SciFi Book Plot Evaluation

Jaime Jimenez, Mary Mays, Lisa Stroh

“Each day means a new twenty-four hours. Each day means everything's possible again. You live in the moment, you die in the moment, you take it all one day at a time.”

—Marie Lu, *Legend*

Background

- After months of stressful participation in the KU Data Analytics Bootcamp, we wanted to escape into a good scifi book. If we were ambitious enough, we might even write a scifi book.
- Goal: Make predictions on scifi books based on the plot description from the Kaggle data set
- We have developed a web page which uses machine learning to predict the success of a book and the subgenre based on the plot description.

Mikołaj on Unsplash

Dataframe cleaned up

The following table includes five percent of the data at random used for training our machine learning models during this project

	Book Title	Rating Score	Rating Votes	Book Description	Subgenre	Rating
235	The Degan Incident	3.96	1949	lonely spaceport worker mcsmith meets exotic beastly alien planet dega begins adventure takes stars beyond working earths spaceport mcsmith meets exotic beastly alien planet dega get know relationship quickly heats begin fall disappears without devins encounter triggers startling change body taken captive held top research facility find late	sf_aliens	1
5192	We	3.91	74652	exhilarating dystopian inspired orwells foreshadowed worst excesses soviet russiayevgeny zamyatins powerfully inventive influenced writers orwell glass enclosed city absolute straight lines ruled powerful benefactor citizens totalitarian society onestate live lives devoid creativity mathematician dreams numbers makes discovery individual twenty sixth century ad classic dystopian forerunner works orwells huxleys new world suppressed many years remains resounding cry individual yet also powerful exciting vivid work science fiction browns translation based corrected text first published sixty years suppression	sf_dystopia	1
905	The Hunter's Mate	4.21	580	still haunted nightmares hunters sacrifice save lives friends last alien expects akrellia menops bounty shes drawn dangerous despite insectoid appearance chalks experiences together unable accept might mean moreeven decides never let sight spent searching thing thing onlya cure imprinting end joined traitorous iriduan halian devils bargain deceiving friends hes comfortable deceptionor growing concerns halians plans execution part halians plan isolates alien world far help hunters problems grow complicated unanticipated metamorphosis leaves body completely changedin way makes impossible ignore growing feelings secrets lies past cant change form seemingly insurmountable	sf_aliens	2

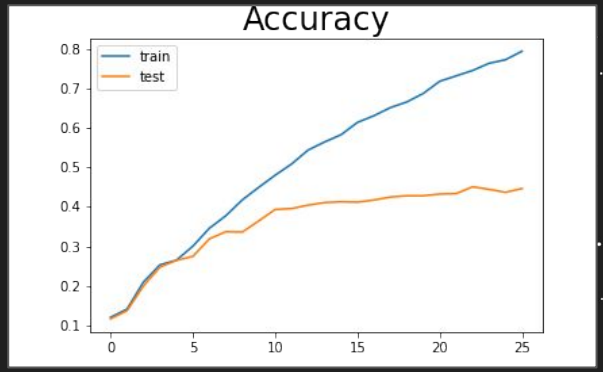
Data Sources

- Kaggle
- Name file for stopwords
- Machine Learning example code
- CSS example code

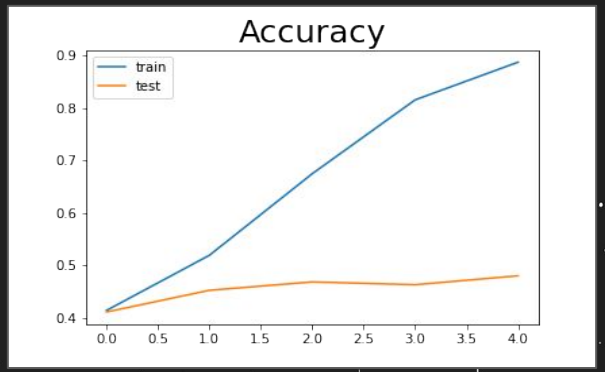
Machine Learning Deep Dive

- Created 3 models which took in description, rating, and subgenre of scifi books.
  - One model predicts the rating and another model predicts the subgenre.
  - The 3rd model used Naive Bayes in PySpark. We chose not to use it, since it was tokenized differently and did not work as well
- The 2 models used RNN Sequential model from TensorFlow
  - Used 4 layers
  - 30 epochs for the subgenre model
  - 5 epochs for the rating model
  - Overfit was a big problem for rating model

