Building a Model to Differentiate Between Related Subreddits

—

Andrew Gossage

Andrew Gossage - Data Scientist

www.andrewgossage.net

Why should we care?

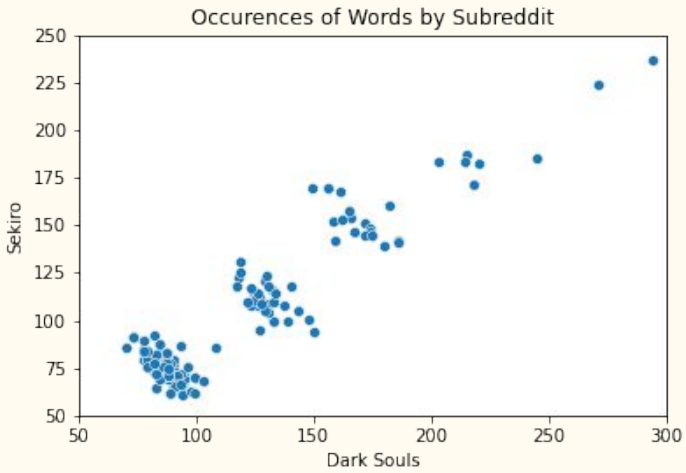
What Data was used?

Data

r/DarkSouls and r/Sekiro

- 30,000 posts
- 10,000 from r/DarkSouls
  20,000 from r/Sekiro
- 20,400 after removing those with no text, all lost were from Sekiro

Data



Words by Sekiro

	count	dark_souls	sekiro	gap
sword	565.0	316.0	249.0	67.0
does	485.0	238.0	247.0	9.0
parry	407.0	210.0	197.0	13.0
better	408.0	218.0	190.0	28.0
butterfly	390.0	208.0	182.0	26.0
good	405.0	224.0	181.0	43.0
try	410.0	234.0	176.0	58.0
times	336.0	164.0	172.0	8.0
attack	336.0	164.0	172.0	8.0
getting	324.0	155.0	169.0	14.0

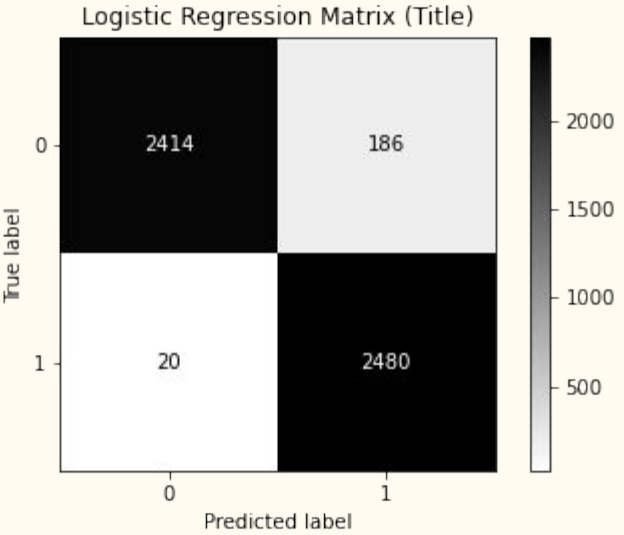
Words by Dark Souls

	count	dark_souls	sekiro	gap
subreddit	8736.0	8736.0	0.0	8736.0
sword	565.0	316.0	249.0	67.0
does	485.0	238.0	247.0	9.0
second	404.0	237.0	167.0	70.0
try	410.0	234.0	176.0	58.0
good	405.0	224.0	181.0	43.0
better	408.0	218.0	190.0	28.0
parry	407.0	210.0	197.0	13.0
butterfly	390.0	208.0	182.0	26.0
soul	313.0	187.0	126.0	61.0

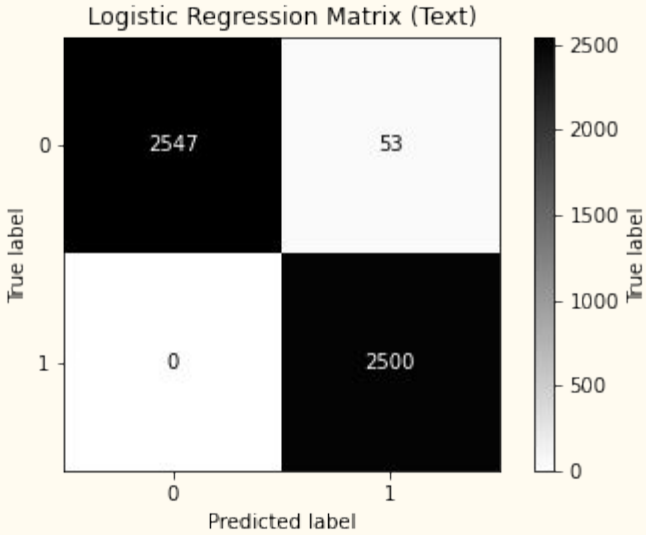
Models

- Logistic Regression Trained on Titles
- Logistic Regression Trained on Text
- Random Forest Trained on Text

