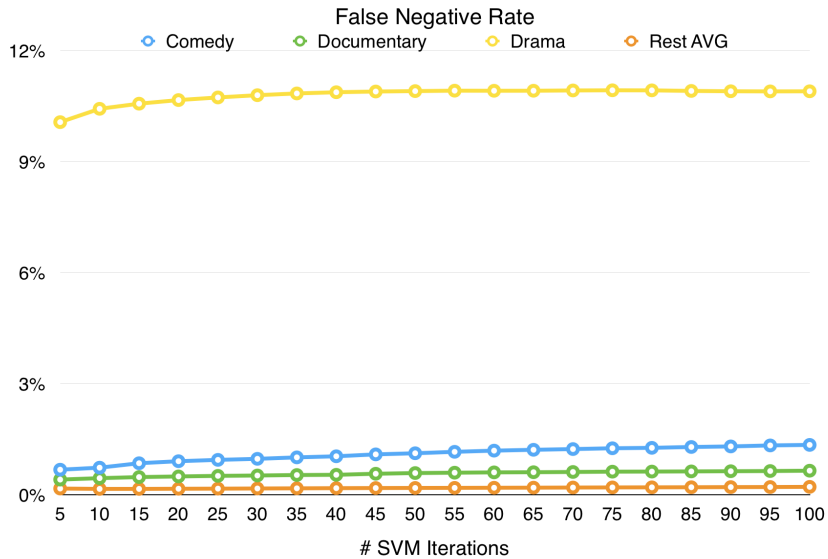
Results - Plot Based



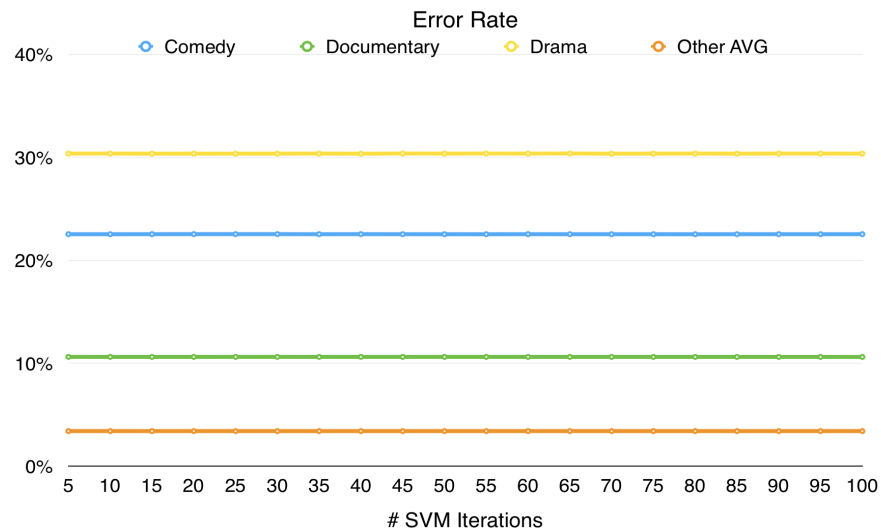
