A hurricane, a star, and an insect walk into a bar

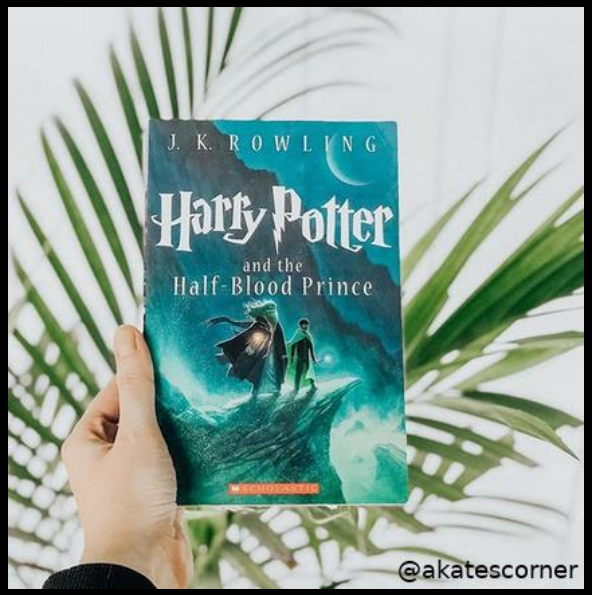
Using Goodreads Reviews to Predict Science Fiction Award Winners

Luke Miller
Jon Kaplan
John Fields

IST718
Dr. Jon Fox
June 11, 2020

"A hurricane, a star, and an insect walk into a bar..."

Popular culture is strongly influenced by so-called 'speculative fiction.' A particularly influential set of books are those that win literary awards.



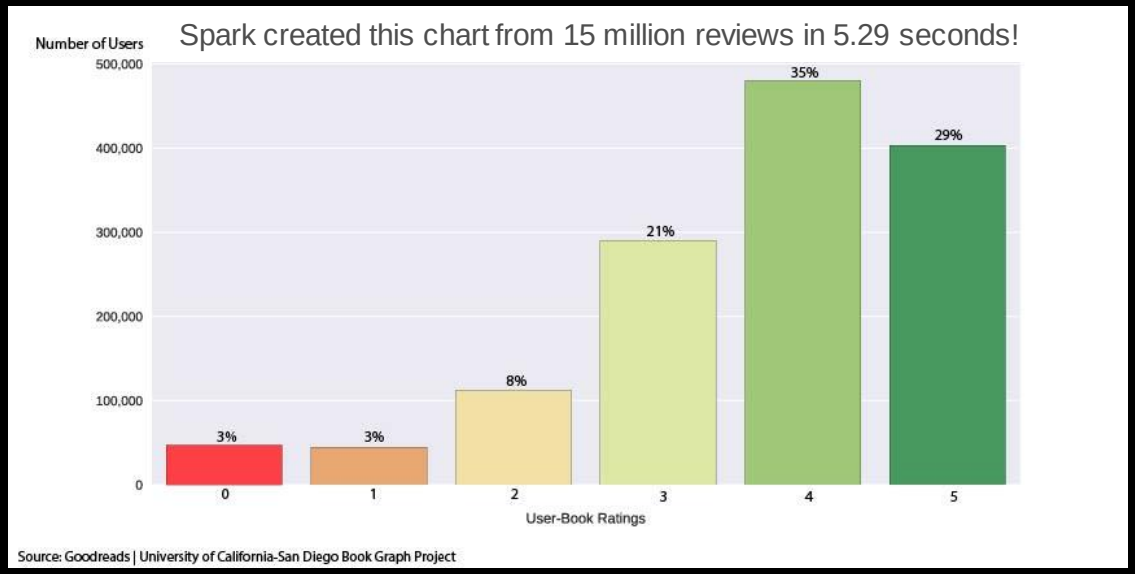
- The three big awards for Fantasy and Science fiction are the Hugo, the Nebula, and the Locus awards.
- We want to try to:
  1. **Predict award nominees**
  2. **Predict award winners**

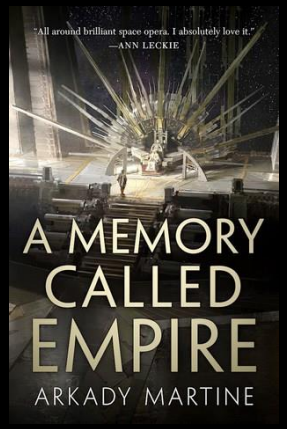
Really big data...

