LCBO Whisky Similarity Analysis

Nelson Siegel

August 2019

Contents

1 Introduction

Finding good whiskies is difficult. Finding good whiskies for good prices is even more difficult! Finding good whiskies for good prices that are available at LCBO is almost impossible.

Thus the intent of this project was to develop a method for finding whiskies that are similar to ones that the user already prefers, but might be cheaper or better value.

In order to do this, data was collected from Reddit on whisky reviews, as well as data from LCBO. This data was combined, then a Word2Vec model was trained on the Reddit reviews. Next, similarity was calculated using Word Mover Distance with the trained Word2Vec model and generated for the entire dataset.

2 Background

Word2Vec is a shallow neural network that calculates similarities between two words by finding how often the occur near each other [1]. By training a Word2Vec model we can then run Word Mover Distance to calculate document similarity.

Word Mover Distance matches up words in two documents in the method that creates the least "difference" between the documents [2]. See Figure 1 for an example of how two completely different documents can have high similarity.

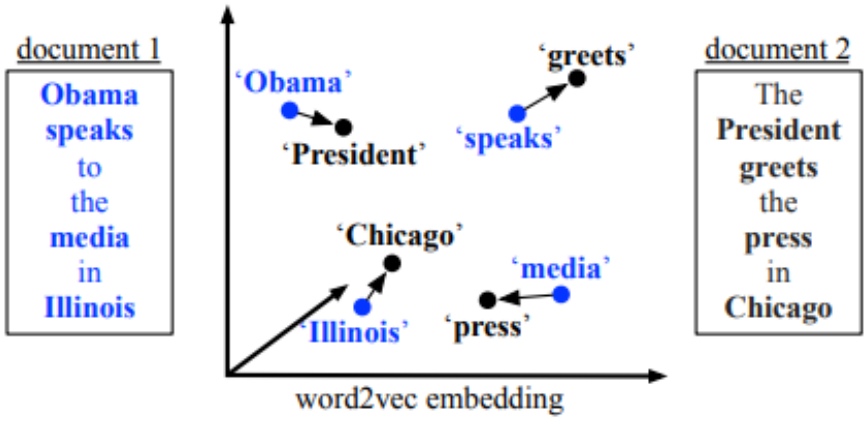

Figure 1: Word Mover distance matches similar words between documents [2].

3 Research Question

The intent of this analysis was to answer the following two questions:

1. Given a whiskey that the user enjoys, how can we tell other whiskeys they will enjoy? The method used to determine is flavour similarity Word Mover Distance with Word2Vec on whisky review text.

2. Based on the prices available for these whiskies at the LCBO, what is the best value purchase the user could choose that they would enjoy the most for the least cost? To do this a Rating per Price is displayed.

4 Methodology

4.1 Data Collection

4.1.1 Reddit Whisky Reviews

In order to collect enough descriptive reviews to perform analysis, the Reddit Whisky Network Review Archive [3] was used to gain a list of Reddit reviews. Then full text was gathered using the praw wrapper for the Reddit api.

4.1.2 LCBO Data

To collect data for LCBO products was more difficult. There is no official documented LCBO api. Thankfully I was able to find an undocumented API at www.foodanddrink.ca, which is another LCBO run site. There are however, a few limitations to the API, most notably that you cannot search by product name but need to include a product number. As such I created a script that tried every possible product number and logged the results.

Another limitation is that a link to the LCBO site is not included from the API. In order to gather links from to the main LCBO api site, another script was required to search the LCBO site for the product number and grab the link.

4.2 Modelling

4.2.1 Whisky Matchup

One of the most difficult aspects of the project was combining the Reddit Whisky Review archive data with the LCBO data. This was difficult because whisky names are written differently. The general method that I ended up using was as follows:

- Generate a list of keywords to search for. This was initially done by finding non-English words in whisky names but was later appended to include n-grams, as well as specific other words that could distinguish them, such as flavour types, special designations, or cask types.

- Each whisky in both Reddit and LCBO was then analysed to extract keywords as well as age from whiskies which had a defined match.
- All whiskies in the two datasets that had the exact same keywords as well as the exact same age were matched. In cases where more than one whisky matched, the python library Fuzzywuzzy was used to use a fuzzy matching algorithm to find the most similar name.

4.2.2 Word2Vec Training

The Word2Vec model was trained on the text of all of the reviews. For each word, a list of similar words can be output. Figure 2 shows an example of the results from the model.

```
1  # Test model
2  word = 'coffee'
3  model.wv.most_similar(positive=word)
```

executed in 9ms, finished 20:07:54 2019-08-06

```
[('espresso', 0.7701224684715271),
 ('milk', 0.7517192363739014),
 ('dark', 0.705661416053772),
 ('nib', 0.7011394500732422),
 ('mexican', 0.6960846185684204),
 ('leather', 0.6845758557319641),
 ('cacao', 0.6811543703079224),
 ('fudge', 0.6733429431915283),
 ('milky', 0.6686280369758606),
 ('bean', 0.6618437767028809)]
```

Figure 2: Results from trained Word2Vec model for the word 'coffee'.

4.2.3 Word Reduction with TF-IDF

In order to run Word Mover Distance in a reasonable amount of time, TF-IDF (Term Frequency, Inverse Document Frequency) was used to reduce the length of whisky descriptions. This was set to a limit of 50.

4.2.4 Word Mover Distance

Word Mover Distance was then run comparing every whisky to every other whisky. This was coded as a multicore process in order to decrease computation time as the number of whiskies exponentially increases the computation.

