An NLP-Based Scotch Whisky Recommender Engine

COMP3004 - Designing Intelligent Agents

Robert Soane

June 2021

Contents

1 Introduction

2 Scotch: Distilled

This project concerns itself with the specific domain of Scotch. To fully understand the problem at hand, a basic understanding of the drink is required. For this reason, in this section I present a brief overview of Scotch, and and a summary of its lexicon. From thereon in whisky-specific terms can be used without explanation.

2.1 A brief overview of Scotch

Distilled spirits have been in production as early as 1310 in France, with fruit wines being fortified by distillation to make brandy. The earliest records of Scotch whisky dating back to 1494. Where brandy is a distillation of fermented fruit juices (wines), whisky is produced by distilling fermented grains [1, 2].

Scotch whisky refers to whisky produced in Scotland fulfilling a set of legal requirements set by the UK Government [3][1]. Grains are allowed to malt (germinate) to develop their sugars, after which the sugars are extracted to produce a syrup called *wort.* The wort is fermented to produce a sweet hop-free beer. The beer is distilled in *pot stills* to increase the alcohol content significantly to produce *new-make spirit.* This is matured in oak casks for a minimum of 3 years, before the whisky is bottled. At this stage the distiller may choose to dilute the whisky to an ABV of no less than 40% [1, 2].

2.2 Whisky and words

When describing whisky, there two important distinctions to be made. Malt versus grain, and single versus blended. *Malt* refers to a whisky wherein the only grain used is *malted barley*, whereas *grain* whisky can be made with mixtures of any grains [4]. *Single* refers to whisky where all whisky in the bottle has come from the same distillery, whereas *blended* whisky is a blend from any number of distilleries. It is important to note that single whiskies are usually themselves blends, but blends of casks from the same distillery [5].

The flavours present in Scotch come from a number of places, and this in turn influences how various whisky's flavour profiles are described. A descriptor often used to describe whiskies is *peat.* To stop the malting process, the grain is heated, some distilleries (particularly those on the island of Islay) use a peat fire to carry this out. This imparts a smokey flavour onto the grain, which carries through to the end spirit. This smokey flavour is often described as *peated* [1, 6].

The maturation process provides another opportunity to add flavour to the drink. The requirement to age all Scotch in oak casks is resource intensive, and has lead to distilleries purchasing used casks from other drinks manufacturers. Traditionally the sherry industry has supplied used casks to distilleries, and more recently bourbon casks have been used. Any cask which has previously held any drink can be used, be it for the entire maturation process, or at the end such as a *sherry cask finish.* These all add their own flavours to the drink, and this is reflected in whisky tasting notes [1, 7].

3 Background and Literature

3.1 Language Models

In general, Natural Language Processing (NLP) tasks require a language model of some form or another. Artificial Intelligence (AI) based methods cannot process text in its native unstructured form, but need to convert the raw text to a structured form suitable for the computer to understand.

The two predominant model types are *syntactic* and *semantic* models. Syntactic methods transform text to a set of 'symbols' which carry no inherent meaning, but can be compared across instances in a dataset, wheras semantic methods (such as those described in subsubsection 3.1.3) retain a general understanding of the text [8].

A dominant syntactic method for transforming unstructured text into a computer-analysable form is the Bag-of-words (BoW) model. The dataset is tokenized (split into individual words), lemmatized (see subsubsection 3.1.1) and k keywords are extracted (see subsubsection 3.1.2) to form our bag of words $b \in \mathbb{R}^k$. Each document[2] is transformed to a vector $v \in \mathbb{R}^k$ such that v_i is the frequency of the word b_i occuring in the document [8, 9, 10].

[1]This report concerns Scotch whisky, and thus *Scotch* and *whisky* are used interchangeably.

[2]It is common to refer to an instance in a text dataset as a *document*

3.1.1 Stemming and Lemmatization

When dealing with text data, it is not uncommon to have multiple forms of the same word. A syntactic model would view the words 'cat' and 'cats' as two different discrete symbols. A method is needed to reduce words to a normal form.