- University of California - San Diego's (UCSD) Book Graph project
  - **This includes Goodreads data from:**

    - 2.4 Million books
    - 830 K authors
    - 15 Million reviews
    - 2001 - 2017
- Also scraped Goodreads for 3600 book records for past winners/nominees and 2020 nominees

Colab/Spark – fast but not friendly to merges/sub-setting





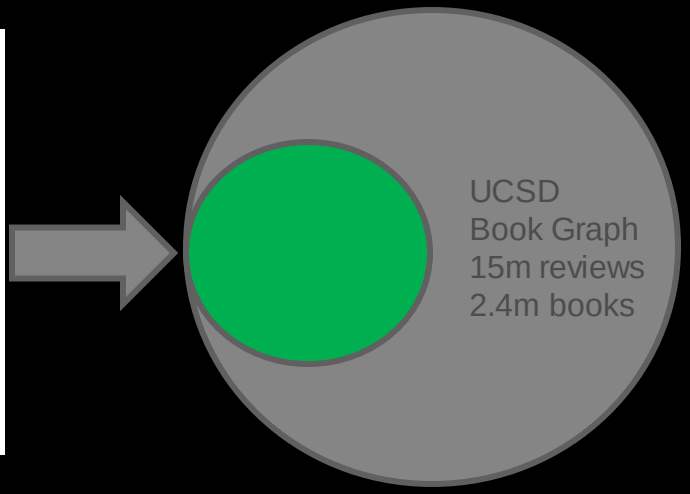
science-fiction	761 people
sci-fi	707 people
fiction	266 people
2019	238 people
scifi	183 people
space-opera	159 people
fantasy	158 people

7 of 4102 "shelves"

Your session crashed after using all available RAM. View runtime logs ×

UCSD Book Graph Fantasy/Paranormal Subset

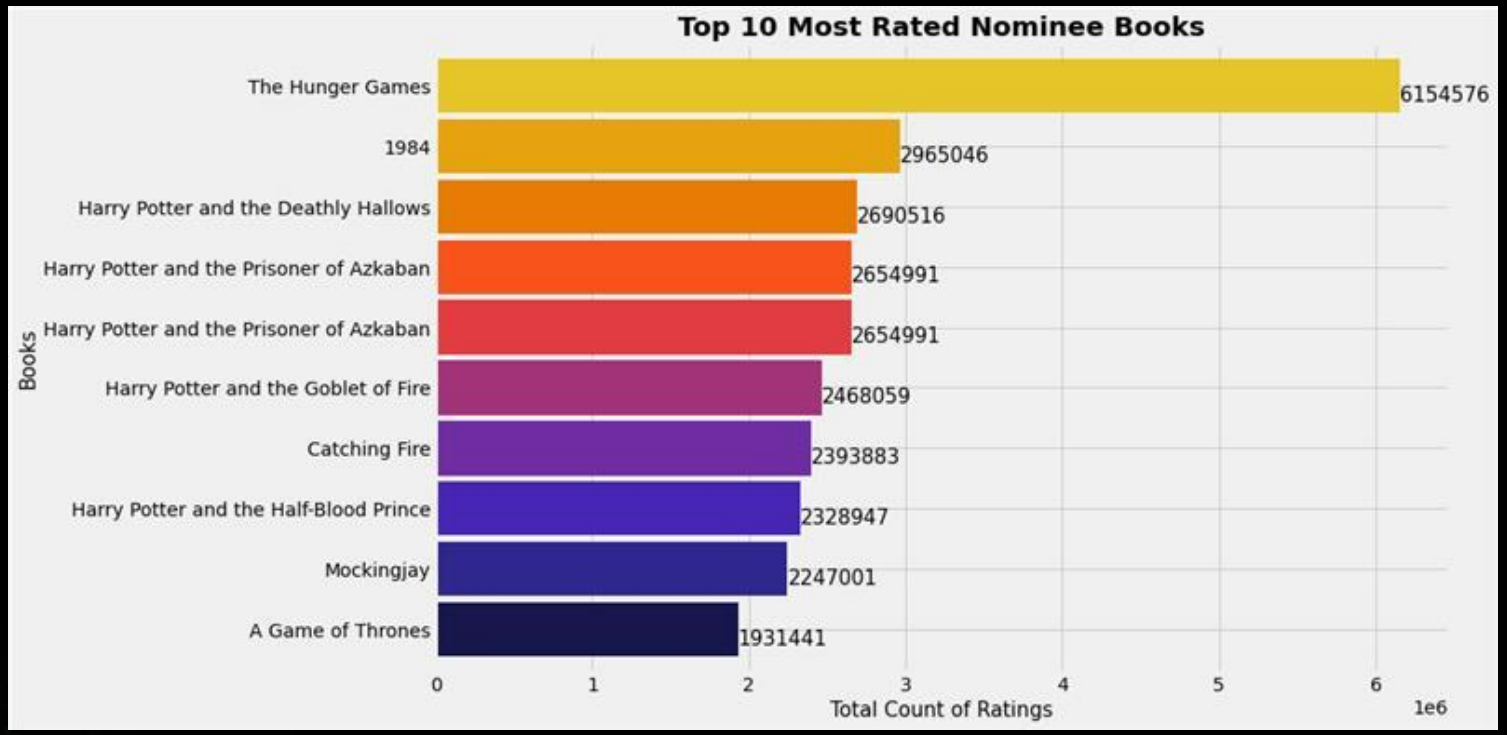
- 3.4m reviews
- 259k books
- 35% of science fiction book awards



