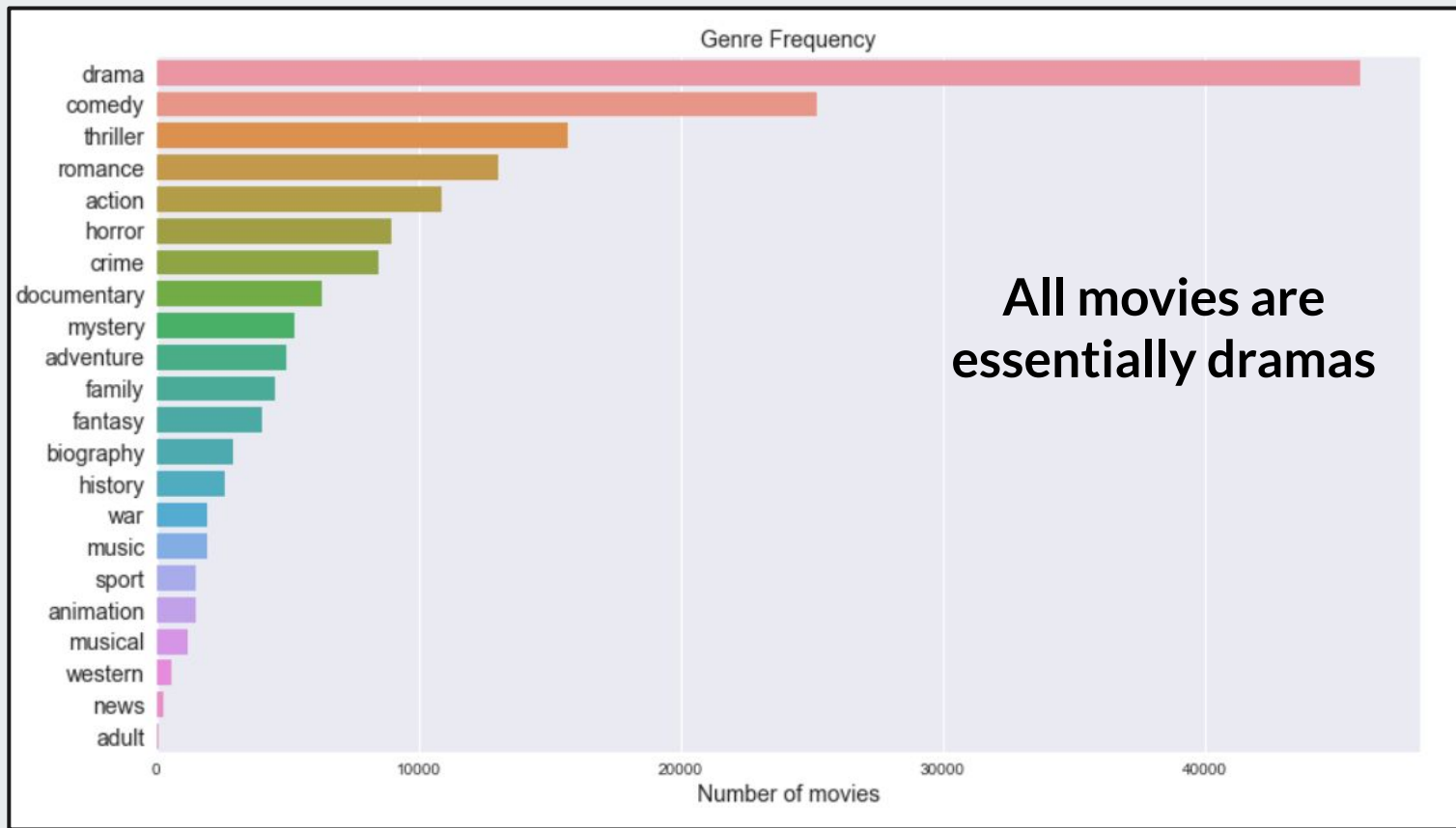# Classify movies genres from text data

by Igor Kirko

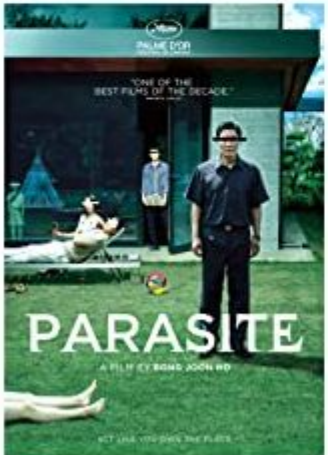# Multi label classification problem

22 unique genres

Genre Frequency

All movies are essentially dramas

Movie can have multiple genres

# Two datasets: Quality vs Quantity

Spoilers

⚠ The synopsis below may give away important plot points.

