Classifying Famous *Fantasy* Stories

By: Samantha Chu

Problem Statement

The *Poets & Writers Magazine* publishes creative writing contests in their magazines each year.

Writing contest:
Write a short story that adds to the story of either Harry Potter, or the Lord of the Rings.

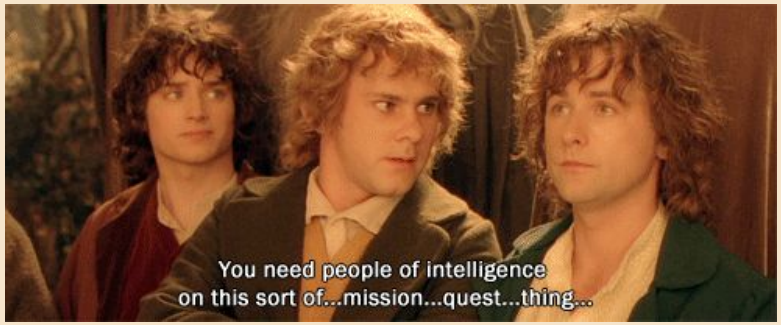
As students begin to submit their short stories online, there is a malfunction with the submissions, and the titles of each story are missing.

How will *Poets & Writers Magazine* sort the stories and categorize them as part of the Harry Potter series or the Lord of the Rings?

Strategy: Collect Data From Subreddits

Operating as a hired Data Scientist for *Poets & Writers Magazine*:

1. Collect 10,000 posts from two different subreddits
2. **Posts** were scraped from these two subreddits using the Pushshift API.
3. Total: roughly 20,000 posts



tolkienfans:
"This subreddit is a space for the Tolkien nerds of reddit to debate and discuss the whole Tolkien mythos. We emphasize serious discussion here over jokey/meme-based posts."

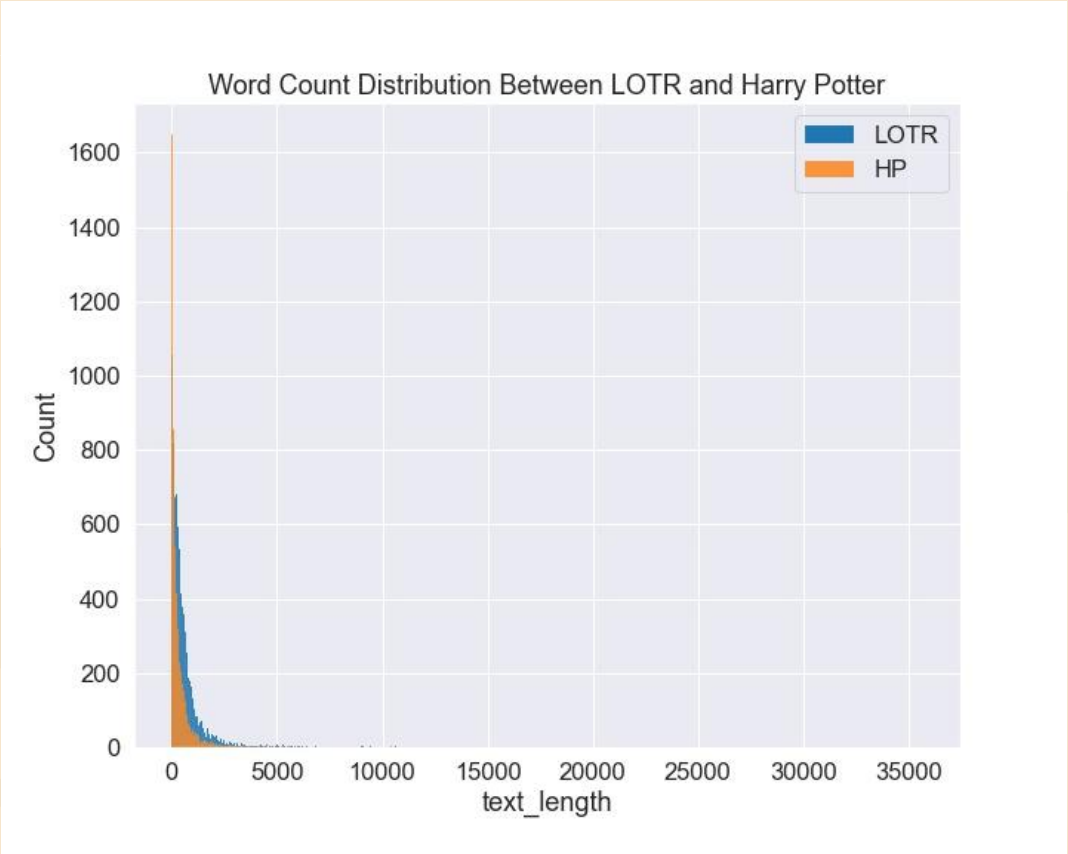
harrypotter:
"The place where fans from around the world can meet and discuss everything in the Harry Potter universe!"