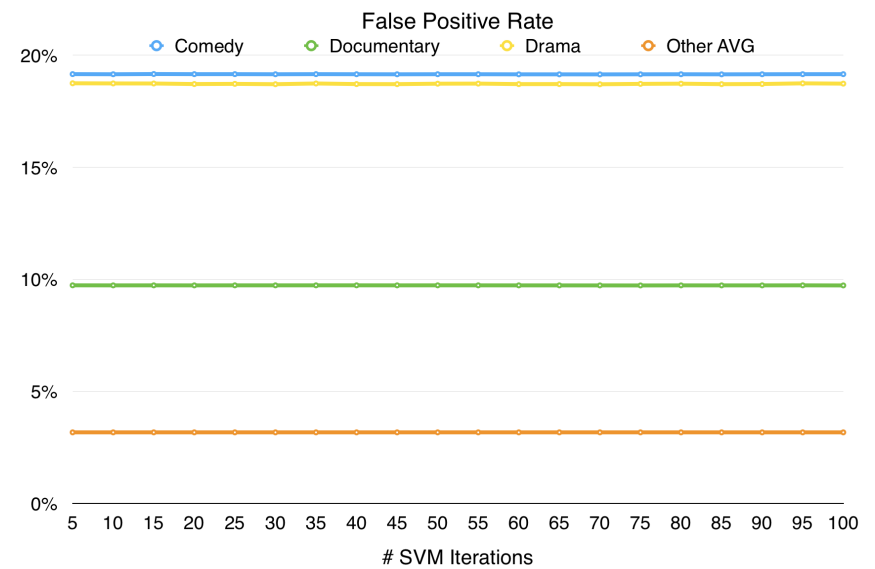
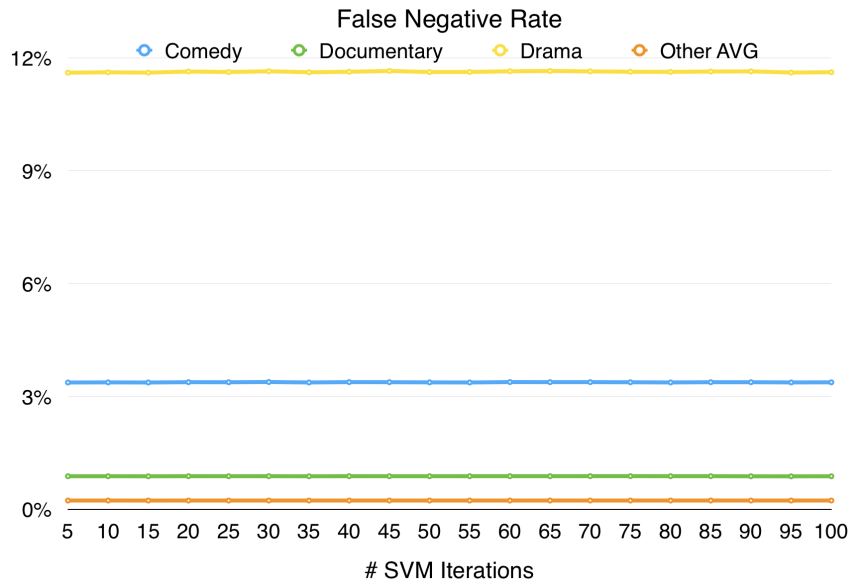
Results - Plot Based



