Food Quality and Preference 83 (2020) 103926

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Fast and automated sensory analysis: Using natural language processing for descriptive lexicon development

Leah M. Hamilton, Jacob Lahne*

Department of Food Science and Technology, Virginia Polytechnic Institute and State University, 1230 Washington St SW, Blacksburg, VA 24061, USA

ARTICLE INFO

Keywords:
Natural language processing
Rapid descriptive methods
Big data
Whisky
Research methodology
Machine learning

ABSTRACT

As sensory evaluation relies upon humans accurately communicating their sensory experience, the diverse and overlapping vocabulary of flavor descriptors remains a major challenge. The lexicon generation protocols used in methods like Descriptive Analysis are expensive and time-consuming, while the post-facto analyses of natural vocabulary in "quick and dirty" methods like Free Choice or Flash Profiling require considerable subjective decision-making on the part of the analyst. A potential alternative for producing lexicons and analyzing the sensory attributes of products in nonstandardized text can be found in Natural Language Processing (NLP). NLP tools allow for the analysis of larger volumes of free text with fewer subjective decisions. This paper describes the steps necessary to automatically collect, clean, and analyze existing product descriptions from the web. As a case study, online reviews of international whiskies from two prominent websites (2309 reviews from WhiskyCast and 4289 reviews from WhiskyAdvocate) were collected, preprocessed to only retain potentially-descriptive nouns, adjectives, and verbs, and then the final term list was grouped into a flavor wheel using Correspondence Analysis and Agglomerative Hierarchical Clustering. The wheel is compared to an existing Scotch flavor wheel. The ease of collecting nonstandardized descriptions of products and the improved speed of automated methods can facilitate collection of descriptive sensory data for products where no lexicon exists. This has the potential to speed up and standardize many of the bottlenecks in rapid descriptive methods and facilitate the collection and use of very large datasets of product descriptions.

1. Introduction

1.1. Descriptions in sensory science

A key task for sensory scientists is the development of standardized, descriptive language for food products—descriptive "lexicons" (Drake & Civille, 2003; Lawless & Civille, 2013). Descriptive lexicons are necessary starting points for methods like Descriptive Analysis and are traditionally developed through repeated product evaluation in focus group settings (Heymann, King, & Hopfer, 2014), requiring a trained panel and many (sometimes hundreds) of man-hours spent tasting and discussing samples. While these methods produce high-quality results, the time and monetary costs of Descriptive Analysis have spurred the development of alternative approaches in recent years (Mónica Bécue-Bertaut, 2014; Valentin, Chollet, Lelièvre, & Abdi, 2012). Many of these alternative methods work to eliminate the need for a standardized lexicon entirely, replacing subject training with post-facto statistical alignment of vocabularies. These approaches produce useful results within product sets, but without a universal vocabulary, there is no simple way to analyze new products or compare results with similar products outside of the researched set. There is still an advantage to having a generalizable descriptive lexicon for frequently-analyzed product categories.

Recently, researchers have begun to develop an alternative approach to descriptive sensory research from another direction: generating descriptive lexicons from existing text data. For example, Bécue-Bertaut, Álvarez-Esteban, and Pagès (2008) used Correspondence Analysis to explore the descriptive-similarity structure of wines reviewed in the 2005 *El Mundo* gastronomic guide. Valente (2016) used a similar guide—the annual John Platter Wine Guide (2008–2014)—to develop a wine wheel—a visualized sensory lexicon (Noble et al., 1987)—for South African Chenin Blanc wines. Ickes et al. (2017) manually compiled website descriptions of rums and used qualitative techniques to develop a lexicon, which they then applied in a traditional Descriptive Analysis activity.

The innovation of these examples is the use of pre-existing sources of text data; they otherwise follow a typical strategy for the analysis of free-text data that is well-established in sensory evaluation (Symoneaux

* Corresponding author.
E-mail address: jlahne@vt.edu (J. Lahne).

https://doi.org/10.1016/j.foodqual.2020.103926
Received 17 October 2019; Received in revised form 14 February 2020; Accepted 29 February 2020
Available online 09 March 2020
0950-3293/

& Galmarini, 2014). The first step is the collection of a set of comments, descriptions, or other text-data about the product set of interest. Following collection, these data are processed to extract descriptive language (e.g., "sweet" is descriptive; "the" is not). The descriptors are then lemmatized (transformed to standardized, dictionary forms called "lemmas"), and finally synonymized (words which probably have the same meaning, like "tart" and "sour", are grouped). In order to reduce the data to manageable dimensionality, infrequent descriptors are removed, usually through a frequency criterion. These steps are the same as those used in traditional comment analysis (Lahne, Trubek, & Pelchat, 2014; Symoneaux, Galmarini, & Mehinagic, 2012; ten Kleij & Musters, 2003). As a multitude of different terms can refer to the same perception, it is difficult to disentangle the meanings of words in an unstandardized vocabulary (Symoneaux & Galmarini, 2014).

The drawback to applying traditional free-text analysis to existing ("big") text data is that the method does not scale well. A product set that consists of several hundred reviews (Mónica Bécue-Bertaut et al., 2008; Valente, 2016) or websites (Ickes, Lee, & Cadwallader, 2017) takes significant researcher time and effort to extract, process, and analyze. This is only a single order of magnitude larger than the average size of product sets for traditional lexicon generation (Lestringant, Delarue, & Heymann, 2019). With the advent of so-called "big data" analyses (Ahn, Ahnert, Bagrow, & Barabasi, 2011; Mikolov, Chen, Corrado, & Dean, 2013; Singh, Shukla, & Mishra, 2018), sensory scientists might reasonably set their sights higher. Datasets of product descriptions several orders of magnitude larger are readily available in the form of user-driven and professional websites dedicated to discussing food and consumer products (e.g., RateBeer.com, WineAdvocate.com). To tackle these datasets, the computational-linguistics field of Natural Language Processing (NLP) provides methodologies for extracting, processing, and analyzing free-text data to obtain descriptive lexicons from existing text data.

1.2. Natural language Processing

NLP operates on human language data to automate language-related tasks such as machine translation, ranking of search engine results, and speech recognition, among many applications (Bates, 1995). There are likewise many possible applications of NLP to sensory data. A sensory scientist might want to identify all the descriptors of a product category, assign probable liking scores to products based on their descriptions, find drivers of liking, or generate the description of an ideal product. The workflow required will depend on the project goals and the starting text (known as a corpus), but most projects will require preprocessing to create counts of relevant words and remove noise.