Pre-Processing Data

- Removed moderator's '[removed]' and '[deleted]' text
- Combined the *title* and *selftext* columns
- Dropped any NA observations
- Removed special characters and created a Word Count column
- Dropped observations that had less than 10 words
- Dropped duplicates
- Created a stemmed text column
- Created a lemmatized text column
- Dropped any observations after stemming/lemmatizing that had less than 10 words.

Exploratory Data Analysis

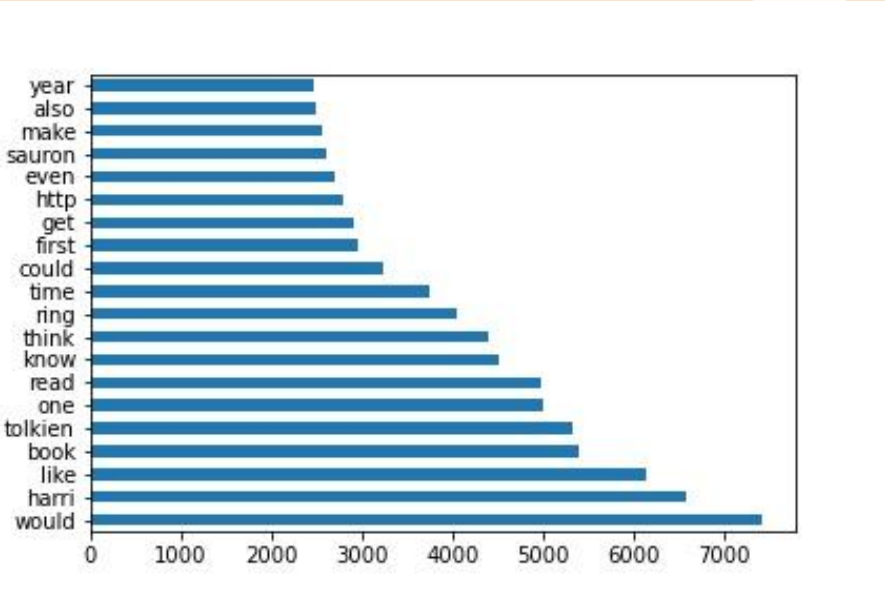
Distribution of word count in posts



Exploratory Data Analysis

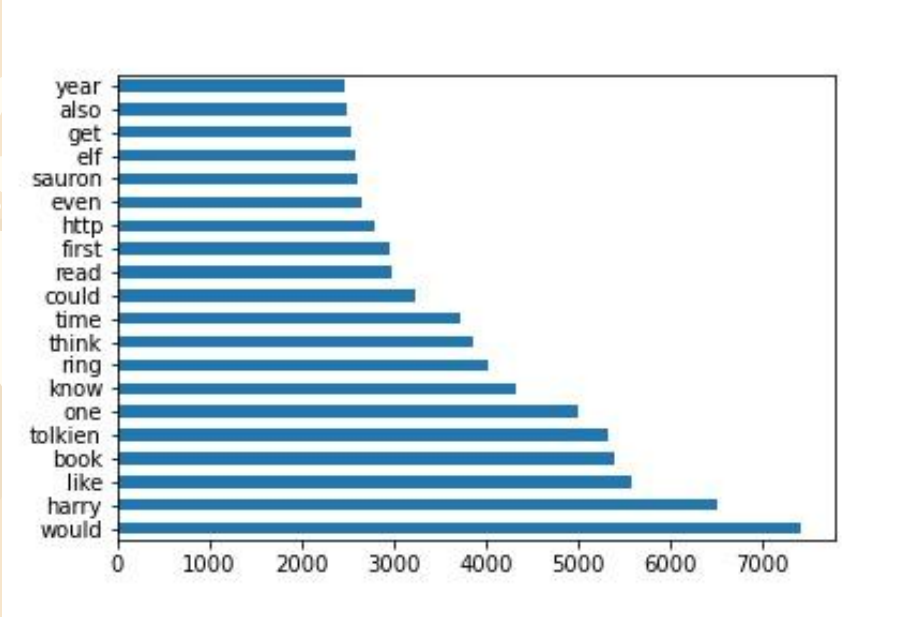
CountVectorized and Stemmed

Top 20 most common stemmed words



CountVectorized and Lemmatized

Top 20 most common lemmatized words



Modeling

01

Multinomial Naive Bayes

- CountVectorized, Stemmed: GridSearch
- CountVectorized, Lemmatized: GridSearch
- TfidfVectorized, Stemmed: GridSearch
- TfidfVectorized, Lemmatized: GridSearch

02

Logistic Regression

- CountVectorized, Stemmed: GridSearch
- CountVectorized, Lemmatized: GridSearch