Synopsis

Ki-woo Kim (Choi Woo-Shik) is a young man living in poverty in the slums of a nameless South Korean city with his family - father Ki-taek Kim (Song Kang-ho), mother Chung-sook Kim (Jang Hye-jin), and sister Ki-jung (Park So-dam). The family lives in a garden unit where they struggle finding a Wi-Fi signal to sneak into in order to get access to the Internet to watch TV, get fumigated on, and have to watch men urinating in the alley outside their home. They make ends meet by doing menial tasks such as folding pizza boxes - and even then, get criticized by the pizza employees for messing up the boxes. Ki-woo's wealthy friend Min-hyuk pays the family a visit, giving them a gift - a rock that is supposed to bring those who have it wealth and prosperity. He then tells Ki-woo that he has been tutoring the teenage daughter of a very wealthy family. He is leaving to study abroad but is in love with the daughter, and knows that any of the other university boys would steal her away. He wants Ki-woo to be her tutor, knowing he will watch over her so that Min-hyuk can propose to her once she graduates high school. Ki-woo knows he isn't qualified since he isn't in college, but Min-hyuk promises to vouch for him, and so he agrees and has Ki-jung forge credentials for him to take to his interview.

Ki-woo interviews at the very wealthy Park family where he meets Mr. Park (Lee Sun-kyun), his scatterbrained wife Mrs. Park (Choi Yeo-jeong), and their children, teenage daughter Da-hye and young son Da-song. Ki-woo realizes he needs to flirt with Da-hye to get the job, and he does - Mrs. Park pays him an exorbitant amount of money and mentions they need an art tutor for Da-song. Da-song had a traumatic incident where he saw a "ghost" in the house and had a seizure and has been needing help with his art. Ki-woo introduces them to Ki-jung, who forges documents for herself and goes by "Jessica," and she too begins making money hand over fist. When the Park's limo driver takes her home, Ki-jung leaves her underwear in the car in order to get him fired: she then suggests her "Uncle" as the new driver - who is really Ki-taek. The last position is that of the housekeeper, Moon-kwang (Lee Jeung-eun). She has worked for the home since before the Park family lived there - she worked for the previous owner, an eccentric architect. In order to get her fired, the Kims exploit her allergy to peaches, causing her to have allergic reactions: they then convince Mrs. Park that she is seriously contagiously ill and cannot be around their son. She leaves, devastated, and Mrs. Kim is given the job, fooling the Park family into hiring the entire Kim family.

The Kim family enjoys their massive increase of income, and when the Park family decides to leave to go camping for Da-song's birthday weekend, they take the opportunity to stay in the huge Park house for the weekend. They spend the evening drinking and eating and making a mess of the place when the doorbell rings: it's Moon-kwang. She claims she was fired so quickly she left without being able to get something and just wants it back. Mrs. Kim reluctantly lets her in, and Moon-kwang runs into the basement and begins screaming, opening a secret passage behind some shelves. She goes to her husband, Geun-sae (Park Myung-hoon), who has secretly been living in the bunker ever since the previous owner moved out in order to hide from loan sharks. She gives him food while Mrs. Kim looks on in horror - she tells Moon-kwang she needs to leave, and as Moon-kwang begs her to let them stay, the rest of the Kim family (who had been eavesdropping) falls off the stairs and into view - and they call each other "dad," etc., which Moon-kwang films on her phone, realizing the con the family has pulled. She threatens to send it to the Park family and uses that so she and her husband can force the Kim family to do their bidding.