First, individual word units (i.e., tokens) are identified in the text by a process known as tokenization (Bird, Klein, & Loper, 2019b; Mullen, 2018). Next, morphologically related forms of words are combined, either by removing inflectional suffixes (e.g., "-ing", "-s") in a process called stemming or by converting each word to its lemma (Manning, Raghavan, & Schütze, 2008). The results of stemming and lemmatizing are improved when each word's part of speech (POS) is known, which can also be computationally determined (Martinez, 2012). The final task of synonymization is known as word-sense disambiguation in NLP. There are multiple kinds of tools to do this task, with older methods using large, expert-curated dictionaries and networks of synonymous word meanings (Fellbaum, 1999), and newer methods using machine learning to calculate a numeric representation of word meaning based on its usage (Faruqui et al., 2014; Stevenson & Agirre, 2018). Most areas of NLP have moved from historical use of dictionaries, rules, and algorithms towards machine learning (Bates, 1995).

1.3. Whisky as a case study

Whisky is one of the most important distilled beverages in the world, with annual global sales in the billions of dollars. Very broadly, whisky is a grain distillate that is aged in oak barrels (Miller, 2019). Different countries produce whiskies following different regulations and traditions. For example, in the United States, whiskey (spelled with 'e') is distilled primarily from unmalted corn and rye mashbills and aged in new, charred-oak barrels (Bryson, 2014). In Scotland, whisky is produced primarily from barley (malted or unmalted) mashbills—often kiln-dried with peat—and aged almost exclusively in pre-used-oak barrels (Jackson, 2017). In Ireland, Canada, and Japan (the other major whisky-producing countries), different production parameters apply. Therefore, even unflavored whisky products present a broad range of sensory profiles.

In contrast to other high-value, sensory-driven products like wine, coffee, or cheese (Heisserer & Chambers IV, 1993; Noble et al., 1987; Spencer et al., 2016), sensory science has not developed a comprehensive, descriptive lexicon for whisky. While a Scotch whisky lexicon was developed by (Lee, Paterson, Piggott, & Richardson, 2001), this lexicon only applies to Scotch whiskies and not to American, Canadian, Irish, or Japanese whiskies. Recent Descriptive Analysis research into American whiskey established several DA lexicons for specific American whiskeys but was not aimed at generalizability (Lahne et al., 2019; Phetxumphou et al., In press). Other types of whisky remain uncharacterized, which explains the lack of comprehensive studies linking whisky production parameters and flavor characteristics (Miller, 2019). However, the absence of validated lexicons has not inhibited professional whisky tasters and enthusiasts from producing descriptions of these products which are published both in print (Bryson, 2014; Jackson, 2017) and online (e.g., Whiskycast.com, WhiskyAdvocate.com).

While these whisky reviews and descriptions do not share the kind of standardized vocabulary that would be facilitated by the existence of a descriptive lexicon, these "communities of taste" do tend to develop broadly shared sensory languages (Hennion, 2015; Shapin, 2016; Teil & Hennion, 2004). Furthermore, these are "big data" sources: there are thousands or tens of thousands of whisky descriptions published online, covering thousands of whiskies. Therefore, by employing NLP methods, it should be possible to derive a descriptive lexicon for whisky as a broad product category from this large set of already published text descriptions without the need to employ a descriptive panel.

1.4. Paper structure

The goal of this report is to demonstrate a workflow (Fig. 1) in which Natural Language Processing methods are applied to a corpus of food-product descriptions with the goal of producing a workable descriptive lexicon. As a case study, we have selected a corpus of online whisky reviews, as the whisky category has some existing lexicons to which we can compare our results but lacks a comprehensive, generalizable lexicon. In the following sections, we describe the process of developing and applying NLP approaches to this dataset. Section 3 describes data collection, Section 4 describes the process of identifying individual, potentially-descriptive terms from full sentences, and Section 5 describes the process of grouping related terms to create a flavor wheel.

In each section, we discuss different NLP tools which could be used to solve the problem in question and explain why and how a particular tool was used. We explain how each tool used compares to more familiar tools for text analysis in sensory science. Finally, we discuss the advantages and disadvantages of our suggested NLP approach and discuss areas for improvement and future investigation. We include several appendices describing tools we mention in this paper and how to obtain and apply them, as well as details of our datasets.

2. Software

Most of the work in this paper was programmed in `R` v3.5.3, a statistical computing language (R Core Team, 2019). Some early data