Subgenre accuracy



Rating accuracy



Machine Learning Deep Dive

In [44]:
```
model = Sequential()
model.add(Embedding(MAX_NB_WORDS, EMBEDDING_DIM, input_length=X.shape[1]))
model.add(SpatialDropout1D(0.2))
model.add(LSTM(100, dropout=0.2, recurrent_dropout=0.2))
model.add(Dense(12, activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
print(model.summary())
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 250, 100)          5000000
_________________________________________________________________
spatial_dropout1d (SpatialDr (None, 250, 100)          0
_________________________________________________________________
lstm (LSTM)                  (None, 100)               80400
_________________________________________________________________
dense (Dense)                (None, 12)                1212
=================================================================
Total params: 5,081,612
Trainable params: 5,081,612
Non-trainable params: 0
_________________________________________________________________
None
```

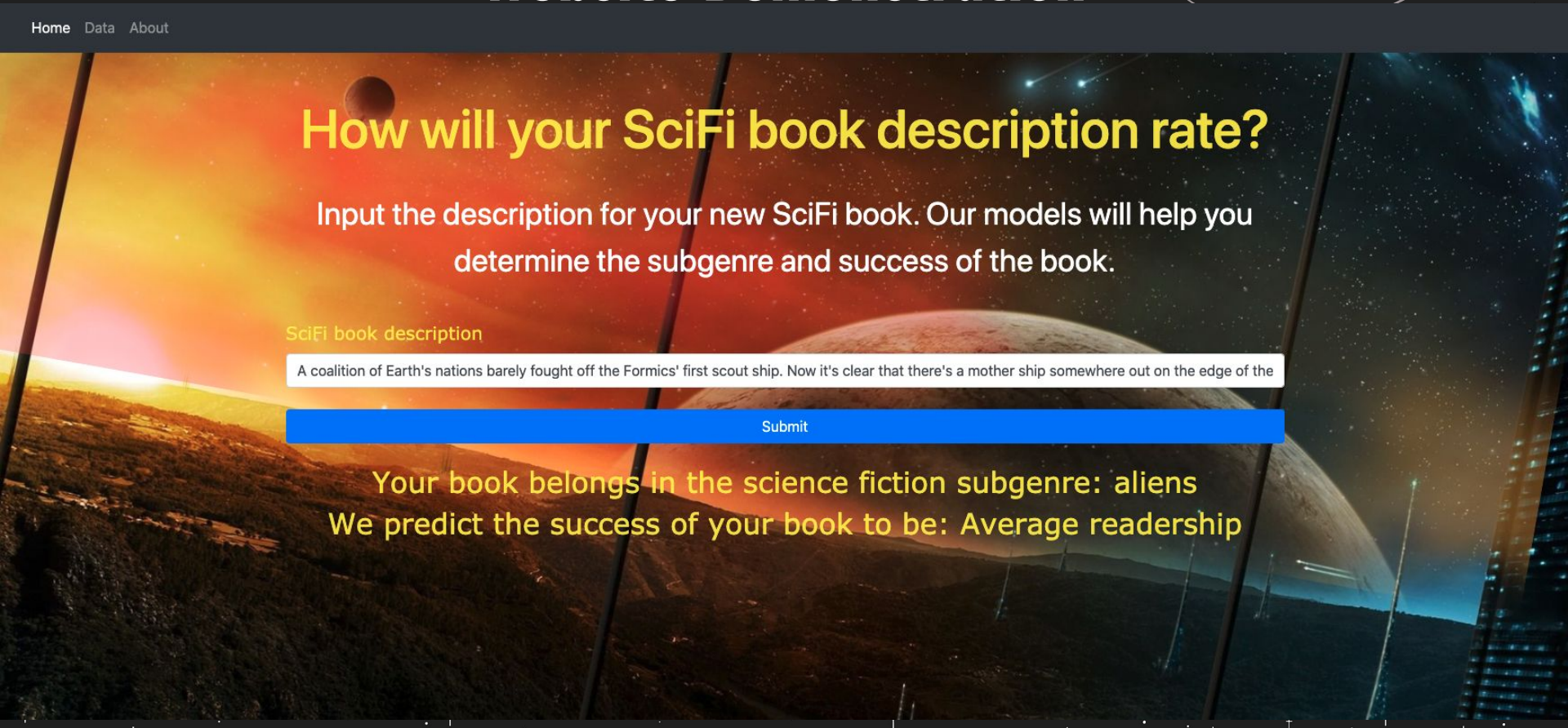
In [62]:
```
epochs = 8
batch_size = 64

history = model.fit(X_train, Y_train, epochs=epochs, batch_size=batch_size,validation_split=0.1,callbacks=[EarlyStoppir
```