- TfidfVectorized, Stemmed: GridSearch
- TfidfVectorized, Lemmatized: GridSearch

03

AdaBoost - Base: Decision Trees

- CountVectorized, Stemmed: GridSearch
- CountVectorized, Lemmatized: GridSearch
- TfidfVectorized, Stemmed: GridSearch
- TfidfVectorized, Lemmatized: GridSearch

04

Random Forest

- CountVectorized, Stemmed: GridSearch
- CountVectorized, Lemmatized: GridSearch
- TfidfVectorized, Stemmed: GridSearch
- TfidfVectorized, Lemmatized: GridSearch

Evaluation Metric: AUC, Optimized Accuracy → Neither predicting LOTR or Harry Potter correctly was more important than the other.

Modeling

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122

Modeling

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122

Modeling

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122

Modeling

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122

Modeling

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122

Modeling

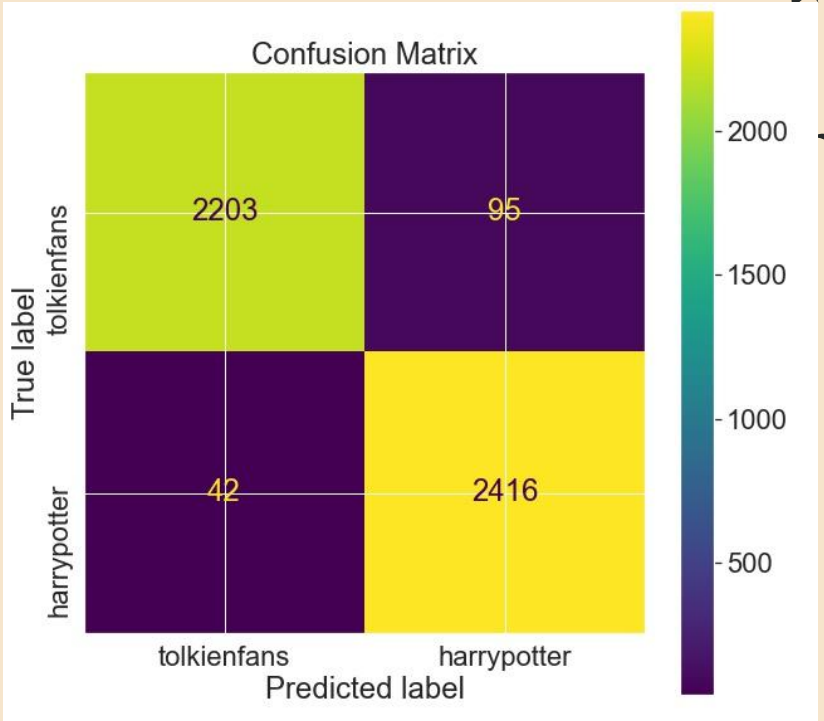
	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122