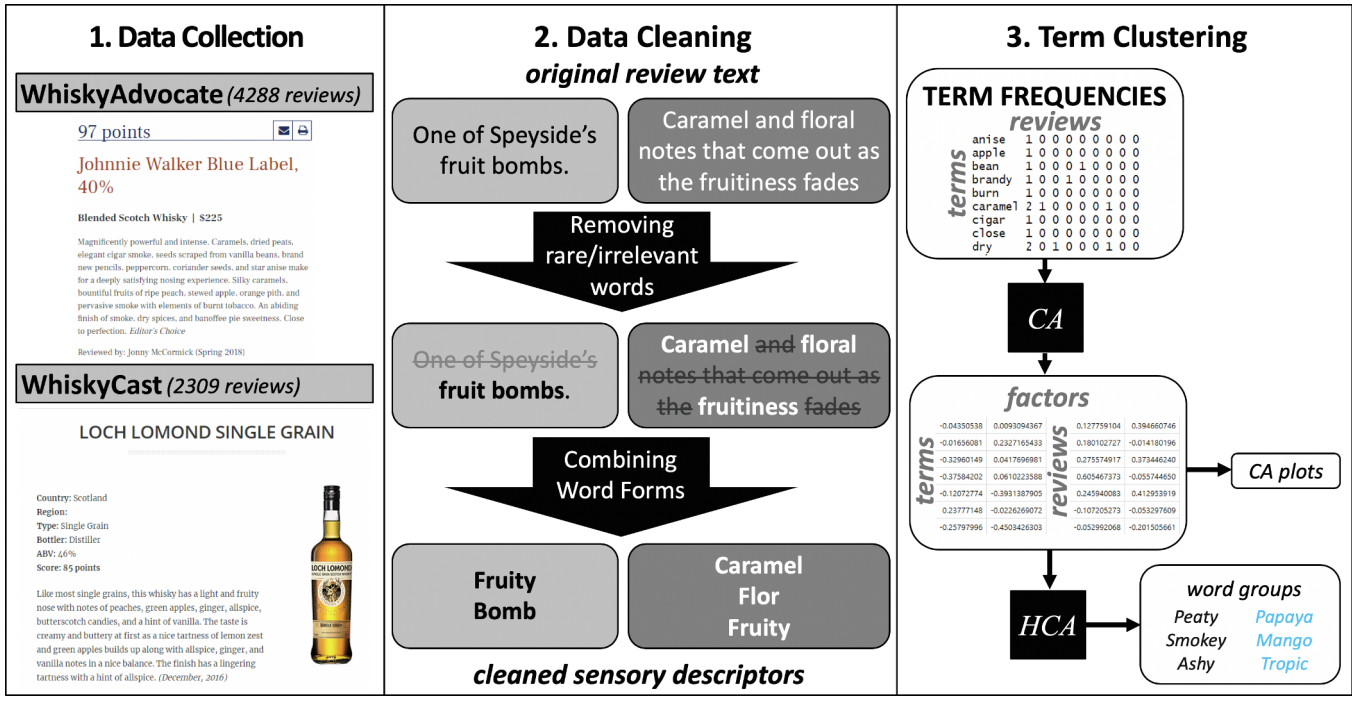

Fig. 1. A workflow diagram for the process outlined in this case study. First, 1) whisky reviews are collected from 2 prominent websites. This process is discussed in Section 2. Then 2) the irrelevant text is removed through a number of processes (e.g. named entity removal, POS-filtering) and redundant forms of the same words are reduced using lemmatization and stemming. This process is discussed in Section 3. Finally, the words are converted to vector representations in a lower-dimensional space and then clustered into groupings to form a flavor wheel. This process is discussed in Section 4.

collection code was written in the programming language `Python` 3.7.2, with the help of the packages `BeautifulSoup4`, `certifi`, `lxml`, and `urllib3`. Later data collection was done in `R` using the package `rvest`.

The R packages used for the final NLP workflow were `ca` (for correspondence analysis), `cleanNLP` (for neural network-based NLP), `hunspell` (for spell-checking), `tidytext` (for a base stop word list and manipulating the format of the text data), and `utf8` (for dealing with nonstandard characters). `cleanNLP` is an `R` framework for accessing several large NLP software packages originally written in other languages and is used here to access `SpaCy`.

`SpaCy` is a `Python` library which can do most standard NLP tasks such as tokenizing, POS-tagging, lemmatizing, and named entity recognition, primarily using pre-trained convolutional neural networks (Honnibal & Montani, 2018). `SpaCy` v2.1.8 is used for all of these tasks in this paper, with the pre-trained English model `en_core_web_sm` v2.1.0. This model was trained on OntoNotes 5.0, an annotated corpus containing telephone conversations, news group posts, news stories, blogs, and transcripts of conversational audio (Weischedel et al., 2013), for POS-tagging and named entity recognition.

All figures containing data were produced in `R`, and all data processing was done in `R`. Code is available on request from the corresponding author.

3. Text collection

Natural language text suitable for NLP can be collected in any number of ways. Traditionally, such qualitative data may be collected through free response questions on a survey or focus group transcripts, and existing text such as customer reviews of products could also be used. In this paper, professional product reviews were collected programmatically from the websites WhiskyAdvocate (WA; http://whiskyadvocate.com/) and WhiskyCast (WC; https://whiskycast.com/).

While this paper focuses on the collection of "big" data from existing web resources, many of the methods in Sections 4–5 are applicable to datasets of any size.

3.1. General workflow for website scraping

While it is possible to manually collect existing product descriptions from web sources (Ickes et al., 2017), the process is time-consuming and thus has limited utility. If it is possible to make a list of all the pages on a website with product information and those pages are formatted consistently (common in modern web design), then a single script can collect all product descriptions from a given website. This process of collecting data from a web page is called "scraping", and the software script is called a "scraper" or "bot" (as opposed to a human user). A scraper is written to collect a specific kind of content from a specific website, and must have a list of Universal Resource Locators (URLs, e.g., https://www.elsevier.com/) pointing to the pages with relevant information. Finding or collecting this list (which can itself be the product of a simpler scraper) is always the first step.

The scraper will then request each page in the list, which is returned as Hypertext Markup Language (HTML). HTML is a text-based specification which uses tags to describe the page layout. The layout varies between websites, so researchers wishing to write text scrapers will need to find the tags containing information relevant to their study. A short primer on HTML and content-scraping can be found in Appendix A, and more comprehensive resources for understanding HTML tags can be found in the beginner-friendly W3Schools web design tutorial (W3Schools, 2019) or the `BeautifulSoup4` documentation on extracting information from HTML trees (Richardson, 2018).

3.2. Case study – scraping whisky reviews

Our case study provides two examples of the website-specific ingenuity that is necessary for developing scrapers, as the corpus was collected from two different whisky review websites with different navigation and HTML structures. WA is a magazine website with a searchable "buying guide" comprised of product reviews written by its