Figure 3 shows the results for the Wild Turkey 101 Whisky.

```
Query:
WILD TURKEY 101 KENTUCKY STRAIGHT BOURBON
1.0
WILD TURKEY 101 KENTUCKY STRAIGHT BOURBON
0.7391308417152184
WILD TURKEY RARE BREED KENTUCKY STRAIGHT BOURBON
0.6976657615110742
FOUR ROSES SMALL BATCH BOURBON
0.6748207106019719
COLONEL E.H. TAYLOR SINGLE BARREL KENTUCKY STRAIGHT BOURBON
0.6670327802999619
WELLER ANTIQUE 107 ORIGINAL WHEATED STRAIGHT BOURBON
0.6599524160202556
EVAN WILLIAMS SINGLE BARREL BOURBON
0.6548025355083453
W. L. WELLER 12-YEAR-OLD KENTUCKY STRAIGHT BOURBON
0.6545186877886962
W.L. WELLER SPECIAL RESERVE BOURBON
0.654487047829863
```

Figure 3: Results from Word Mover Distance for the Wild Turkey 101.

4.3 App Development

The app was developed as a Python Dash App. You can see an image of the App interface in Figure 4

4.3.1 App Data

To do this some of the data needed to be exported in the proper format. The following is a list of the datasets loaded by the app:

- whisky_tfidf.parquet : Basic whisky dataframe used to generate selector.
- whiskyinfo.parquet : Details about a whisky such as price, ABV, etc.
- similarities.parquet : Table showing every whisky's similarity to every other whisky.
- itemlinks.parquet : Table that includes an LCBO link for each whisky.
- reviewlist.parquet : Table that includes reddit review information and links for each whisky.

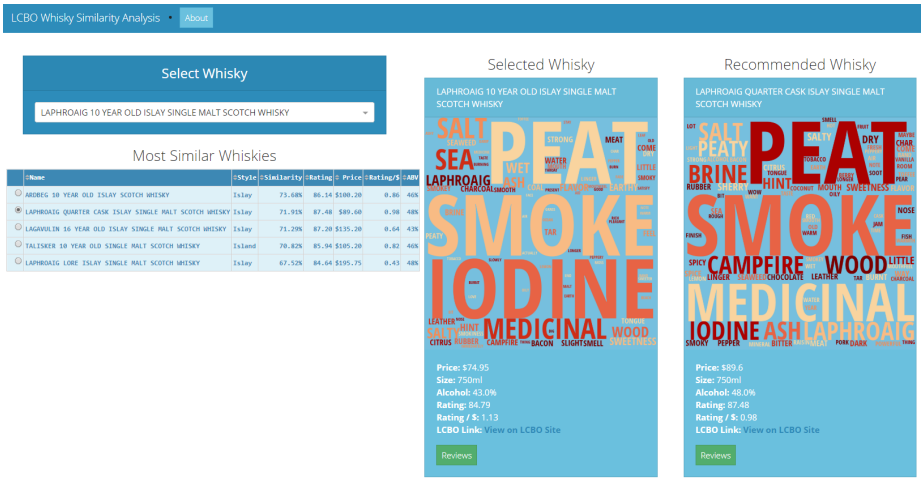

Figure 4: App interface showing suggestions for Laphroaig 10.

4.3.2 App Function

Within the App, there are several functions used to grab the required data to display to users. These are:

- getReviews : Given an itemnumber, returns all reddit review information to populate a table.
- getLCBOLink : Given an itemnumber returns a link to the LCBO website.
- show_top_similarities : Returns table data of top n similar whiskies to selected itemnumber.
- getwhiskydesc : Given an itemnumber, returns all info to display to the user in a markdown format.

5 Results

The results perform much better than I was expecting. Despite not guiding the algorithm at all by holding back regional styles, most suggestions fall within style boundaries, with exceptions being those whiskies that are outliers for their style, as would be expected. Figures 5 and 6 show examples of how Bourbons and Irish whiskies both recommend mostly the same category.

The app can be viewed at https://lcbo-whisky-similarity.herokuapp.com/.

	Name	Style	Similarity	Rating	Price	Rating/$	ABV
◉	WILD TURKEY 101 KENTUCKY STRAIGHT BOURBON	Bourbon	72.51%	82.78	$42.95	1.93	50%
○	FOUR ROSES SMALL BATCH BOURBON	Bourbon	70.05%	87.72	$48.20	1.82	45%
○	OLD FORESTER	Bourbon	69.41%	83.65	$33.20	2.52	43%
○	WELLER ANTIQUE 107 ORIGINAL WHEATED STRAIGHT BOURBON	Bourbon	68.59%	85.15	$35.95	2.37	54%
○	EVAN WILLIAMS SINGLE BARREL BOURBON	Bourbon	67.81%	82.36	$54.55	1.51	43%

Figure 5: Suggestions for Wild Turkey Rare Breed.

Figure 6: Suggestions for Redbreast 12.

6 Discussion

I am happy with the results and the recommendations given, as well as the app interface. Exploring the data through the app reveals a few issues with data cleaning that I would like to address, such as a couple of whiskies appearing twice, or some reviews being matched to other whiskies.

I would also like to improve on the process by implementing a database and having scripts run incrementally to capture new reviews and products.

References

[1] Chris Nicholson. A beginner's guide to word2vec and neural word embeddings, 2019. https://skymind.ai/wiki/word2vec, Last accessed on 2019-08-24.

[2] Edward Ma. Word distance between word embeddings, 2019. https://towardsdatascience.com/word-distance-between-word-embeddings-cc3e9cf1d632, Last accessed on 2019-08-24.

[3] Misc Reddit Users. Reddit whisky network review archive, 2019. https://docs.google.com/spreadsheets/d/1X1HTxkI6SqsdpNSkSSivMzpxNT-oeTbjFFDdEkXD30o, Last accessed on 2019-08-24.

List of Figures