IN HARDWOOD GROVES

The same leaves over and over again!
They fall from giving shade above,
To make one texture of faded brown
And fit the earth like a leather glove.

Before the leaves can mount again
To fill the trees with another shade,
They must go down past things coming up.
They must go down into the dark decayed.

They *must* be pierced by flowers and put
Beneath the feet of dancing flowers.
However it is in some other world
I know that this is the way in ours.

When was that poem written? Ask my computer.

Chaim Gluck, General Assembly, Data Science Immersive

Overview

- Conception
- Collection
- Cleansing
- Count Vectorization
- Creation of Models
- Conclusions

Conception: The Idea

- Can a computer be trained to identify when a poem was written based on it's style?
- If it can, with what reliability?
- What changes can we observe between the poems in different eras?

Collection

- No suitable digital corpus is available.
- When there is no data, take “The Road Not Taken”, Webscrape!
- I collected **34,000** poems from a poem collection website

Cleansing

- Interesting Challenge: manually removing the HTML tags
- Go words, not stop words
- Deciding which classes to create and how to create them

Count Vectorization

- Over **42,000** features
- The models worked best with all of the words
- Using ngrams didn't help the model's performance. This warrants further exploration; words next to each other should contribute to style.

Creation of Models

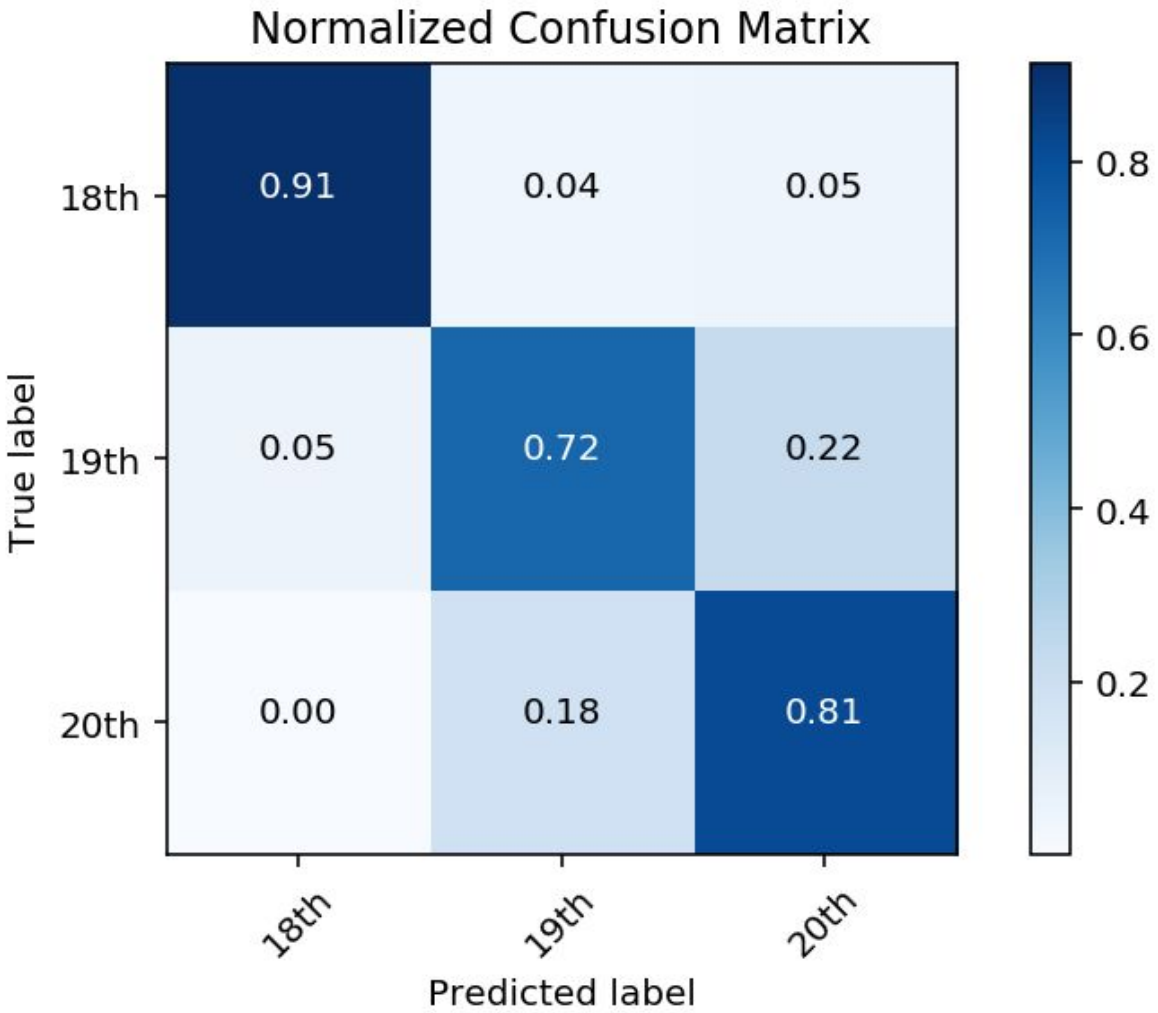
- Ensemble methods create multiple decision trees to model which features determine class.
- Naive Bayes models assume the independence of all features and calculate the likelihood of each poem belonging to each class given their words.

Conclusion

- The best performing model, with an 81% accuracy score, was a Multinomial Naive Bayes Classifier.
- The distinction between the 18th century and the other classes is clearest, with 91% recall. The 19th and 20th, scored 72% and 81%, respectively. Given how language has evolved, it makes sense that it is more difficult to distinguish between these.

Conclusion, Continued

- The style of poetry between different eras is distinct enough that a computer can identify the likely era.
- Word usage, in particular frequency, is enough of an indicator to model this distinction.
- More deliberate choices regarding class breakdown and class membership will help increase the model's performance.


Normalized Confusion Matrix
True label
18th
19th
20th
0.91
0.04
0.05
0.05
0.72
0.22
0.00
0.18
0.81
18th
19th
20th
Predicted label
0.8
0.6
0.4
0.2

Some Interesting Insights

- Religious poetry seems to have faded over time and clearly helps the model predict an older poem.
- If I had to describe these lists in one word, I would say the 18th century feels revelatory, the 19th romantic, and the 20th descriptive.

	18th		19th		20th
thy	0.012041	thy	0.005893	like	0.006553
shall	0.005759	thou	0.005669	love	0.003508
god	0.005050	like	0.004788	man	0.003484
ye	0.004369	thee	0.004430	old	0.003389
thou	0.004315	shall	0.004258	eyes	0.002969
let	0.004051	love	0.004154	night	0.002842
wi	0.003748	day	0.002937	day	0.002834
lord	0.003626	heart	0.002825	little	0.002779
love	0.003626	little	0.002810	said	0.002580
like	0.003586	life	0.002676	heart	0.002565
man	0.003356	soul	0.002511	know	0.002501
thee	0.003107	away	0.002384	light	0.002461
heart	0.002884	eyes	0.002377	life	0.002438
day	0.002817	man	0.002377	come	0.002374
grace	0.002439	night	0.002370	men	0.002350
oer	0.002182	old	0.002370	long	0.002342
soul	0.002061	sweet	0.002347	time	0.002342
th	0.001973	long	0.002280	shall	0.002239
life	0.001872	light	0.002272	away	0.002231
eyes	0.001852	oer	0.002123	say	0.002033