WINE REVIEWS

NPL ANALYSIS AND CLASSIFICATION

DATA SETS USED

- The Kaggle dataset used was scrapped off of WineEnthisuast website (https://www.winemag.com/?s=&drink_type=wine) https://www.kaggle.com/zynicide/wine-reviews
- Data contains wine review information including a description of the wine, and the wine variety.

OBJECTIVE

- The objective of this exercise is to test whether it is possible for a neural network to predict a certain wine variety (e.g. Pinot Noir, Grigio, etc) from an expert's description.

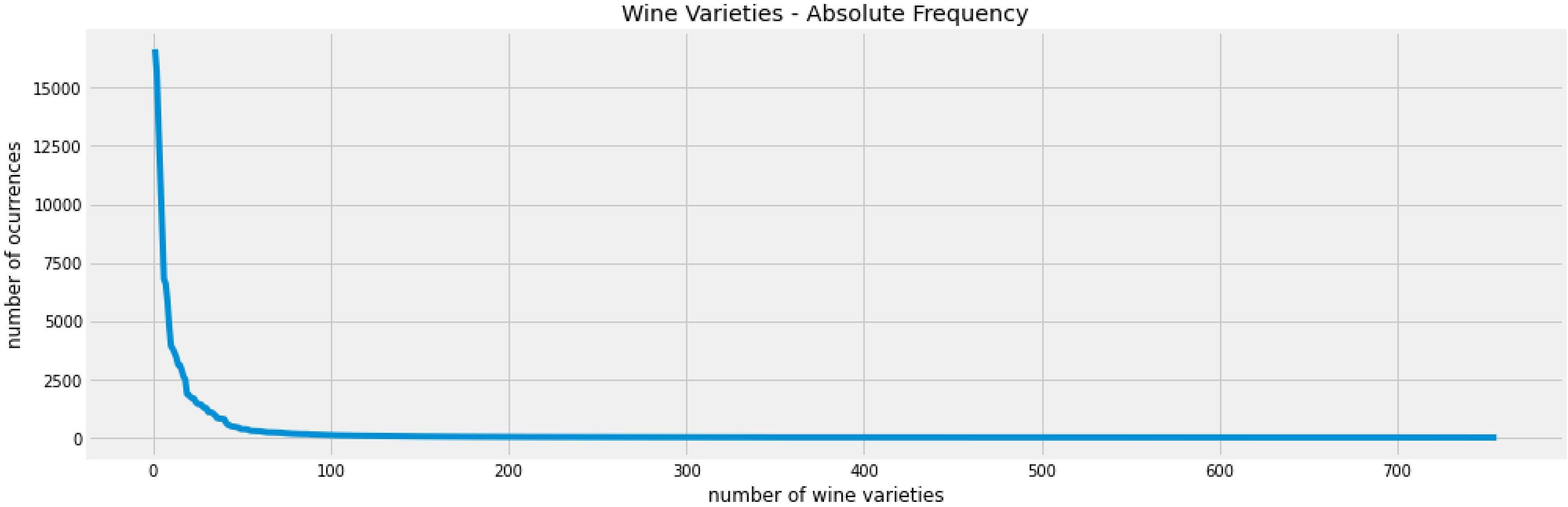
EXPLORATORY ANALYSIS

- After initial cleaning, we were left169k unique reviews.
- Below a few examples:

	description	variety
0	Aromas include tropical fruit, broom, brimston...	White Blend
1	This is ripe and fruity, a wine that is smooth...	Portuguese Red
2	Tart and snappy, the flavors of lime flesh and...	Pinot Gris
3	Pineapple rind, lemon pith and orange blossom ...	Riesling
4	Much like the regular bottling from 2012, this...	Pinot Noir