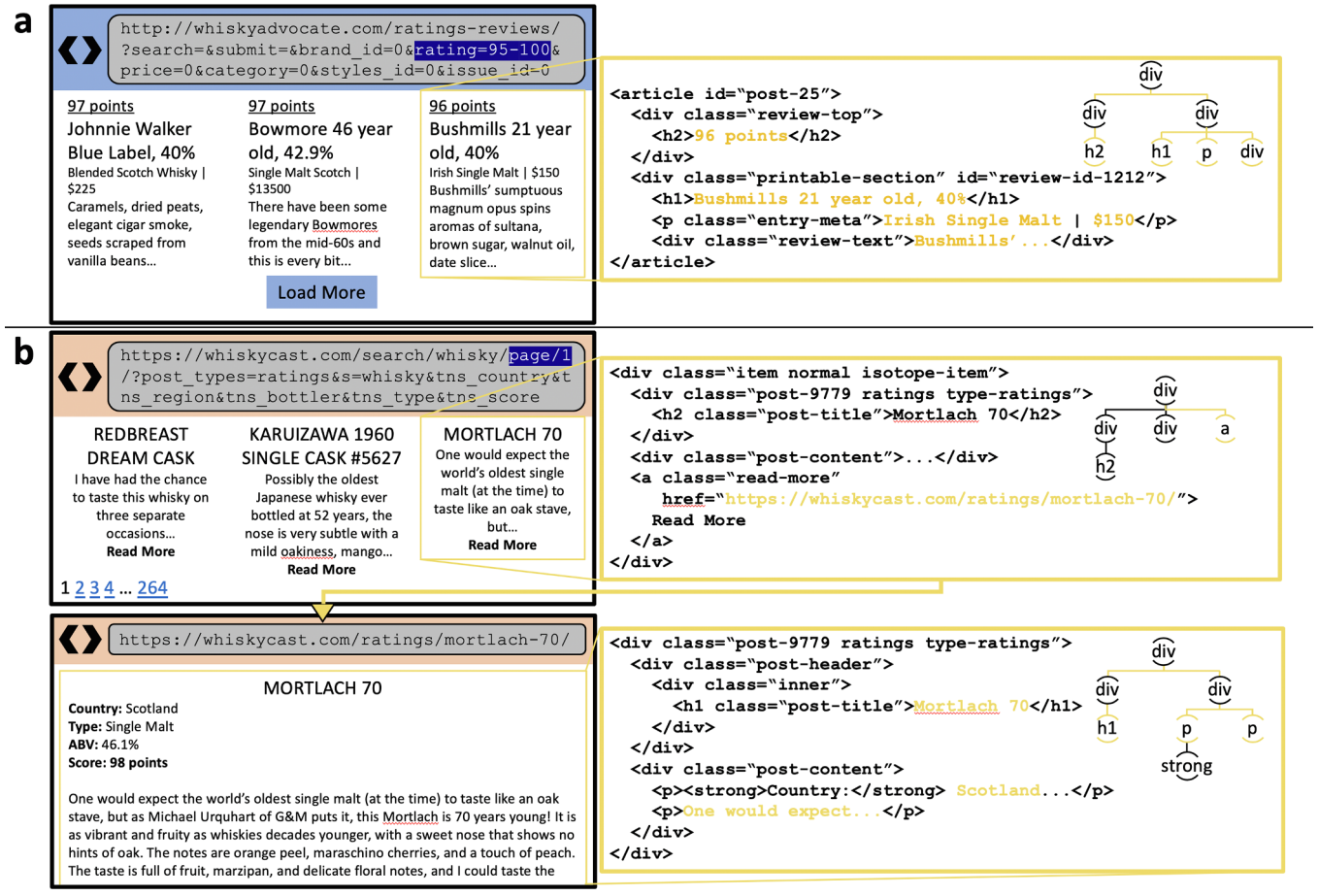

Fig. 2. A series of cartoons demonstrating the differences in the scraping workflow between the two websites scraped in this study: a) WhiskyAdvocate and b) WhiskyCast. In both diagrams, the blue-highlighted text was manipulated in the search page URLs to collect information about each whisky. Also in both diagrams, the text collected at each stage of the workflow is represented by a single product or page, with the data that was saved highlighted in yellow. In b, only URLs were collected from the search pages (highlighted in yellow), so the pages then had to be visited individually to collect product data.

professional tasting staff (Shanken, Lindenmuth, Schwenk, Barton, Simmons, & Kostro, 2018). WC is a podcast website that houses the "tasting notes" of its producer in a searchable database (Gillespie, 2018). The search page URLs of both websites were manipulated to access all of the reviews.

WA was easier to scrape, as its search page provides the option to filter by various product categories. For instance, reviews were categorized by bins of quality scores, with each product falling in one of five possible bins: 95–100, 90–94, 80–89, 70–79, and 60–69. These 5 values were inserted into the search query URL (Fig. 2a, highlighted in blue) to make a total of 5 query URLs. While each page only displays the first 6 results to the user on load, all results for a category are pre-loaded as hidden elements, so all WA reviews were collected from these 5 pages. Fig. 2a shows an example review with the extracted text shown in yellow. See Appendix A for more explanation of the HTML tags visible in Fig. 2.

The process for scraping WC was slightly more involved. The search page loads 10 products at a time and the user must navigate to the next page in order to see more results (Fig. 2b). We determined that the search strings "whisky" and "whiskey" both returned all reviews on the website. Thus, WC required the following multi-stage process. 1) The first page of results for the search string "whisky" was loaded. 2) The product review links on the page of results were added to a list (Fig. 2b, yellow text). 3) The number of pages of results (264 at the time of scraping) was also pulled from this page. 4) Steps 1 and 2 were repeated for each page number between 2 and 264. 5) Each link in the list was visited to scrape the review.

The product reviews from WA and WC comprise the corpus used in sections 4 and 5. Additionally, Wikipedia (https://en.wikipedia.org/wiki/) and Flaviar (https://flaviar.com/distilleries/) were used to compile distillery information and determine product country of origin, which is only used to characterize the corpus. The majority of the dataset (3384 reviews or 51.3%) is comprised of single malt whiskies, followed by bourbons (925 reviews or 14%). The smallest non-empty review is 6 words long, the longest is 400 words long, and the median review length is 75 words (example review in Fig. 3).

The oldest reviews from WA are from the Winter 1992 issue with the most recent from Fall 2018 (Shanken et al., 2018). The dataset contains all reviews on the website as of November 16th, 2018 and was collected using the `R` statistical environment (R Core Team, 2019) and the package `rvest`, based on code from Koki (2018). The 4288 reviews in this corpus are comprised of 2296 reviews of Scottish whiskies (53.5%), 912 of American (21.3%), 227 of Canadian (5.29%), 211 of Irish (4.9%), and 87 of Japanese (2.0%), with 555 (12.9%) of the reviews of whiskies from other countries or of unknown origin.

The oldest dated tasting notes from WC are from October 2009, and all notes housed on the website as of October 4th, 2018 were collected using `Python`, `BeautifulSoup4`, and `urllib3`. The 2309 reviews in this corpus cover 1282 Scotch (55.5%), 538 American (23.3%), 150 Irish (6.5%), 114 Canadian (4.9%), and 74 Japanese (3.2%) whiskies, with 151 (6.5%) from other countries.