2:03 | Trailer   17 VIDEOS | 269 IMAGES
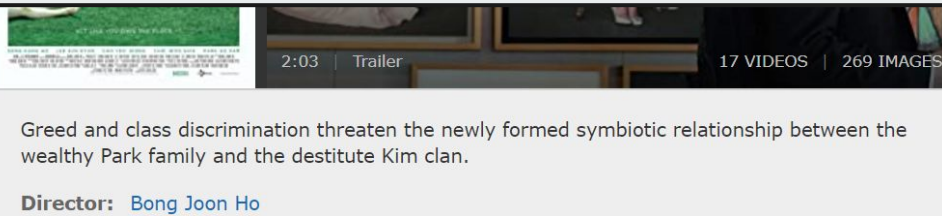
Greed and class discrimination threaten the newly formed symbiotic relationship between the wealthy Park family and the destitute Kim clan.

Director: Bong Joon Ho

1. **Short summary**
   Length   50-200 symbols
   Quantity 90k items

2. **Synopsis**
   Length   500-10 000 symbols
   Quantity 11k items

Solving the problem:

Pretraining **- Word2Vec**

Architecture **- Recurrent Neural Networks**

Evaluation **- F1-score**

RNN architecture

Embedding	use weights from w2v model. w2v dimension = size of embedding layer. *Trainable=False*
LSTM Bidirectional	prove to understand context better
Dropout	prevent quick overfitting
BatchNormalization	provides regularization, reduce generalization error
Output	*activation='sigmoid'* probability of occurrence does not depend on other classes
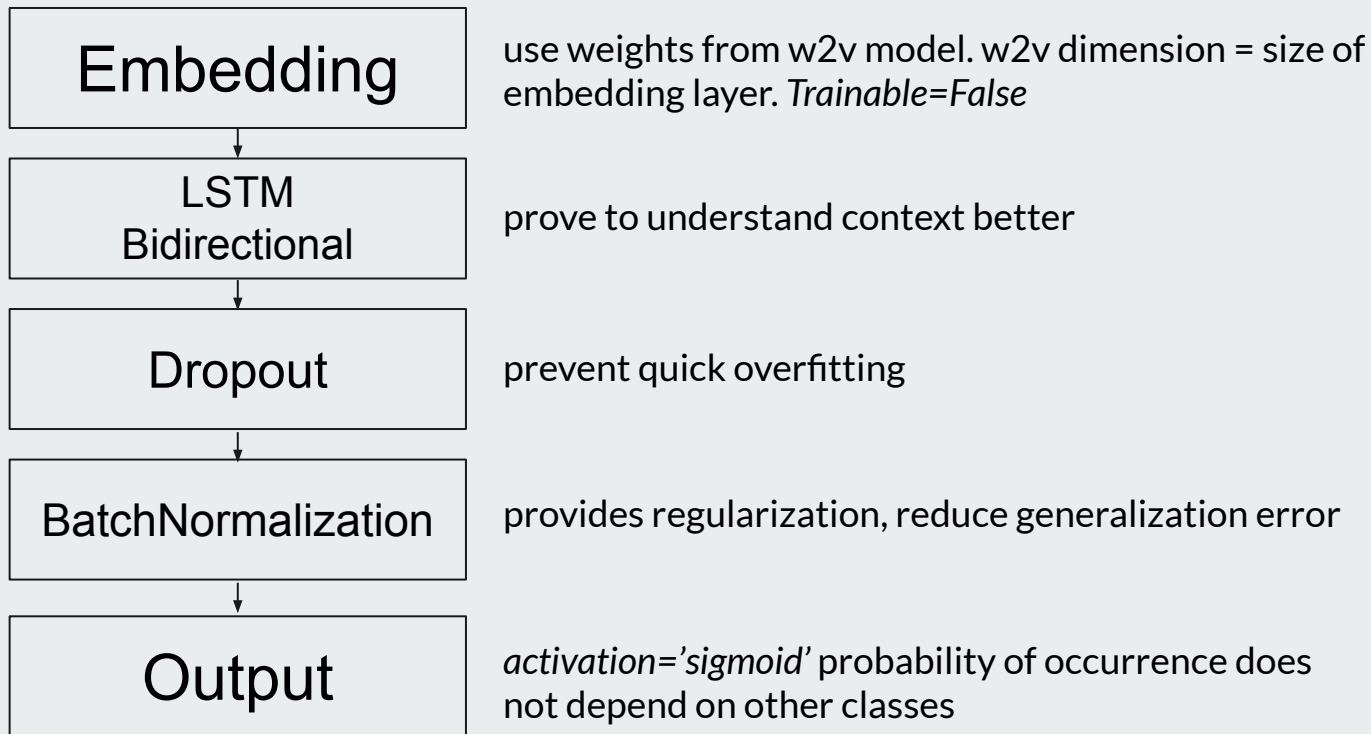
Result for each movie is a list of probabilities

```
predictions_2[1]

array([0.00631773, 0.00827481, 0.00323178, 0.02500881, 0.39342663,
       0.03074602, 0.00798566, 0.789581  , 0.01913959, 0.02047212,
       0.00702533, 0.04181992, 0.04003656, 0.01180624, 0.03862275,
       0.00330225, 0.45375243, 0.01017664, 0.0774362 , 0.00385976,
       0.0020665 ], dtype=float32)
```