John

Observe

Nominated Books



Observe

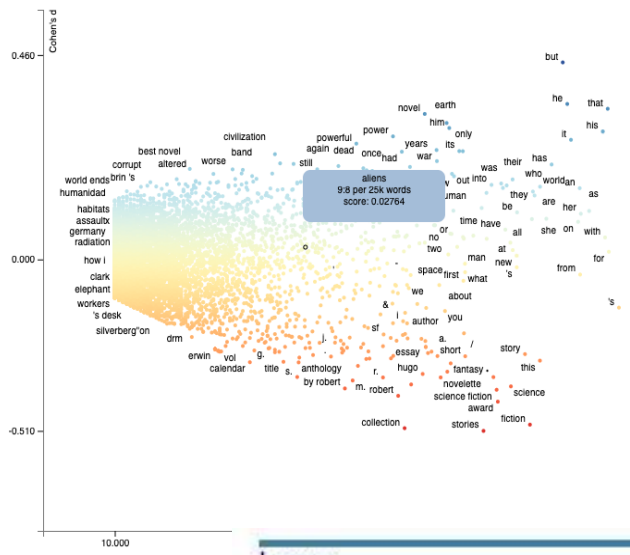
Nominated Books

Hugo Awards Nominees

Book Descriptions Word Cloud

Observe

Nominated Books



Search the chart

data

data

Locus
Life is cheap, and **data** is cheaper.

And a hunted **data**-vampire has followed Boris to where she is forbidden to return.

Locus
When multiple terrorist acts bring the world to its knees, the future of mankind rests in the hands of a few, mostly overweight, **data** center geeks.

Two **aliens** have wandered Earth for centuries.

Nebula
Now, long years later, a second alien race has been discovered, but again the **aliens'** ways are strange and frightening...

Locus
Now, long years later, a second alien race has been discovered, but again the **aliens'** ways are strange and frightening...

Locus
Aliens have landed in New York.

Now that plague is heading for Earth, and threatens humans and **aliens** alike.

Nebula
government will sell treasury debt to SPACE **ALIENS**!!!

https://jonkaplan18.github.io/jonkaplan18/index.html

Jon

Observe

Demographics

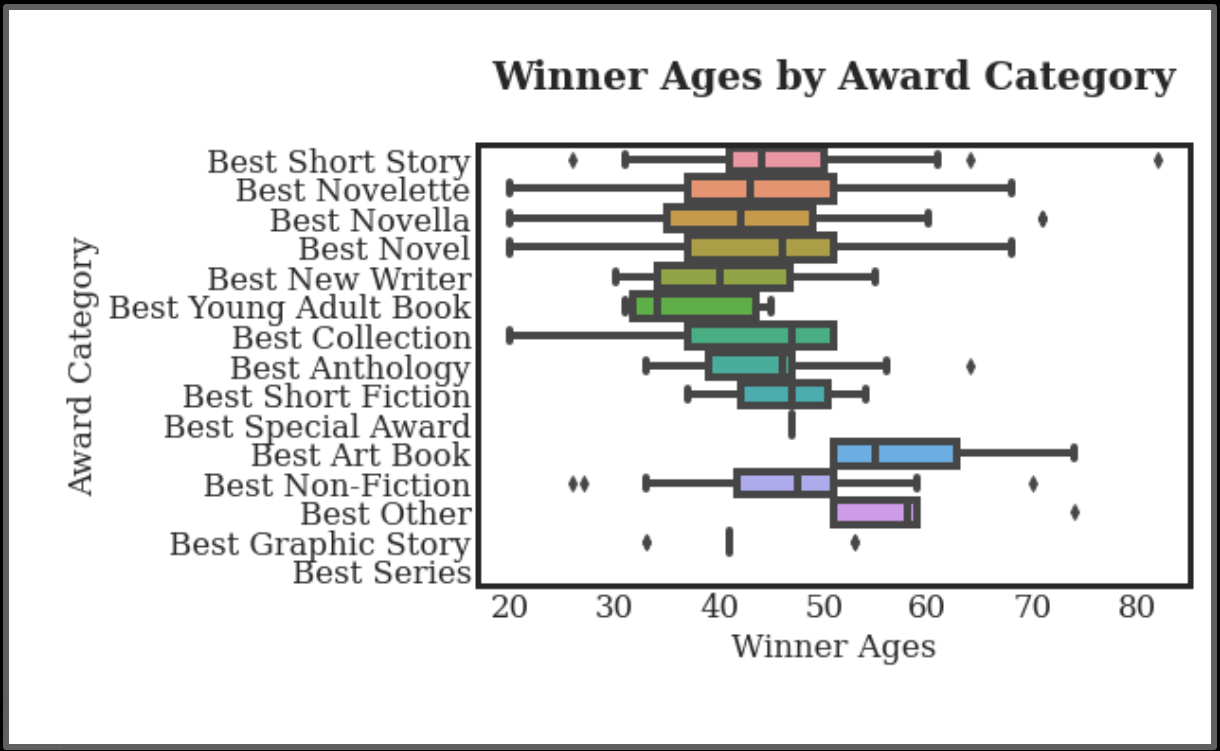
Occupations:

- Writer
- Librarian
- Anthropologist
- Publisher
- Psychologist
- Linguist
- Hypnotist
- Klingon language expert
- Lawyer
- Aerospace engineer
- Chemical engineer
- Futurist
- Truck Driver
- Marine biologist
- Psychic medium
- Occultist

Observe

Demographics

Average Age:
44 years old

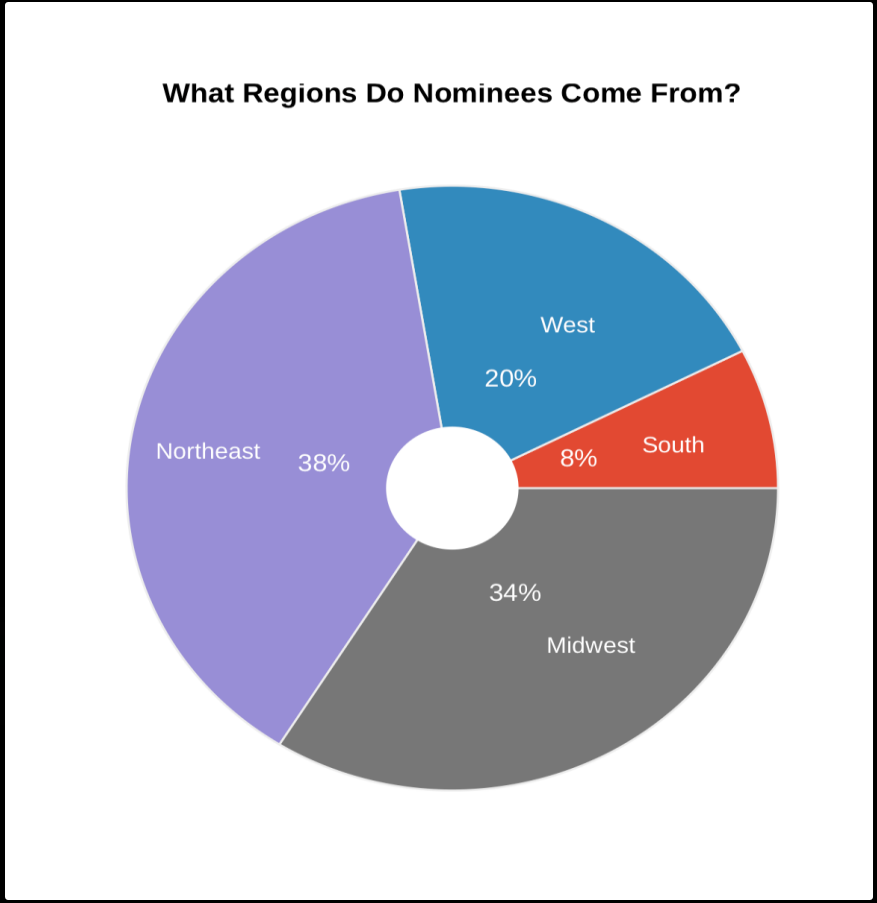


Jon

Observe

Demographics

Authors by USA Region:
Mostly Northeast & Midwest

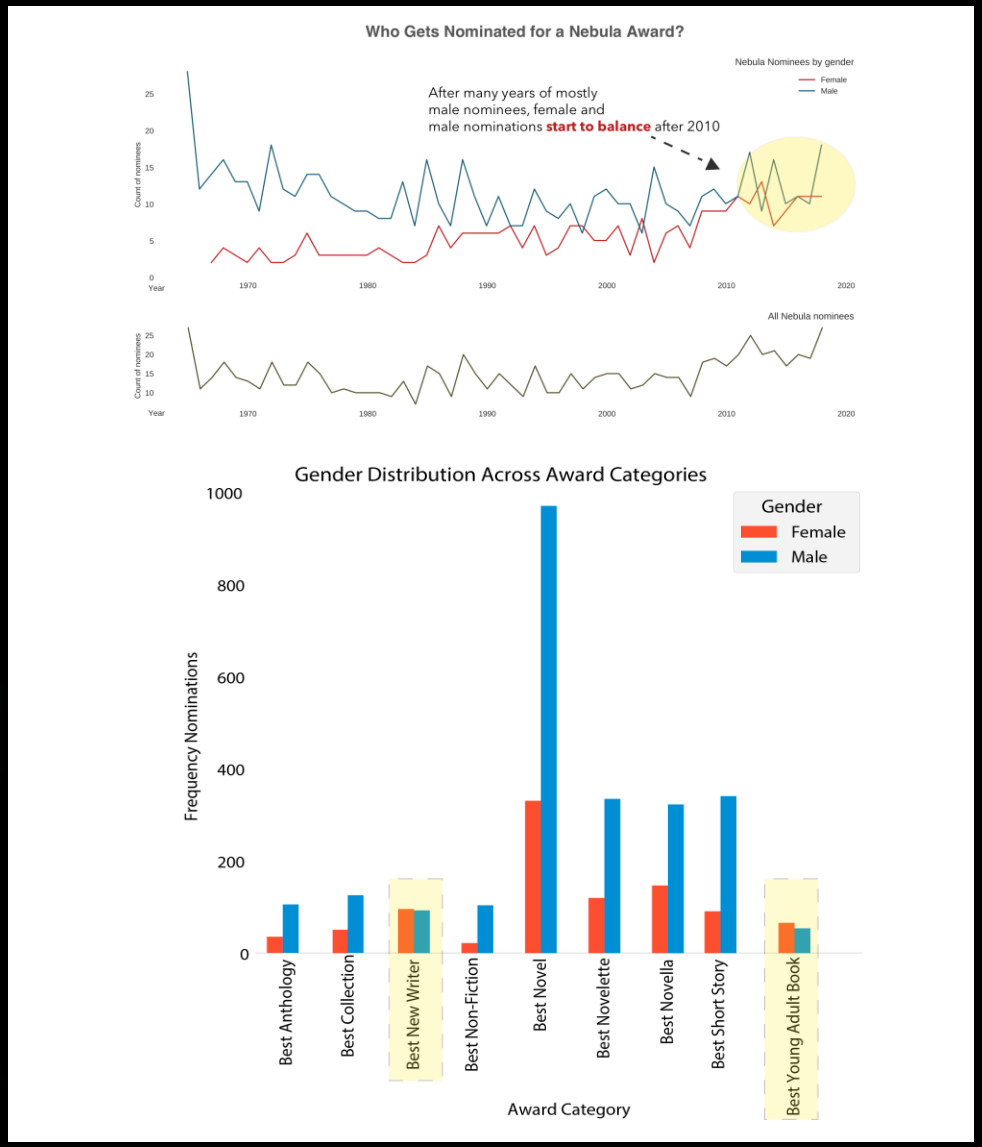


Jon

Observe

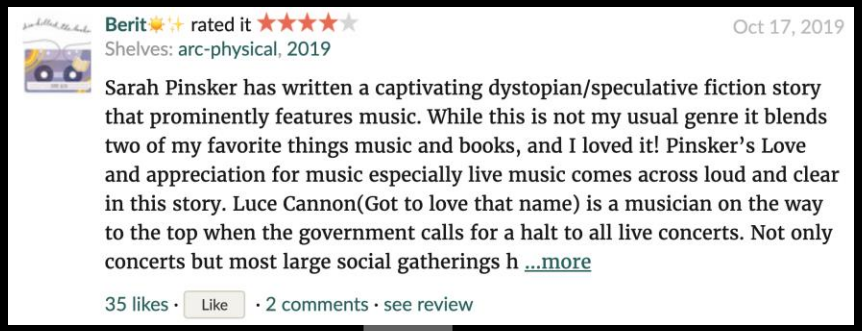
Demographics