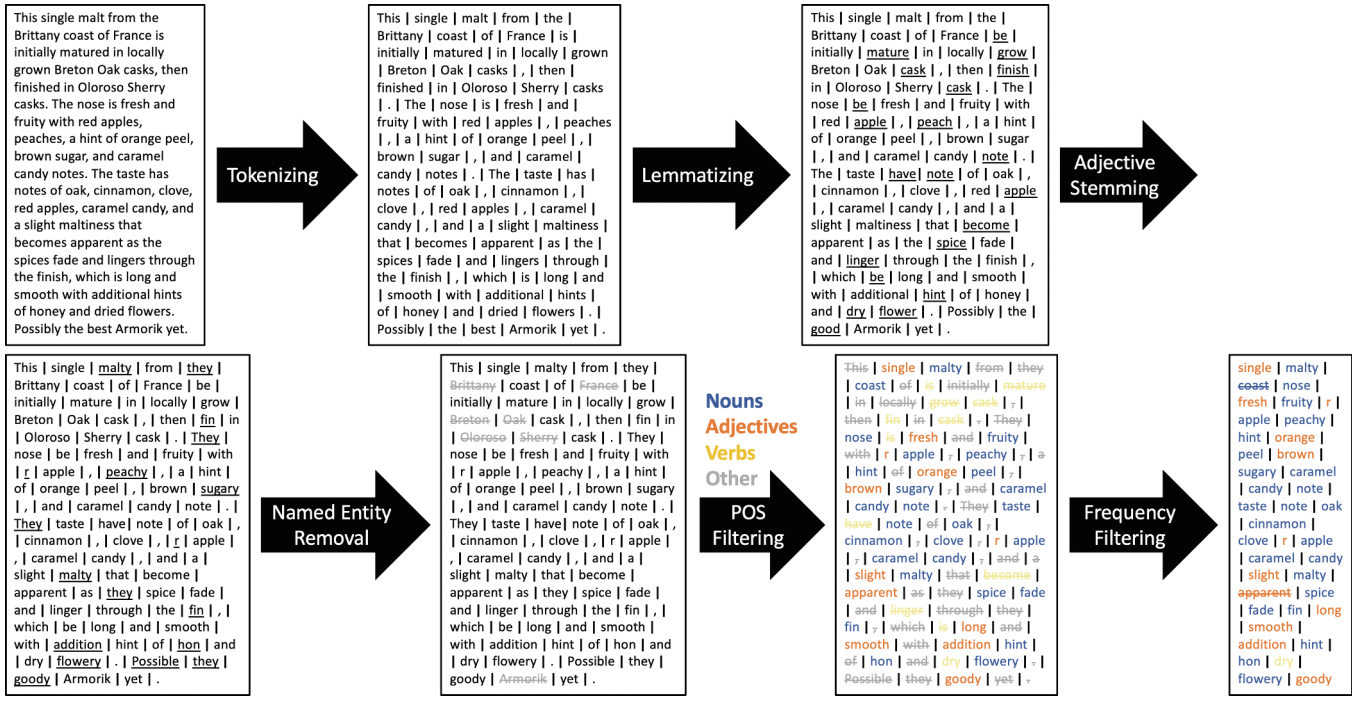

Fig. 3. A workflow diagram outlining the steps in the text cleaning workflow utilized in our case study, using WC's review of Armorik Double Maturation as an example. The removed tokens are representative of the actual automated output of each step rather than an idealized gold standard. Note that there are annotation errors included, such as oak and sherry being removed as named entities or "casks" being tagged as a verb.

3.3. Advantages and limitations

Scrapers provide rapid, low- or no-cost data collection and large sample sizes at the expense of precise experimental controls. The whiskies were not tasted blind or in controlled environmental conditions—in fact, each review may comprise an amalgamated impression of several tastings from the same reviewer. We do not have repeated evaluations of samples by multiple independent tasters (although consumer review websites may). This large amount of data also contains a large amount of irrelevant text, which requires similar processing to any free text data (but at a larger scale). One must carefully consider the representativeness of the data source. In the current case, scotch single malt whiskies are over-represented, but there are more whiskies of all types than would be practical with traditional lexicon development methods. In exchange for decreased experimental control, we have descriptions of more than 2000 products spanning the globe and the past 20 + years. An overview of bias and data quality in big data can be found in Bruce and Bruce (2017).

The scraping methods outlined here are not universal, and the differences between WA and WC demonstrate that there is no one-size-fits-all solution. In particular, the need for a list of information-containing pages will limit the number of websites that can easily be scraped. For example, BeerAdvocate (https://www.beeradvocate.com/), a large consumer beer review website that was an early target for web scraping, will only show a limited number of results for search queries. Chaulagain, Pandey, Basnet, and Shakya (2017) cover some of the available tools for scraping websites where URL manipulation and HTML navigation are not sufficient.

Finally, there are important legal, ethical, and courtesy concerns regarding data scraping. Website terms of service may prohibit certain uses of the website content, especially commercial ones. Automated requests for a large number of web pages can overload website servers and cause issues for other users, so some websites may block page requests from bots or repeated requests in a short timeframe. It is best practice to limit the rate of page requests, usually to a maximum of one per second (still much faster than traditional data collection in sensory science!), though some websites may have more specific instructions for bots. While there is currently very little specific law or legal precedent about web scraping and big data, a recent review of the legal and ethical concerns in web scraping can be found in Gold and Latonero (2018), while a more pragmatic view of web scraping actions likely to have short-term repercussions has been written by Lawson (2015).

4. Text processing and cleaning

Natural language text needs to be cleaned and processed before analysis, although the specific steps will vary based on the input text and the desired output. For sensory science, the desired output is generally either a minimal list of relevant descriptors, as in a lexicon development workflow (Ickes et al., 2017; Valente, 2016), and/or occurrence counts for relevant descriptors, as in comment analysis of solicited text (Monica Bécue-Bertaut & Lê, 2011; Lahne et al., 2014; Symoneaux et al., 2012; Symoneaux & Galmarini, 2014). The cleaning steps in an automated text analysis workflow are very similar to the manual process outlined in papers like (Monica Bécue-Bertaut & Lê, 2011; Symoneaux et al., 2012). First, the text is split into individual words or relevant phrases (i.e., tokenized), then the words are converted into standard forms to reduce the number of unique tokens (i.e., stemmed or lemmatized), and then various methods are used to eliminate irrelevant or non-descriptive words. These steps will be necessary for most automated analysis of sensorial text, whatever the desired output.

4.1. A general workflow for automated text cleaning

Computers store text as a series of characters, mainly comprised of letters, numbers, whitespace (e.g. spaces and tabs), and punctuation. This text may have some issues that need addressing before beginning NLP, but this can be minimized with a carefully written scraper, and a discussion of tools to strip special characters or irrelevant page elements is outside the scope of this paper.

4.1.1. Identifying word units – tokenizing

To begin NLP, individual word units (tokens) must first be identified from the continuous string of characters in a process called tokenization. For English, this process is trivial since word units are broken up by whitespace and punctuation. Less trivially, one could choose to treat multi-word units at tokens, such as only tokenizing at specific words or punctuation (more practical in descriptive text written according to specific formatting instructions) or treating every pair of words as a token rather than every individual word (called bigram tokenization, extended to larger sets in ngram tokenization). These alternative methods will not be covered in depth here, and the interested reader can turn to (Bird et al., 2019b) for a general overview of tokenization and (Mullen, 2018) for a practical overview of tokenizing with `R`.