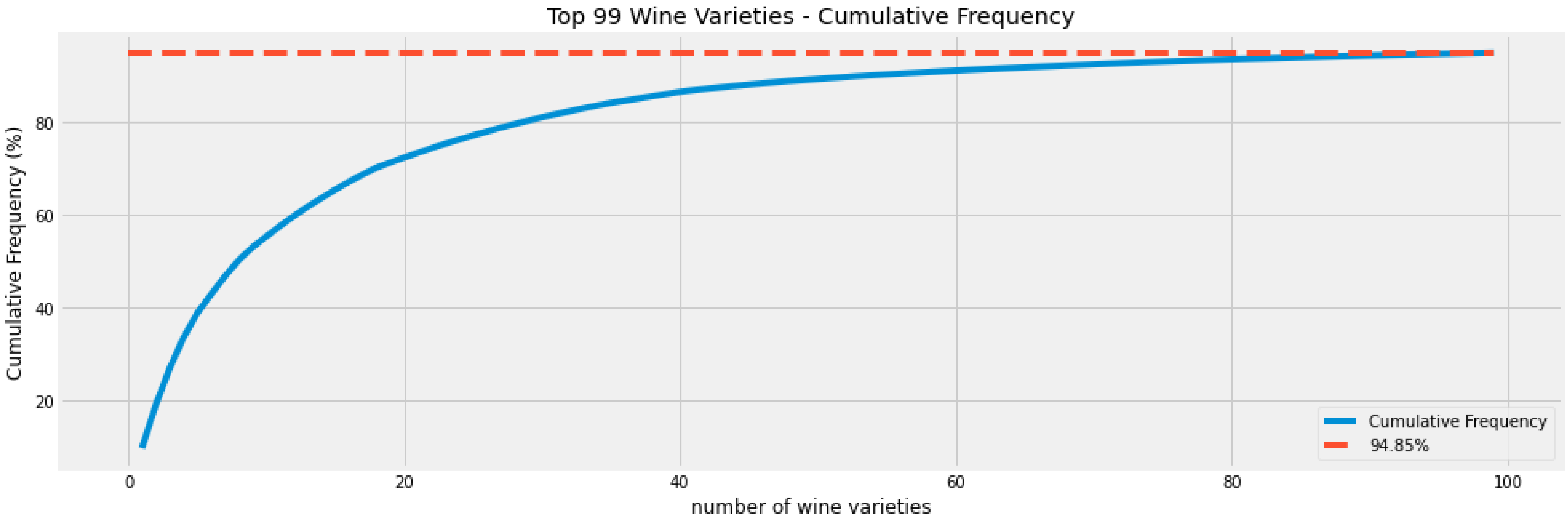
EXPLORATORY ANALYSIS

- There are 707 wine varieties listed in the dataset, with a highly uneven frequency distribution.
- Most of them have no more then 100 reviews, which would not be nearly enough to train a Neural Network appropriately.



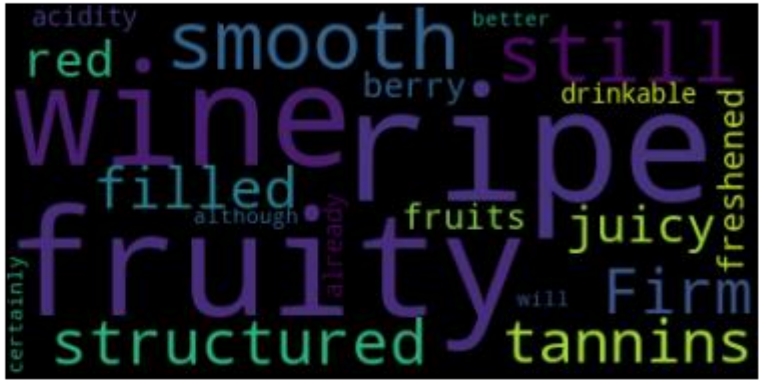
EXPLORATORY ANALYSIS

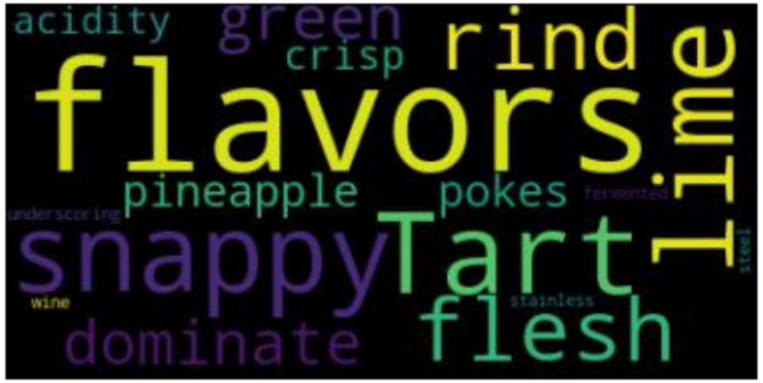
- If all wine varieties with less then 100 reviews are excluded, the dataset will still contain around 95% of the data and we are left with only the top 99 most common varieties in the dataset