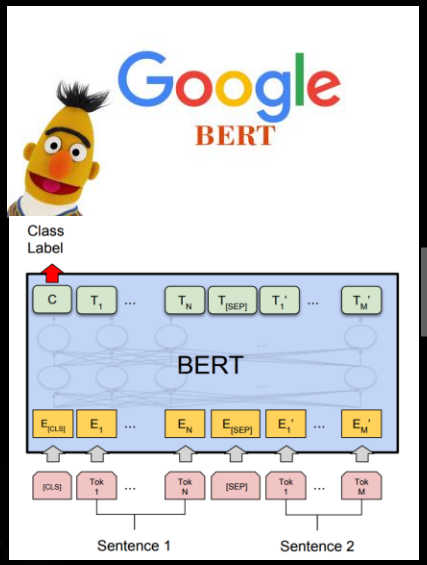
Gender:
Male-heavy, but changing

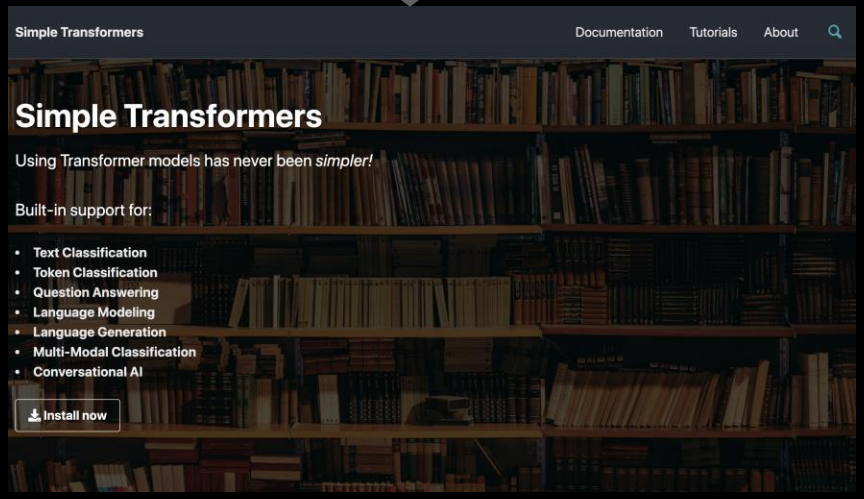


Jon

Model – picking nominees with BERT







75% ACCURACY

John

Model – picking nominees with BERT + sklearn

	model	params	fit_time_avg	score_time_avg	f1_avg	accuracy_avg	cv_elapsed
0	RandomForestClassifier	{"max_depth": 2}	0.175958	0.025919	0.491692	0.967309	0:00:00.410419 elapsed
1	RandomForestClassifier	{"max_depth": 10}	0.263998	0.047518	0.539130	0.966471	0:00:00.633970 elapsed
2	RandomForestClassifier	{"max_depth": 100}	0.281441	0.051397	0.547964	0.961861	0:00:00.677039 elapsed
3	SVC	{"kernel": "linear", "C": 1}	165.117654	0.013786	0.496810	0.967477	0:05:30.274687 elapsed
4	SVC	{"kernel": "rbf", "C": 1}	0.109521	0.055421	0.496801	0.967393	0:00:00.339447 elapsed
5	SVC	{"kernel": "rbf", "C": 10}	0.263779	0.067080	0.523079	0.966555	0:00:00.672040 elapsed
6	SVC	{"kernel": "poly", "C": 1}	11.360120	0.017424	0.518310	0.967896	0:00:22.766108 elapsed
7	GaussianNB	{"priors": [0.03, 0.97]}	0.005442	0.005059	0.578774	0.911484	0:00:00.027622 elapsed