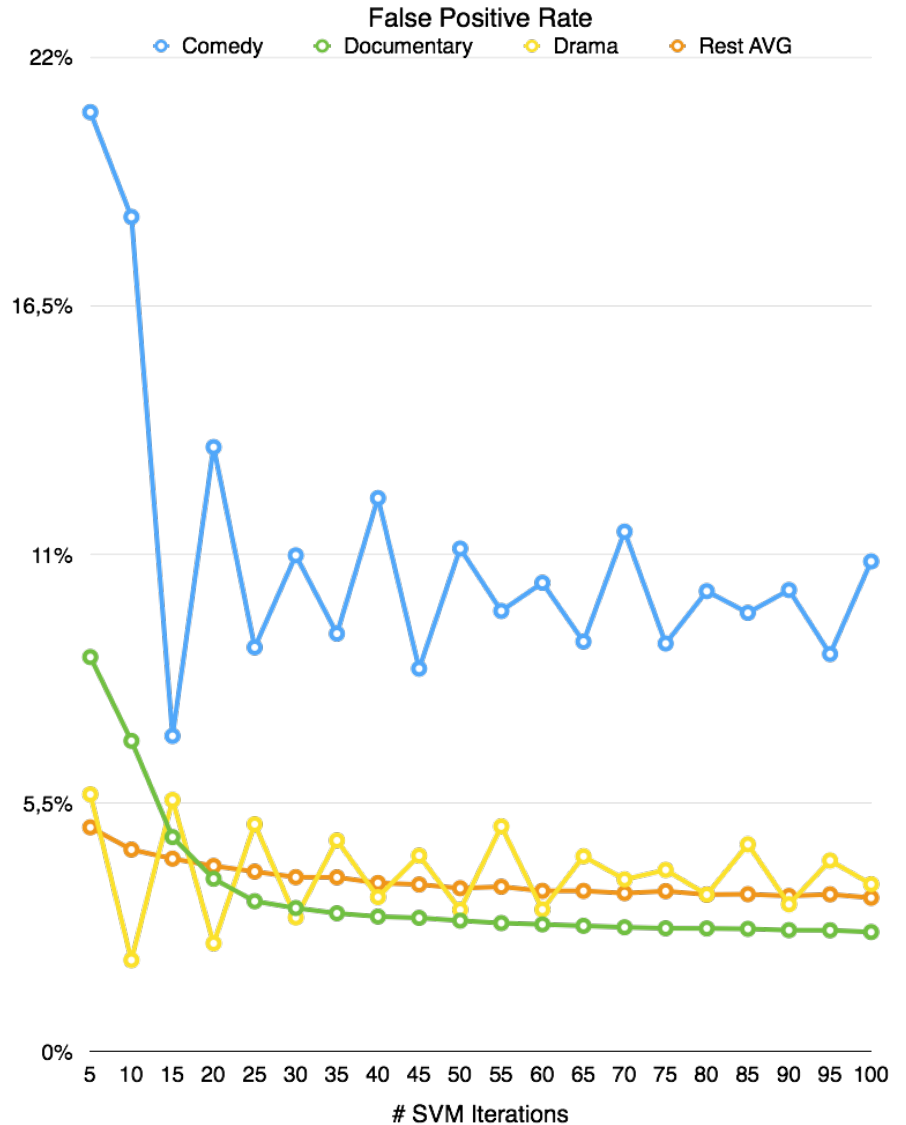
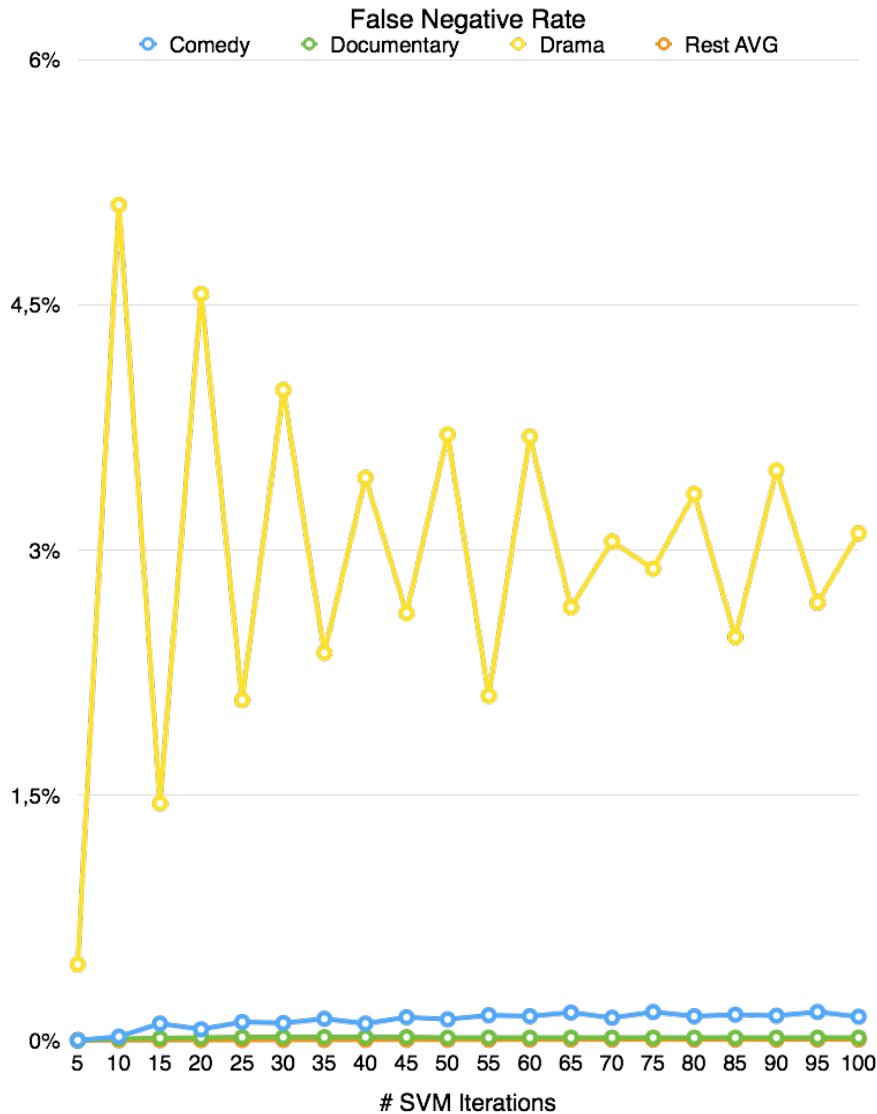
Results - Keyword Based