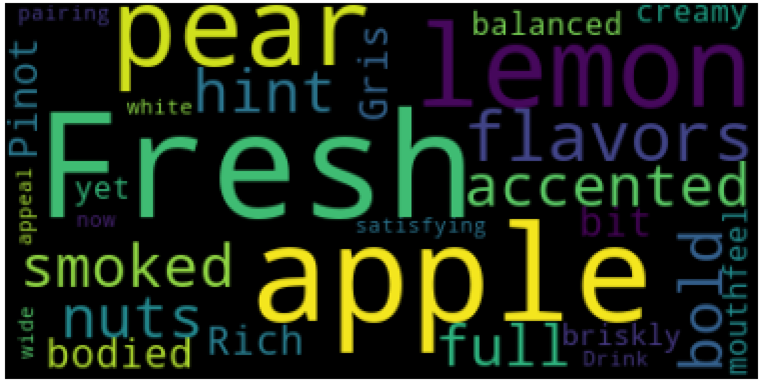


EXPLORATORY ANALYSIS

- After additional cleaning, words that were too common to be meaningful were deleted, as well as any actual wine varieties (it would give away the answer to the NN)
- As an illustration, here are three wine reviews' word clouds showing the most common words in each of them.

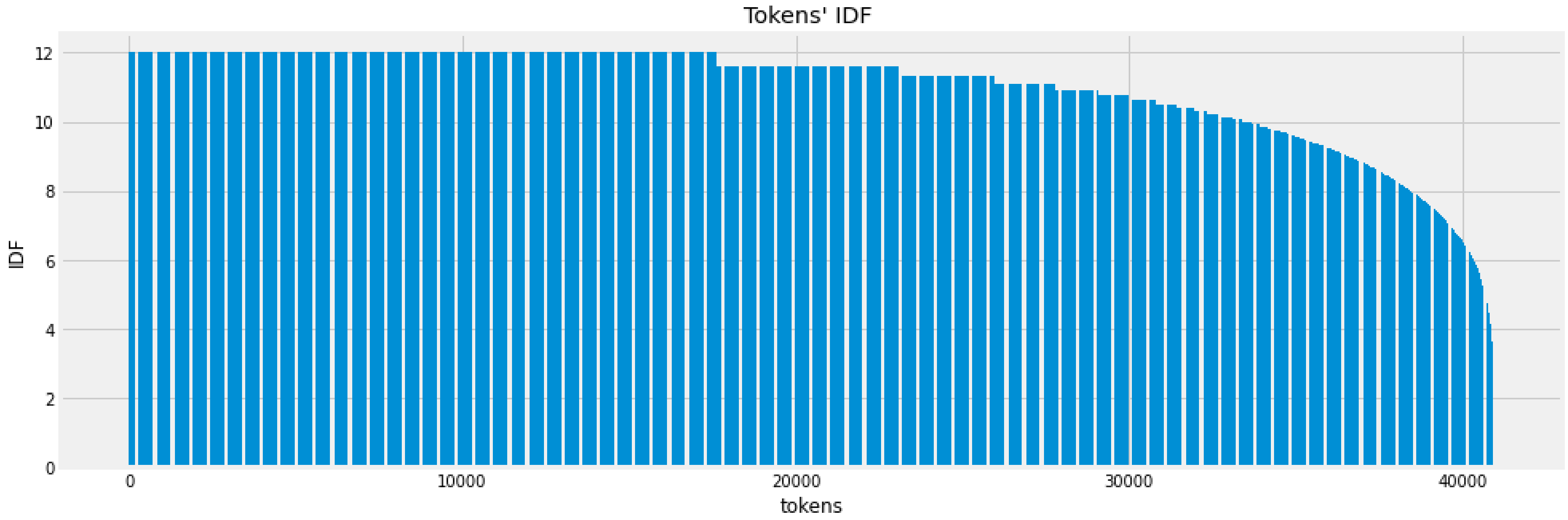






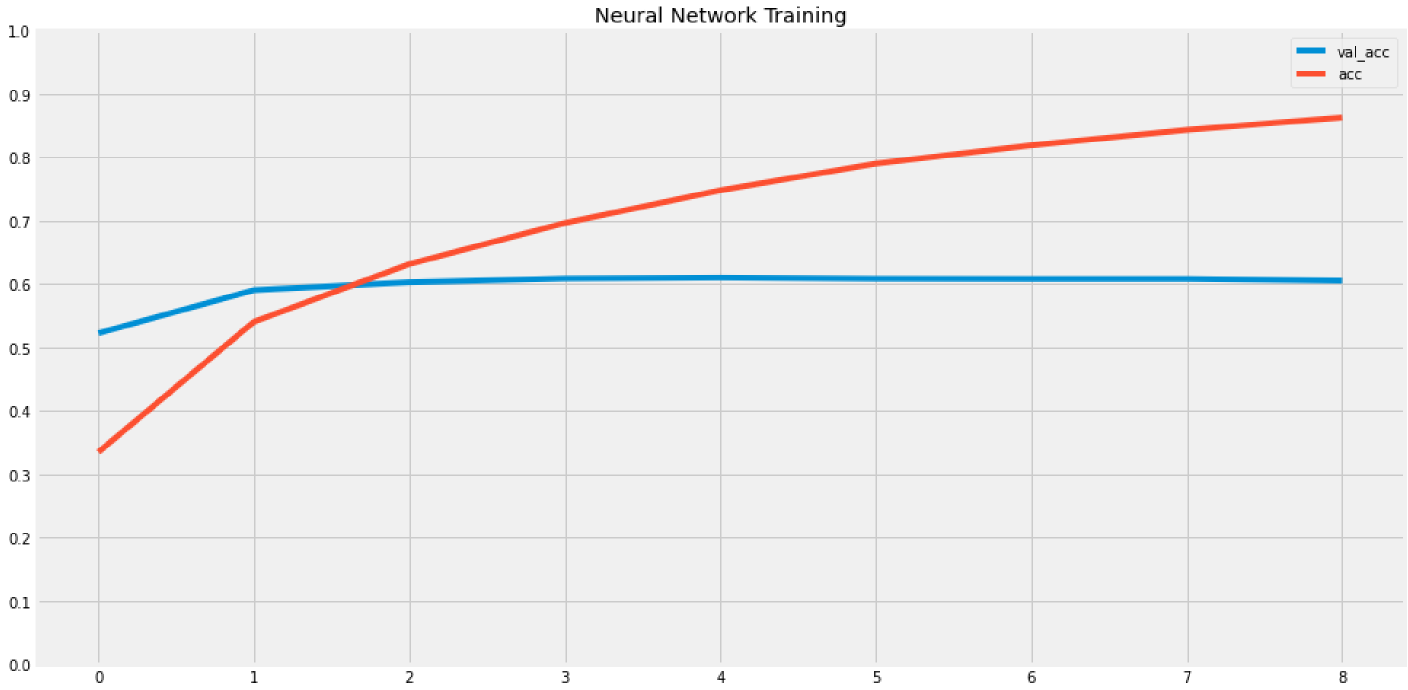
EXPLORATORY ANALYSIS

- With a clean set of tokenized terms, the Inverse Document Frequency (IDF) of each term was calculated.
  - *IDF is a measure of the rareness of a term*
- Most of the over 40k terms left are similarly infrequent as shown in the IDF chart below:



RESULTS

- Upon learning from the data, a Deep Neural Network was able to correctly predict the wine variety of a sample 61% of the time, given 99 possible varieties to choose from.



FURTHER INVESTIGATION

Can results be improved if other variables such as the review score is used?

THANK YOU