4.1.2. Reducing redudant word forms – stemming and lemmatizing

After tokenizing, it is often worthwhile to combine near-duplicate tokens, such as pairs of word forms like “fruity” and “fruit” or “drying” and “dry”. This can be done with a stemmer, which looks for pre-determined letter combinations such as “-y”, “-ing”, or “-tion” at the ends of words and removes them, leaving behind a word stem which may or may not be a complete word. Lemmatization is an alternative to stemming, in which each word is converted to its base English form, usually with a dictionary or POS-dependent rules. Dictionary lookup lemmatizers are easy to implement but no longer commonly used, as they are slow and even a very large dictionary such as WordNet (Fellbaum, 1999) may not contain every word in your text (e.g. “bready” is missing from WordNet). Another approach, variously called rules-based, algorithmic, or heuristic lemmatization, is similar to stemming in that it applies some set of rules to the word endings but returns a lemma rather than a stem. Good rules-based approaches are fast and capable of handling out-of-vocabulary (OOV) words. Algorithms can also be combined with a table of irregular exceptions, to minimize the table size and lookup cost while maintaining OOV flexibility. Other approaches such as neural networks are rarely necessary for English text. An overview of stemming can be found by (Manning et al., 2008) and an overview on lemmatizing with emphasis on rules-based lemmatization can be found by (Romero, 2019).

The results of stemming or lemmatizing can be improved with information on the POS of each word, due to the presence of homonyms. An example relevant to sensory descriptions is “clove”, which could lemmatize to “cleave” if treated as a verb or remain “clove” if treated as a singular noun. POS-tagging requires information on the surrounding words or sentence in order to disambiguate between the possible POS tags for a word, and automated taggers typically use statistical modeling or a trained machine learning model. For example, `SpaCy` uses a type of machine-learning algorithm called a convolutional neural network (CNN) for this purpose, which tags each individual token-in-context with a POS, using knowledge of the word and surrounding words. This means that in the phrase “will dry out your throat”, “dry” can be tagged as a verb, while in “dry and sandy”, “dry” is tagged as a noun.

4.1.3. Filtering for relevant terms

A reasonably sized dataset such as that in our example will have thousands to hundreds of thousands of unique words after stemming and lemmatizing. Many application-specific methods exist for identifying important words in a dataset, the most basic of which are based around frequency of occurrence. Words which occur only once or a few times in a large corpus are likely to be irrelevant noise, while words which occur in every document (i.e., are used to describe every product) are not likely to differentiate between documents (or products). This problem parallels the language-selection problem in a DA panel, wherein terms that are used by only a few panelists or which do not differentiate between products are removed to limit the lexicon size. Sensory scientists using manual text analysis have previously used a variety of methods to select important descriptors (Monica Bécue-Bertaut & Lê, 2011; Ickes et al., 2017; Lahne et al., 2014; ten Kleij & Musters, 2003), and any of these criteria could be applied to an NLP workflow.

Other common filtering steps involve removing any words in a stoplist or removing specific tokens based on annotation (e.g., removing a specific POS). Generic stoplists, or stop-word lists, are included with most NLP tools, and contain words that are common and (thought to be) irrelevant to the meaning of the corpus (Bird, Klein, & Loper, 2019a; Flood, 1999). These lists typically contain pronouns (e.g., he, it, I), conjunctions (e.g., and, or), and other words which usually convey very little meaning. Specific applications can benefit from custom-made stoplists, or the knowledge that certain POS may be irrelevant: for example, prepositions are unlikely to be relevant sensory descriptors. These are not the only possible filtering methods, however, and a different cleaning workflow may be necessary for different text sources such as the low density of descriptive terms found by Valente (2016) or the higher level of typos and other noise present in social media data (Singh et al., 2018).

4.2. Case study – extracting descriptors from whisky reviews

`SpaCy` was chosen as our primary NLP tool because of its ability to easily tokenize, tag POS, find named entities, and lemmatize in one step. `CoreNLP` could easily have been used instead, but `SpaCy` was chosen as it is faster and `Python`-based (rather than `Java`-based). Fig. 3 shows an overview of all NLP steps described in Sections 4.2.1-3, by way of an example review.

4.2.1. Tokenizing

The text was tokenized using `SpaCy`’s algorithmic (i.e., rule-based) tokenizer (Fig. 3), the rules for which are detailed in the `SpaCy` manual (Honnibal & Montani, 2018).

4.2.2. Stemming and lemmatizing

Descriptive language about food has unique features (Monica Bécue-Bertaut & Lê, 2011), including the use of nouns and adjectives as descriptive terms (e.g., a wine may be described as both “vibrant” [adjective] and “blackberry” [noun]). To process the whisky reviews in this dataset it was necessary to find ways to group adjective and noun forms of the same word (e.g., malt and maltiness, Fig. 3). Since no lemmatizer we tested combined adjective and noun forms, we first used `SpaCy`’s lemmatizer and then ran the results through a custom adjective stemmer written with regular expressions.

The `SpaCy` lemmatizer uses a set of 21 POS-specific rules. The POS tags and lemmas for an example review can be seen in Fig. 3. Note that different instances of the same word could receive different POS tags (e.g., “finished” as a verb and “finish” as a noun in Fig. 3).

The adjective stemmer removed adjective endings (e.g., “-ful”, “-y”) from all tokens (full list in Appendix B). As this often resulted in stems which were not complete words, the stemmer would then add an “-e” to the end of any token not recognized by the `hunspell` dictionary. “Smoky”, which was stemmed to “smok”, was converted back into “smoke” by this step, while red, which was stemmed to “r”, was not converted to “re”, as neither form was in the dictionary. A “-y” ending was then added to every token if doing so created a dictionary word, so “smoke” was converted to “smokey”. This improved the number of readable stems, and maximized the number in the adjective form. As can be seen in Fig. 3, this process is not perfect—some stems such as “hon” for “honey” or “goody” for “good” are not immediately understandable—and this may or may not be a problem depending on the intended use of the output.