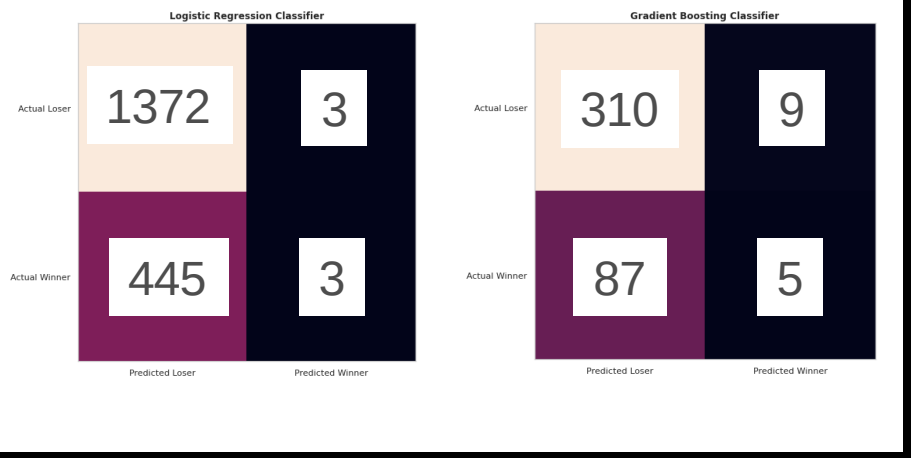
92-97% ACCURACY

Luke

Model – Picking Winners

Imbalanced: 75% at picking non-winners

```
0    2268
1     762
```

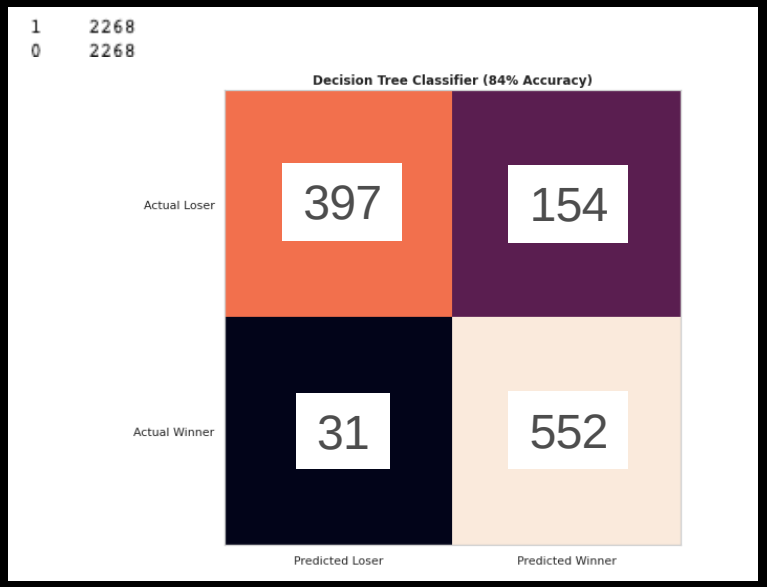


5 true positives

```
Accuracy of Logistic regression classifier on training set: 0.74
Accuracy of Logistic regression classifier on test set: 0.76
Accuracy of Decision Tree classifier on training set: 0.92
Accuracy of Decision Tree classifier on test set: 0.68
Accuracy of K-NN classifier on training set: 0.78
Accuracy of K-NN classifier on test set: 0.69
Accuracy of LDA classifier on training set: 0.74
Accuracy of LDA classifier on test set: 0.76
Accuracy of GNB classifier on training set: 0.74
Accuracy of GNB classifier on test set: 0.76
Accuracy of SVM classifier on training set: 0.74
Accuracy of SVM classifier on test set: 0.76
```

Balanced: 84% at picking winners

```
1    2268
0    2268
```

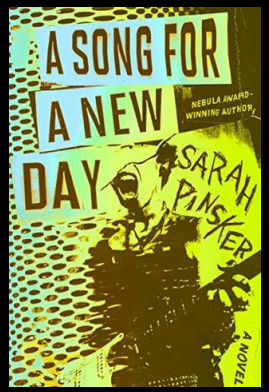


552 true positives

```
Accuracy of Logistic regression classifier on training set: 0.56
Accuracy of Logistic regression classifier on test set: 0.52
Accuracy of Decision Tree classifier on training set: 1.00
Accuracy of Decision Tree classifier on test set: 0.84
Accuracy of K-NN classifier on training set: 0.80
Accuracy of K-NN classifier on test set: 0.70
Accuracy of LDA classifier on training set: 0.61
Accuracy of LDA classifier on test set: 0.58
Accuracy of GNB classifier on training set: 0.56
Accuracy of GNB classifier on test set: 0.55
Accuracy of SVM classifier on training set: 0.55
Accuracy of SVM classifier on test set: 0.53
```

Recommend – if we just had more compute...

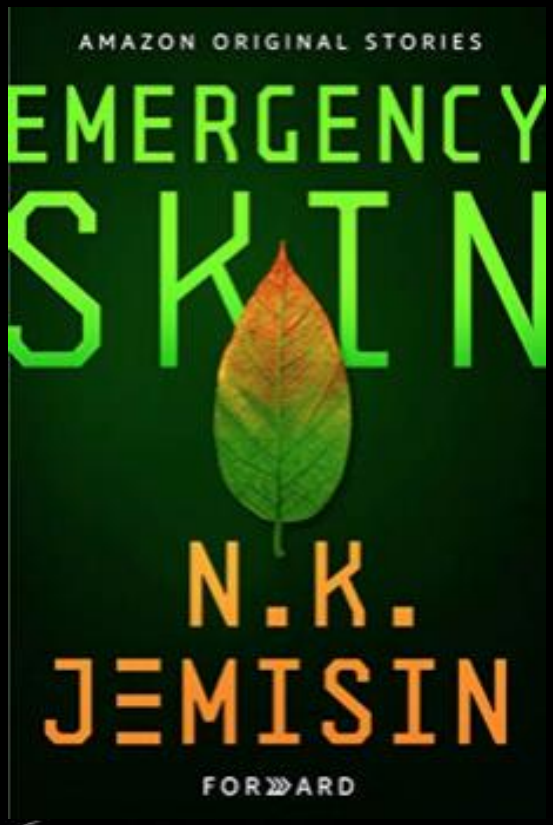
SciFi Book Awards	Accuracy
Predict Nominees (2007-2017)	75% (BERT) 92-97% (BERT + sklearn)
Predict Winners (2020)	84% (sklearn balanced)

