Identifying Fund Securities with NLP

Goal

Given a database full of identified fund securities and a list of unidentified fund securities, create a way to identify the unidentified fund securities based on the description only.

Fund Descriptions

Fund Descriptions can come in many different forms depending on the source of the data. The description could be concatenated, words abbreviated, and even enriched with more data than needed. This is why NLP would be a great solution for this problem we are trying to solve.

What words are the most important parts of a fund description?

Database Description

'JP Morgan US Small Company R6'

Unidentified Fund Description

'6130 JP Morgan US Small Company Fund R6'

Unidentified Fund Description Broken Down:

- Fund Family: 'JP Morgan'
- Fund Class: 'R6'
- Other Words in Description: '6130, US, Small, Company, Fund'

Feature Engineering

When identifying fund descriptions, Fund Family and Fund Class are very important parts of a fund description. This is why it is important to identify/classify these 2 features of a fund description.

Description	Family	Class
Prudential International Equity A	Prudential	A
Schwab Investor Money Fund	Schwab	No Class
Templeton Growth A	Templeton	A

How do we identify Fund Family and Fund Class?

Using a list of Fund Family Names and Fund Classes, we could search our descriptions for matching names or classes within the description. But if the data is available a better option is to use training sets to train models to identify Fund Family and Fund Class.

What to do with this new data(Family and Class)?

Because we have identified the Fund Family and Fund Class of a description, we can now build our NLP model on a smaller subset of data by filtering out the data in our database based on the Fund Family and Class.

Example:

Unidentified Description: 'Aston Small Cap Fund N'

Filtered Database Data (Family = 'Aston' and Class = 'N'):

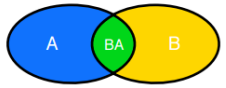
Description	Family	Class
Aston/ABN Amro High Yield Bond N	Aston	N
Aston/TCH Investment Grade Bond N	Aston	N
Aston Small Cap N	Aston	N

We then vectorize the descriptions from our smaller dataset and use an NLP model (Naïve Bayes Model in this example) to help us identify the best matching security.

	abn	amro	aston	bond	cap	grade	high	investment	n	small	tch	yield
Aston/ABN Amro High Yield Bond N	1	1	1	1	0	0	1	0	1	0	0	1
Aston/TCH Investment Grade Bond N	0	0	1	1	0	1	0	1	1	0	1	0
Aston Small Cap N	0	0	1	0	1	0	0	0	1	1	0	0

This result is called **Bayes' theorem**

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$



Benefits of creating a filtered NLP model :

- Smaller subset of data is used to create model
- More accurate results
- Performance (vectorization, model creation, and model calculation is faster)
- Less memory is used (vectorization is done on a small subset of descriptions vs the entire database of descriptions)

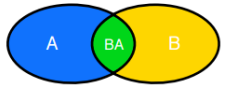
With our created model, we can feed the unidentified description into the model so the model can provide us the best matching fund security in the database.

Unidentified Description: 'Aston Small Cap Fund N'

	abn	amro	aston	bond	cap	grade	high	investment	n	small	tch	yield
Aston/ABN Amro High Yield Bond N	1	1	1	1	0	0	1	0	1	0	0	1
Aston/TCH Investment Grade Bond N	0	0	1	1	0	1	0	1	1	0	1	0
Aston Small Cap N	0	0	1	0	1	0	0	0	1	1	0	0

This result is called **Bayes' theorem**

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$



NLP Identified Database Description: 'Aston Small Cap N'