Observations:

- Stemming slightly **better** than Lemmatizing
- Best parameters for MultinomialNB:
  - CountVectorizer
    - Max_df = 0.9
    - Max_features = 5,000
    - Min_df = 2
    - Ngram_range = (1, 2)
  - TfidfVectorizer
    - Max_df = 0.8
    - Max_features = 5,000
    - Min_df = 1
    - Ngram_range = (1, 2)

Modeling

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122



Modeling

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122

	coefs	features
29	-13.065787	1st place
59	-13.065787	3rd place
93	-13.065787	accio
107	-13.065787	action tv
159	-13.065787	alan rickman
161	-13.065787	albu
162	-13.065787	albu dumbledor
224	-13.065787	amp auto
240	-13.065787	andromeda
250	-13.065787	animagu

Smallest values

	coefs	features
4427	-4.518841	tolkien
4877	-4.677564	would
3670	-4.745826	ring
3485	-4.855662	read
2514	-4.923723	like
3056	-4.989271	one
551	-5.117048	book
2394	-5.118815	know
3769	-5.186874	sauron
4325	-5.290931	think

Largest values

Modeling

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122

Modeling

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122

	coefs	features	e^coef
1985	-8.551078	harri	0.000193
2098	-5.434038	hogwart	0.004365
3306	-5.078030	potter	0.006232
4679	-4.703263	voldemort	0.009066
1239	-4.497534	dumbledor	0.011136
4011	-4.462378	snape	0.011535
2140	-4.162239	hp	0.015573
2049	-4.000079	hermion	0.018314
2001	-3.669953	harri potter	0.025478
4833	-3.660520	wizard	0.025719

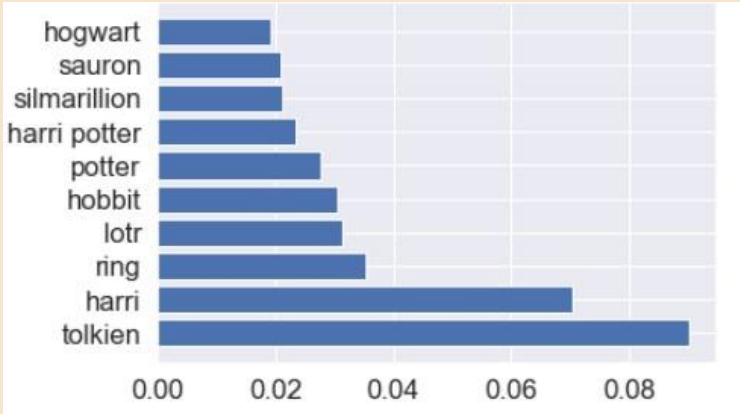
Smallest coefficients

	coefs	features	e^coef
4424	12.481889	tolkien	263521.093484
2088	6.996582	hobbit	1092.891788
2606	6.917830	lotr	1010.125899
3669	6.413539	ring	610.048982
3948	5.666969	silmarillion	289.156808
3768	5.182288	sauron	178.089904
1338	4.452551	elv	85.845660
1282	3.966184	earth	52.782748
2787	3.777674	middl earth	43.714237
2851	3.589088	morgoth	36.201061