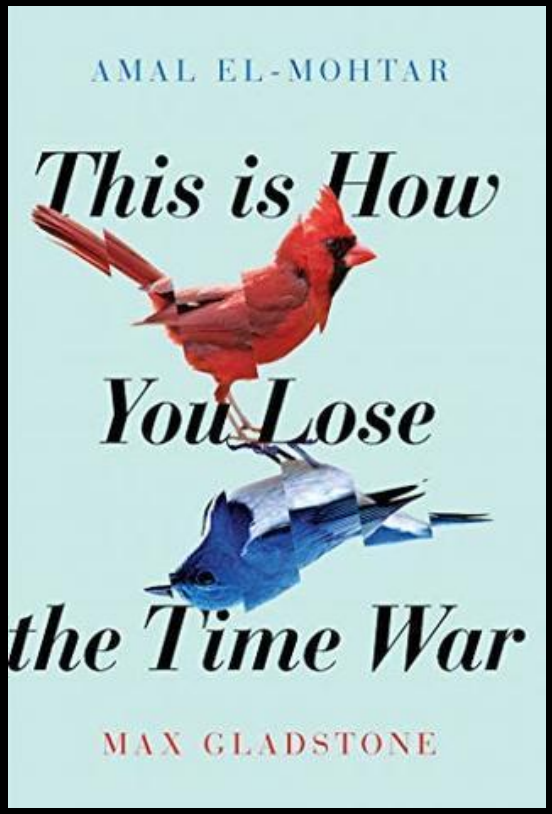


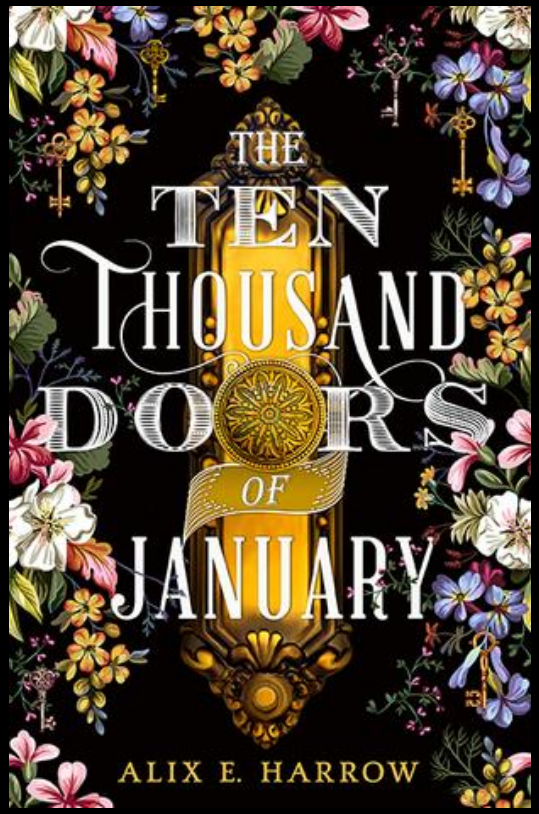
Challenges	
Spark syntax	Great for fast queries but very hard to subset and merge
Colab resources	GPU support ok but limited RAM
BERT/simple transformers	Good results on text classification but requires large compute (GPU's)

Luke

Hugo 2020 Predicted Winners







Check our predictions
August 1, 2020

Jon

Resources and Data

- https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home
- Devlin, J., Ming-Wei, C., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Ithaca: Cornell University Library, arXiv.org. Retrieved from https://search-proquest-com.libezproxy2.syr.edu/docview/2118630252?accountid=14214
- https://gluon-nlp.mxnet.io/examples/sentence_embedding/bert.html
- https://github.com/ThilinaRajapakse/simpletransformers
- https://github.com/huggingface/transformers
- https://simpletransformers.ai/[8] https://medium.com/swlh/simple-transformers-multi-class-text-classification-with-bert-roberta-xlnet-xlm-and-8b585000ce3a

- Nominees Dataframe
- **nom_awards** = pd.read_csv('https://www.dropbox.com/s/1mll9m9r27wy5iz/nominees_stats_awards_all.csv?dl=1') #this is the original GoodReads scraped dataset without additional demographic info.
- nom_awards2 = pd.read_csv(https://www.dropbox.com/s/ysuekksts86beag/final_nominations_df.csv?dl=1'') #this is the version with the merged demographic info from Worlds Without End scrape