1. Apply threshold at which we can assign genre to a movie.

2. Find best performing threshold value

Accuracy

Short summary

```
F1-measure: 0.5171 for 0.1 threshold
--------------------------------------
F1-measure: 0.5764 for 0.2 threshold
--------------------------------------
F1-measure: 0.5797 for 0.3 threshold
--------------------------------------
F1-measure: 0.5546 for 0.4 threshold
--------------------------------------
F1-measure: 0.5039 for 0.5 threshold
--------------------------------------
F1-measure: 0.4219 for 0.6 threshold
--------------------------------------
F1-measure: 0.3165 for 0.7 threshold
--------------------------------------
F1-measure: 0.1952 for 0.8 threshold
--------------------------------------
F1-measure: 0.0742 for 0.9 threshold
```

Synopsis

```
F1-measure: 0.5383 for 0.1 threshold
--------------------------------------
F1-measure: 0.5888 for 0.2 threshold
--------------------------------------
F1-measure: 0.5932 for 0.3 threshold
--------------------------------------
F1-measure: 0.5804 for 0.4 threshold
--------------------------------------
F1-measure: 0.5402 for 0.5 threshold
--------------------------------------
F1-measure: 0.4658 for 0.6 threshold
--------------------------------------
F1-measure: 0.3493 for 0.7 threshold
--------------------------------------
F1-measure: 0.2015 for 0.8 threshold
--------------------------------------
F1-measure: 0.0566 for 0.9 threshold
```

Predictions vs actual

Short summary

actual	predictions
[drama]	[drama]
[comedy, romance]	[comedy, drama, romance]
[action, adventure]	[action, comedy, crime, drama, thriller]
[drama]	[comedy, drama]
[horror]	[drama, horror]
[adventure, animation, comedy, family, fantasy]	[adventure, comedy, family, fantasy]
[comedy, drama, romance]	[comedy, drama]
[drama]	[drama]
[comedy, music]	[documentary, music]
[horror, thriller]	[horror, thriller]
[comedy, crime]	[comedy, drama, thriller]

Synopsis

actual	predictions
[documentary]	[biography, documentary, news]
[action, comedy, sport]	[action, comedy, documentary, sport]
[action, adult, adventure, comedy, drama, musi...	[action, documentary, sport]
[biography, documentary, news]	[biography, documentary, news]
[crime, documentary, fantasy, mystery]	[animation, crime, documentary]
[documentary]	[biography, documentary]
[documentary]	[biography, documentary, news]
[documentary]	[documentary]
[biography, comedy, documentary]	[biography, documentary, news]
[animation, documentary, fantasy, news, thriller]	[documentary, fantasy]
[action, adult, biography]	[biography]

Conclusion and Future work

Synopsis model showed slightly better results.
Overall accuracy for both models **~ 0.6** is good considering number of labels

Ways to improve model:

Train on more examples

Try different RNN architectures

THANK YOU