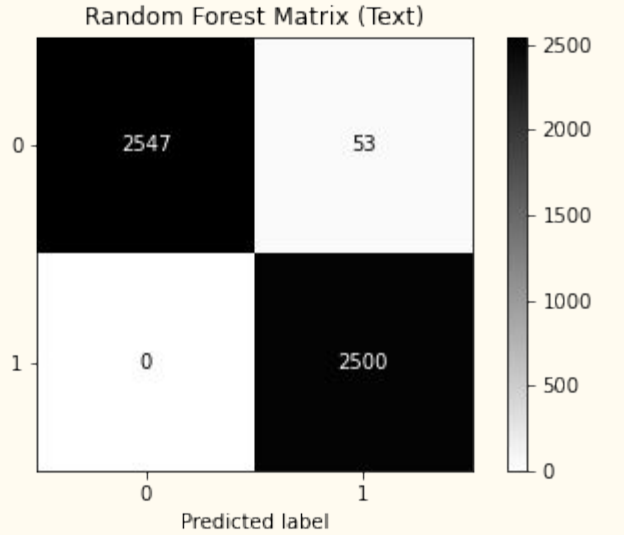
Model Performance



Accuracy: ~0.96



Accuracy: ~0.99

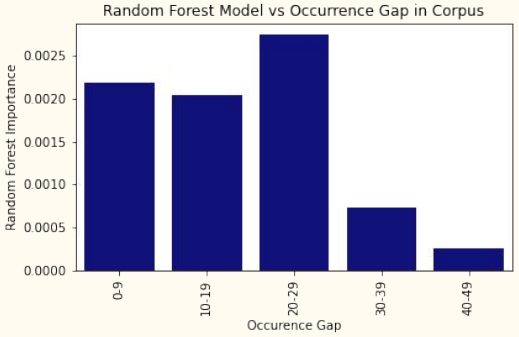


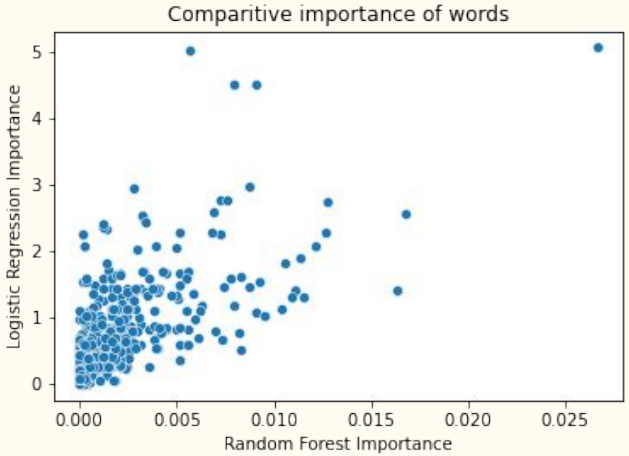
Accuracy: ~0.99

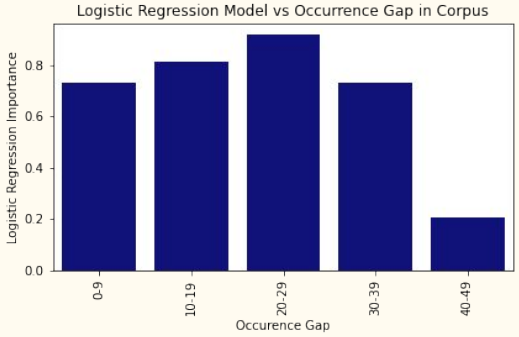
Are the Models Identical?

Entry	Random Forest Importance	Logistic Regression Coeficients	Logistic Regression Importance
deleted	0.026606	-5.066097	5.066097
gauntlets	0.016747	-2.562937	2.562937
bar	0.016296	-1.419464	1.419464
blighttown	0.012711	2.755551	2.755551
gwyn	0.012700	2.285476	2.285476
grind	0.012141	-2.074051	2.074051
humanity	0.011499	1.306646	1.306646
warrior	0.011330	-1.890887	1.890887
mean	0.011040	1.408761	1.408761
skills	0.010916	-1.308653	1.308653
prayer	0.010566	-1.837720	1.837720
reddit	0.010423	-1.137326	1.137326
teach	0.009544	-1.032802	1.032802
thinking	0.009241	-1.533699	1.533699
strength	0.009109	1.096729	1.096729
mod	0.009033	-4.515762	4.515762
shield	0.008747	1.455295	1.455295
build	0.008736	2.971656	2.971656
different	0.008317	1.614436	1.614436
flame	0.008194	0.787640	0.787640

No But There is a Strong Correlation.







Demonstration

Conclusion

We were able to create and deploy a model to be used in predicting the relationship between text and one of two video games.

A large dataset is necessary when trying to predict between related subjects.

The models trained on text where better because they are able to use more words to differentiate.

Next Steps

Create a model that predicts between several targets.

Tie this model to a web crawler.

Perform more in depth testing to ensure applicability to texts outside of reddit posts.