Largest coefficients

Modeling

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	Multinomial NB	CountVectorizer	Stemmed	0.971961	0.971194	0.997536
1	Multinomial NB	CountVectorizer	Lemmatized	0.967475	0.966779	0.997227
2	Multinomial NB	TFIDF	Stemmed	0.975606	0.970353	0.997294
3	Multinomial NB	TFIDF	Lemmatized	0.975606	0.967830	0.997025
4	Logistic	CountVectorizer	Stemmed	0.990397	0.968461	0.995426
5	Logistic	CountVectorizer	Lemmatized	0.991588	0.966358	0.995728
6	Logistic	TFIDF	Stemmed	0.985139	0.972876	0.996521
7	Logistic	TFIDF	Lemmatized	0.986331	0.972456	0.996698
8	RandomForest	CountVectorizer	Stemmed	0.935791	0.927881	0.986320
9	RandomForest	CountVectorizer	Lemmatized	0.975957	0.958999	0.992629
10	RandomForest	TFIDF	Stemmed	0.979532	0.959420	0.992220
11	RandomForest	TFIDF	Lemmatized	0.937824	0.936712	0.986401
12	AdaBoost	CountVect	Stemmed	0.976237	0.960261	0.993695
13	AdaBoost	CountVect	Lemmatized	0.975466	0.961102	0.993525
14	AdaBoost	TFIDF	Stemmed	0.972873	0.953532	0.991864
15	AdaBoost	TFIDF	Lemmatized	0.954227	0.947645	0.988122



Feature Importances

Conclusions/Recommendations

1. To the *Poets & Writers Magazine*, use the Multinomial Naive Bayes model to predict whether the story is a Harry Potter series story or a Lord of the Rings story.

2. Words that are the most predictive include: “dumbledore, tolkien, ring, hobbit, sauron, harry, hogwart” -- Flag these words

3. Harry Potter predictive words tended to lean towards names, whereas the predictive words for LOTR were more content based.

Thanks

Sources

Reddits

- **Tolkien:** https://www.reddit.com/r/tolkienfans/
- **Harry Potter**: https://www.reddit.com/r/harrypotter/

API

- https://github.com/pushshift/api

Appendix

Custom Stopwords

harry

tolkien

hp

potter

lotr

No Names Models

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	MultinomialNB	CountVectorizer	Stemmed	0.952381	0.950615	0.991785
1	Logistic	TFIDF	Stemmed	0.971174	0.955702	0.992427
2	RandomForest	TFIDF	Stemmed	0.965381	0.932175	0.982628
3	AdaBoost	CountVect	Stemmed	0.949484	0.938745	0.986431

No Names Models

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	MultinomialNB	CountVectorizer	Stemmed	0.952381	0.950615	0.991785
1	Logistic	TFIDF	Stemmed	0.971174	0.955702	0.992427
2	RandomForest	TFIDF	Stemmed	0.965381	0.932175	0.982628
3	AdaBoost	CountVect	Stemmed	0.949484	0.938745	0.986431

Logistic: Smallest coefficients

	coefs	features
2094	-5.994882	hogwart
4678	-5.421240	voldemort
1242	-5.043886	dumbledor
4023	-4.740321	snape
2046	-4.672461	hermion
4832	-4.224682	wizard
4696	-3.617568	wand
3983	-3.520462	siriu
3730	-3.452524	ron
3873	-3.303405	seri

	coefs	features
2083	8.181247	hobbit
3681	6.871067	ring
3960	6.327872	silmarillion
3781	5.335286	sauron
1339	4.790458	elv
1284	4.500190	earth
2792	4.261726	middl earth
2857	3.848599	morgoth
1748	3.828539	frodo
2791	3.806508	middl

Logistic: Largest coefficients

No Names Models

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	MultinomialNB	CountVectorizer	Stemmed	0.952381	0.950615	0.991785
1	Logistic	TFIDF	Stemmed	0.971174	0.955702	0.992427
2	RandomForest	TFIDF	Stemmed	0.965381	0.932175	0.982628
3	AdaBoost	CountVect	Stemmed	0.949484	0.938745	0.986431

No Names Models

	Model	Transformer	Stemmed/Lemmed	Train_acc	Test_acc	AUC
0	MultinomialNB	CountVectorizer	Stemmed	0.952381	0.950615	0.991785
1	Logistic	TFIDF	Stemmed	0.971174	0.955702	0.992427
2	RandomForest	TFIDF	Stemmed	0.965381	0.932175	0.982628
3	AdaBoost	CountVect	Stemmed	0.949484	0.938745	0.986431