Porter proposed an algorithm for removing word suffixes to aim for a normal form, this is called *stemming*. With no semantic understanding, the algorithm searches for specific suffix patterns and removes them until it is unable to [11].

A more semantic approach would be *lemmatization*. Instead of algorithmically removing word endings lemmatization aims to normalise words to a real word root, that is a lemmatizer would reduce the word to the dictionary form of the word [12]. A lemmatizer implementation in Python is the WordNetLemmatizer in the Python Natural Language Tool Kit (NLTK), which queries the WordNet corpus to find the root word [9, 13].

3.1.2 Keyword Extraction

For syntactic methods, keyword extraction is key. For the purposes of this report, a keyword is a word of particular relevance or importance, and from which we might extract useful information. Keyword extraction refers to strategies based on which those important words can be ranked, and only the most relevant kept.

TF-IDF Onesuch method, is Term Frequency Inverse Document Frequency (TF-IDF). This is commonly used with BoW, and is implemented in Scikit-Learn [14]. TF-IDF is a statistic for scoring a words importance based the word's how frequently it occurs in a document, and how frequently it occurs in the dataset [15].

Scoring as such aims to penalise words that occur too frequently across a document, boosting the scores of words in an individual document for which they with disproportionately high frequency.

Graph based keyword extraction

3.1.3 Word2vec

3.2 Recommender engines

3.2.1 Collaborative filters

3.2.2 Content Driven Recommendations

3.3 Machine learning applications to Whisky

4 Approach

5 Implementation

6 Results

7 Discussion

8 Conclusion

References

[1] K. Jacques, T. Lyons, and D. Kelsall, *The Alcohol Textbook 4th edition*, 4th ed. Nottingham University Press, 2003.

[2] M. Pyke, "THE MANUFACTURE OF SCOTCH GRAIN WHISKY," *The Distillers Company Ltd., Glenochil Research Station, Menstrie, Clackmannanshirc, Scotland)*, vol. 71, pp. 209–218, 1965.

[3] "The scotch whisky regulations 2009," *Legislation.gov.uk*, 2009. [Online]. Available: https://www.legislation.gov.uk/uksi/2009/2890/contents/made

[4] P. Valaer, "Scotch Whisky," *Industrial and Engineering Chemistry*, vol. 32, no. 7, pp. 935–943, 1940.

[5] B. C. Smith, C. Sester, J. Ballester, and O. Deroy, "The perceptual categorisation of blended and single malt Scotch whiskies," *Flavour*, vol. 6, no. 1, pp. 1–9, 2017.

[6] G. N. Bathgate, "The influence of malt and wort processing on spirit character: the lost styles of Scotch malt whisky," *Journal of the Institute of Brewing*, vol. 125, no. 2, pp. 200–213, 2019.

[7] J. Mosedale and J.-L. Puech, "Wood maturation of distilled beverages," *Trends in Food Science & Technology*, vol. 9, no. 3, pp. 95–101, mar 1998. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924224498000247

[8] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48–57, may 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6786458/

[9] E. L. Steven Bird, Ewan Klein, *Natural Language Processing with Python.* O'Reilly Media Inc, 2009.

[10] Y. Zhang, R. Jin, and Z. H. Zhou, "Understanding bag-of-words model: A statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

[11] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, mar 1980. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/eb046814/full/html

[12] K. Jayakodi, M. Bandara, I. Perera, and D. Meedeniya, "WordNet and cosine similarity based classifier of exam questions using bloom's taxonomy," *International Journal of Emerging Technologies in Learning*, vol. 11, no. 4, pp. 142–149, 2016.

[13] "What is wordnet?" 2010. [Online]. Available: https://wordnet.princeton.edu/

[14] P. Fabian, G. Varoquaux, A. Gramfort, M. Vincent, B. Thririon, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[15] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1, pp. 29–48, 2003.