Product Design

- Pulled and transformed csv files in jupyter notebook
- Cleaned, and tokenized book descriptions
- Ran machine learning models in jupyter notebook
- Exported ML models to h5 files and exported tokenizer to pickle file
- Created flask app to run models based on user input
- Used axios in javascript for click event of submit button and posted to flask app which then returned ML predictions
- Display was completed using html and javascript
- Graphs done with matplotlib
- Bootstrap and css for design
- Deployed to Heroku

Website Demonstration



Discussion

01

General Conclusion

- Selecting the best model for a learning task and tuning the model is difficult and time consuming
- Book descriptions correlate with Scifi subgenres

02

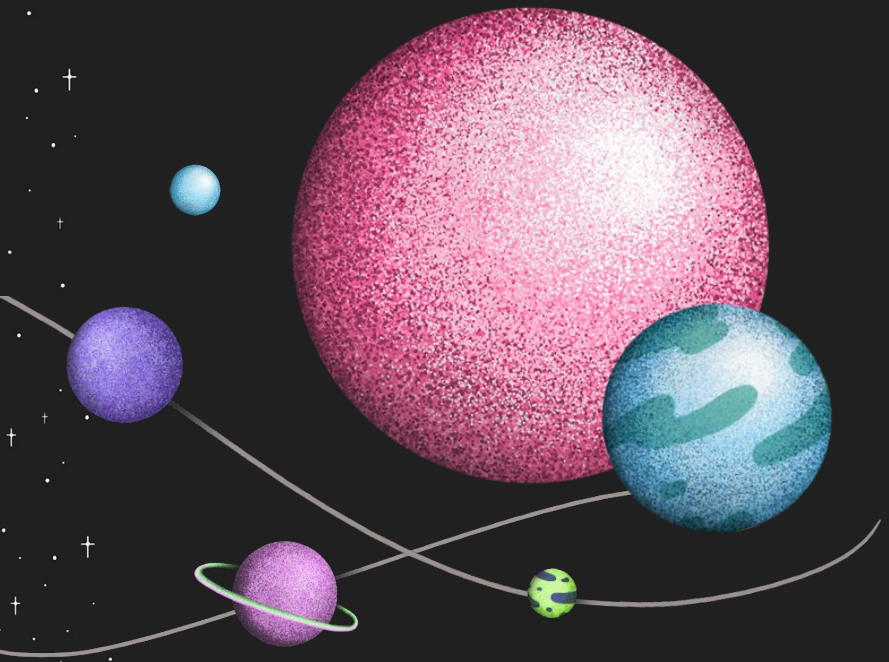
Challenges

- Finding the right machine learning model.
- Switching between pySpark and Keras and SKlearn
- Getting the flask app and APIs working properly

03

Future Goals

- Get more data to train the model - especially for the rating model
- Expand the website and ML models to all book genres and not just Scifi



THANKS!

Deployment link:
https://scifybook.herokuapp.com/
Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**