4.2.3. Term filtering

At this point, there were 11,569 unique word roots in the dataset, many nondescriptive. `SpaCy` was used to find named entities (for our dataset, largely numbers and proper nouns), which were removed,

reducing the dataset to 9004 unique roots. The POS annotations were also used for filtering. As in Bécue-Bertaut and Lê (2011), we retained nouns and adjectives, but we also kept any instances of the lemmas in Appendix B tagged as verbs. This resulted in 6190 unique tokens. The short whitelist of verbs was collected by manual investigation of the most common verbs. While some of the lemmas on this list are participle forms of verbs like "drying" or "charred", many are nouns tagged incorrectly as verbs, indicating that there is potential for improvement in POS taggers for sensory text in the future ("casks" in Fig. 3 is an example of this which was not added to the whitelist). For this paper, no stop word list was used, because most stop words were removed during POS-filtering.

Finally, a frequency cutoff was imposed: terms were removed which appeared in < 1% of documents (65 or fewer reviews) to reduce the number of unique tokens from 6190 to 407. Some fraction of the 5783 words which were removed by frequency and POS filtering were uncommon descriptors such as "lactic", "petrichor", and "turnip", but most were irrelevant words such as "legitimate" and "overprice" or versions of words with misspellings or unusual punctuation. The handling of the remaining irrelevant nouns, adjectives, and verbs is discussed in section 5.2.

Some less-common categories of whisky in the data set, such as whiskies from certain countries or in less renowned product categories, comprise less than 1% of the reviews and their characteristic descriptors may have been filtered out. In the future, a more nuanced filter could be used here to generate a more representative lexicon, such as keeping words with at least a certain prevalence in reviews for a known category (e.g., Japanese, rye), or balanced groups could be made with sub-sampling or over-sampling.

4.3. Advantages and limitations

The final workflow described in this paper relies heavily on a pre-trained multifunction neural network (`SpaCy`), which is common in modern applications of NLP. While heuristic, algorithmic, rules-based, or dictionary-based methods exist for many of the steps which are less computationally intensive, they are less accurate. For example, neural network-based POS taggers benefit from sentence context. The other main advantage of multifunction language models is that many steps can be done at once and can thus complement each other. Adding a new step to a `SpaCy` workflow (e.g., identifying place names, extracting different parts of speech, finding intensity modifiers, dependency parsing) can be accomplished with minimal coding.

`SpaCy` advertises itself as "Industrial-Strength" NLP that is easy-to-use for practical applications rather than being oriented towards researchers, and as such is still undergoing major updates (Honnibal & Montani, 2018). This means that the algorithms are ever evolving and code written with `SpaCy` can remain relatively current by updating the underlying language model. However, results are not reproducible across versions, sometimes significantly so. Slight differences in the annotations and lemmas being produced by a newer model lead to differences in the resulting occurrence counts which create ripple effects through the dimension reduction, clustering, and scree plots, eventually creating different categories in the final flavor wheel.

Unfortunately, most existing NLP tools were not trained on sensory descriptions and can struggle to accurately tag such descriptions. Words like "clove" and "peppery", while common in sensory science, may not be recognized by language models or present in dictionaries. Most lemmatizers are programmed with separate base lemmas for nouns and adjectives (e.g. "pepper" and "peppery"). Finally, while POS-filtering removes many non-descriptors, there is no existing tool which can specifically tag words for sensory relevance. Some of these are problems that are likely easily fixed in the future with transfer learning on a sensory-specific dataset, while others (like identifying words of sensory relevance) are larger undertakings.

5. Term grouping

In order to measure any specific sensory attribute of a product, a methodology must have some way of accounting for humans' tendency to use idiosyncratic expressions and synonyms. In Descriptive Analysis, there is a lengthy training process wherein the assessors generate and agree upon a single vocabulary to use for the sensations caused by the product (Heymann et al., 2014). Methods such as CATA don't require training but require the researchers to provide a list of possible distinct product attributes that is interpretable by the study participants (Jaeger et al., 2015; Meyners & Castura, 2014, p.). A growing body of methods instead has participants describe products in their own words which then have to be grouped *post facto* by either human coders (Lahne et al., 2014) or statistical methods (Kostov, Bécue-Bertaut, & Husson, 2014; Perrin et al., 2008).

The same problem exists when using existing descriptive text such as product reviews, but in most cases the scale of the dataset is beyond what is feasible to manually group. Thus, it is desirable to have some way of automatically grouping the data into categories of synonymous, related, or at least correlated words.

5.1. Word sense disambiguation and synonymy in NLP

The oldest method of automatically determining word meaning and synonymy was use a manually-curated dictionary which contained links between synonyms, namely WordNet (Fellbaum, 1999). However, one can easily see that WordNet's "synonyms" are not based on sensory perception for flavor terms, as it defines "tart" and "lemony" as synonyms while a pair such as "peaty" and "smoky" are not related at all.

Many NLP applications have found that a general dictionary-based approach is not sufficient for domain-specific text, so there are many other methods of grouping related words. Many modern approaches are based on trained word vectors (called embeddings). Similar to the embeddings underlying the CNNs discussed in section 4.1, word vectors trained on large enough corpora can represent word meaning in a high-dimensional space. A spatial clustering algorithm can then be used to acquire sets of synonyms (Stevenson & Agirre, 2018). Unfortunately, this requires a very large corpus in the domain of interest (typically more than 1 billion words), and unlike manually-curated databases, it cannot easily disambiguate between multiple possible meanings or uses of a word (Mikolov et al., 2013).

Neural networks are not the only way to acquire word vectors which are based on word meaning or at least word co-occurrence in the corpus. By treating each document (or review, or product) as an observation and each word as a variable, a matrix of per-document word counts can be constructed. Any method of dimensionality reduction that is appropriate for count data can be used to get word vectors for clustering. In computational linguistics, a method called Latent Semantic Analysis (LSA) is the most common Singular Value Decomposition (SVD)-based approach. LSA is a Principal Components Analysis (PCA) conducted on the covariance matrix of non-normalized data, and is historically preferred (to, e.g., PCA on the correlation matrix) for reasons of computation time (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Meanwhile, some statisticians have used methods such as Correspondence Analysis (CA), an SVD method which determines optimal continuous scales for categorical data, to calculate word and document vectors (Mónica Bécue-Bertaut et al., 2008; Greenacre, 2017). Finally, other methods such as Latent Dirchlet Allocation (LDA) have been used with great success, but these require *a priori* knowledge of data structure such as the number of groups/synonyms to be made, and are not deterministic (Blei, Ng, & Jordan, 2003).