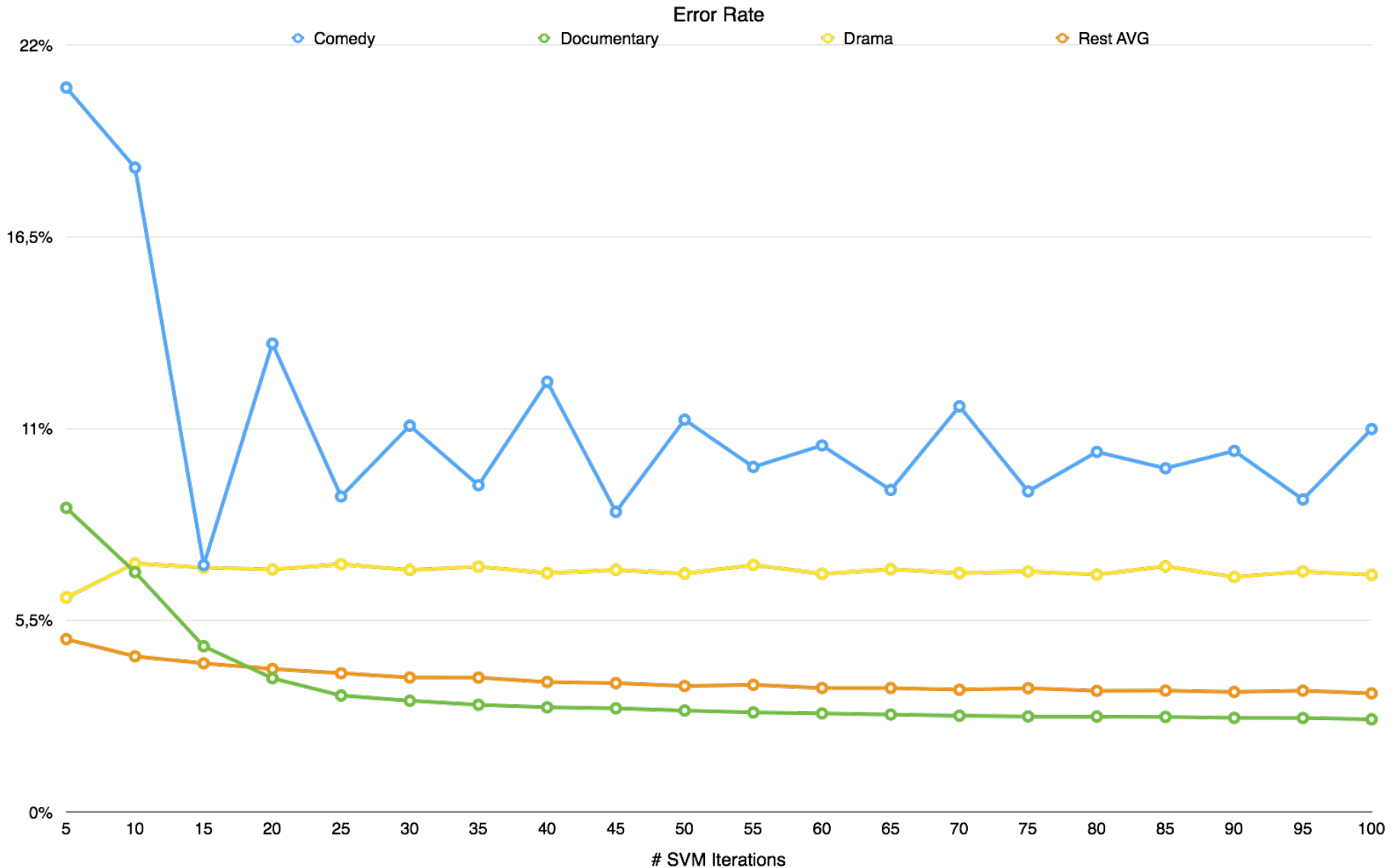
Results - Actors Based



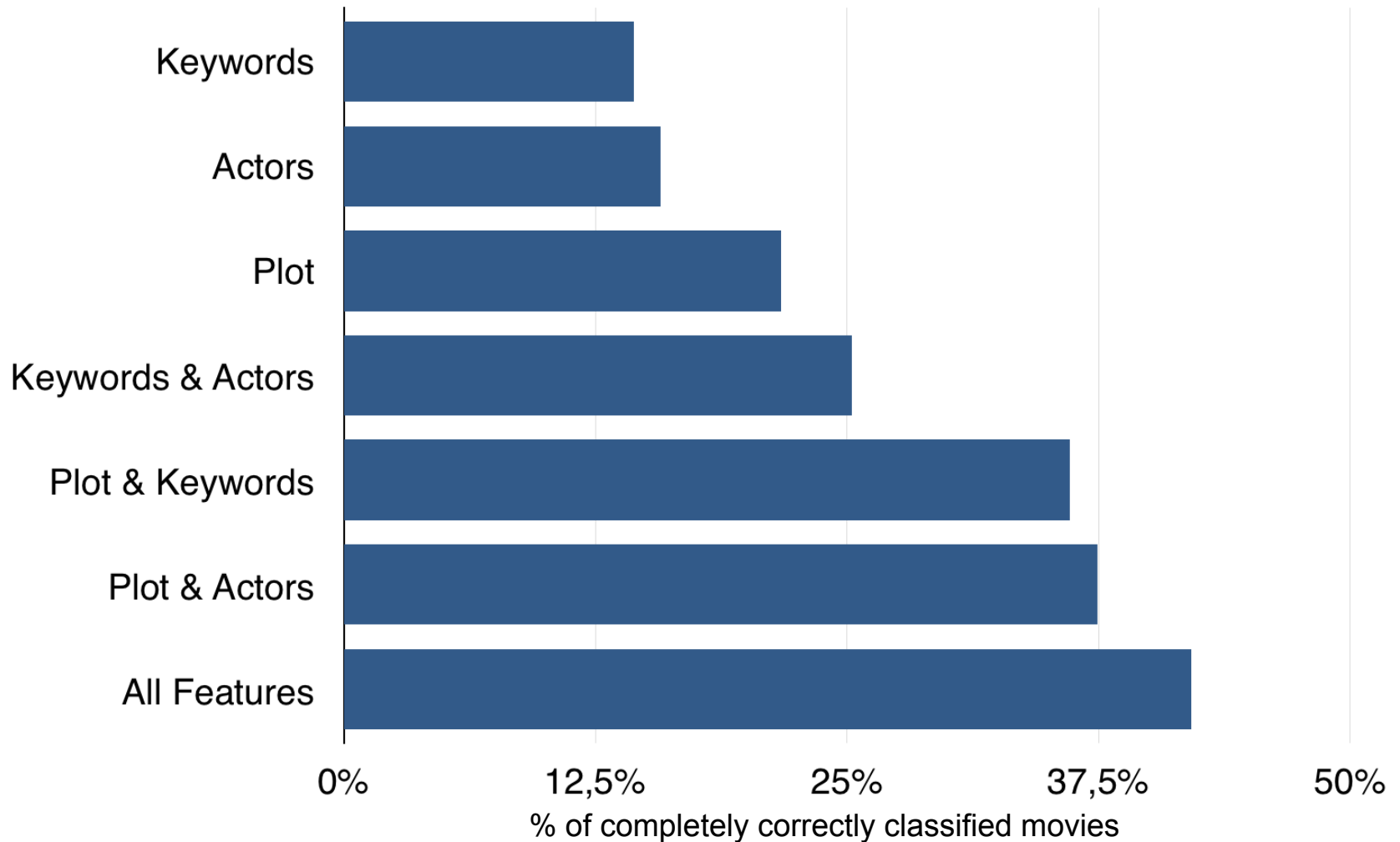
Results - Combined Features



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”