All of these methods can effectively reduce the dimensionality of the dataset to produce word vector representations, but none provide a single gold standard method for determining the number of dimensions to keep. This is an outstanding problem of unsupervised learning in

statistics and data science, too large to cover here, so the interested reader is referred to (James, Witten, Hastie, & Tibshirani, 2013). A very common method in many fields including sensory science is the scree plot, in which the amount or percentage of variation explained by each dimension in an SVD analysis are plotted in descending order and visual inspection is used to determine the optimal cutoff where the plot starts to "flatten" (James et al., 2013).

Finally, the semantic word vectors must be grouped with synonymous or related words, so some sort of clustering is required. K-means is popular for this purpose in the computer science literature, namely for its computational efficiency (Xu & Tian, 2015), but it is also non-deterministic and the number of clusters must be set by the analyst *a priori*. Agglomerative hierarchical clustering analysis (HCA) is not computationally efficient for large data sets due to the large number of pairwise comparisons that must be made, but it does allow for scree plots and other semi-systematic methods to determine the number of clusters, and the solution for every possible number of clusters is found in a single analysis (James et al., 2013).

5.2. Case study – grouping related whisky descriptors

The workflow in section 4.2 resulted in a large table of descriptor counts present in each review. The list of terms was relatively clean, but still contained an unwieldy number of terms (n = 407) with many duplicates present. A clustering algorithm which could group together related descriptive terms was needed.

First, CA was used to reduce dimensionality of the data. Fig. 4 shows the first two axes of this analysis on all 407 terms, with visualization of terms whose summed contributions to the inertia of the first two axes was in the top 20th percentile (n = 82). The 5 production countries and 5 types of whisky most common in the data are also plotted as supplemental points at the barycenter of the reviews belonging to those categories (Greenacre, 2017).

It is already clear that the most-contributing terms are related to production variables such as mashbill or oak-aging. Thus, for processes like peating which impart a single characteristic note ("smokey" or "peaty"), the associated words are all synonyms. For moonshine and craft whiskies, which are less aged, the mashbill has more of an impact on flavor and those terms (e.g. "rye", "wheat", "grainy") are highly associated despite referring to different flavors (or the production variable itself, in the case of "mashbill"). It is likely that some of these terms would become more distinct in a larger data set (especially with more products in the less-represented categories). The mostly barrel-aged rye whiskies are close to moonshines in this plot, which may indicate that they differ from the majority Scotch single malt dataset in similar ways, rather than being indistinguishable from each other.

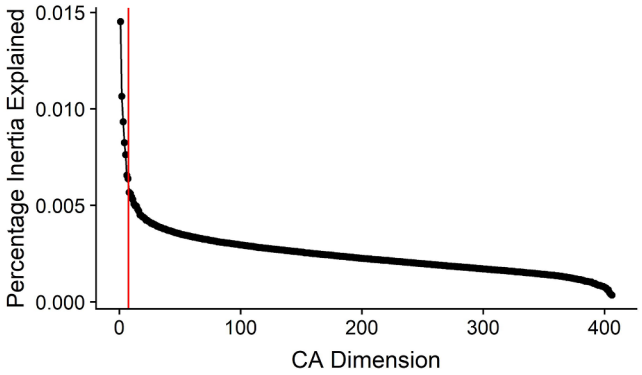

Fig. 5. A scree plot showing the percentage of total inertia explained by each factor in the CA solution. The red line shows the cutoff point, and dimensions to the left were kept (n = 7) while the rest were discarded.

A scree plot of the variation explained by each dimension can be seen in Fig. 5, and k = 7 dimensions were kept for the sake of term clustering. This cutoff is not at the "elbow" of the plot but instead after the last large drop in the scree plot. These decisions are very hard to make with large-dimensional matrices, as there is often no single clear cutoff. Appendix C demonstrates and briefly discusses the difficulty of making such decisions with very high-dimensional data and the impact on later steps, by way of investigating the dendrograms resulting from

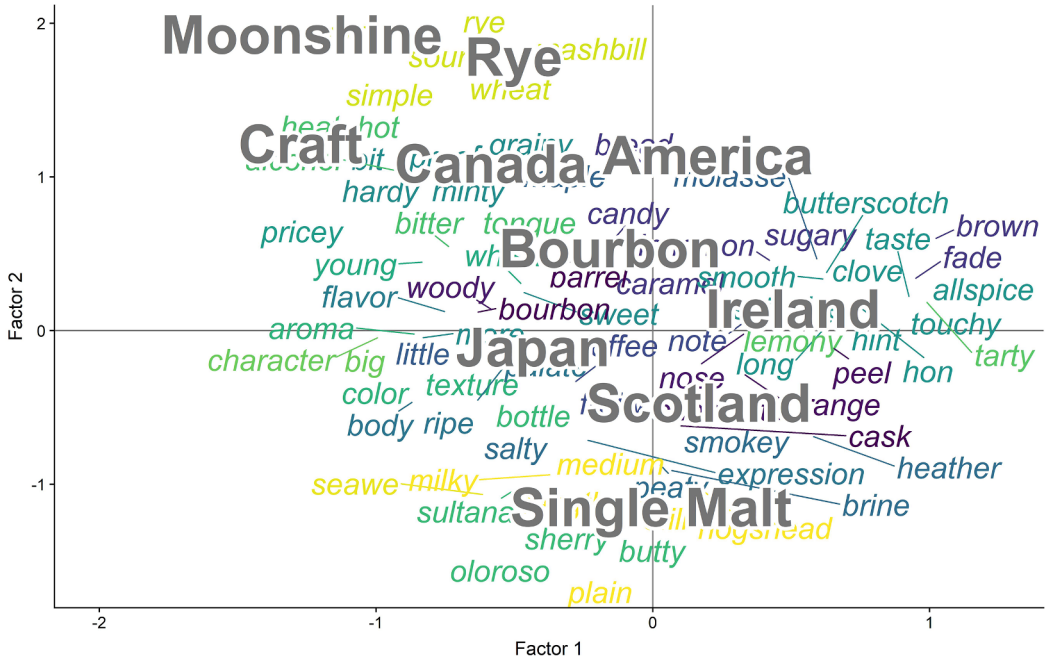

Fig. 4. A 2D CA biplot is shown based on the term frequencies of the 407 most common nouns, adjectives, and verbs in the reviews. The terms shown are only those whose summed contributions to the inertia of the first two axes was in the top 20th percentile (82 terms), and are plotted at principal coordinates. The 5 production countries and 5 types of whisky most commonly represented in the data are also plotted as supplemental points at the barycenter of the reviews (in standard coordinates) belonging to those categories.

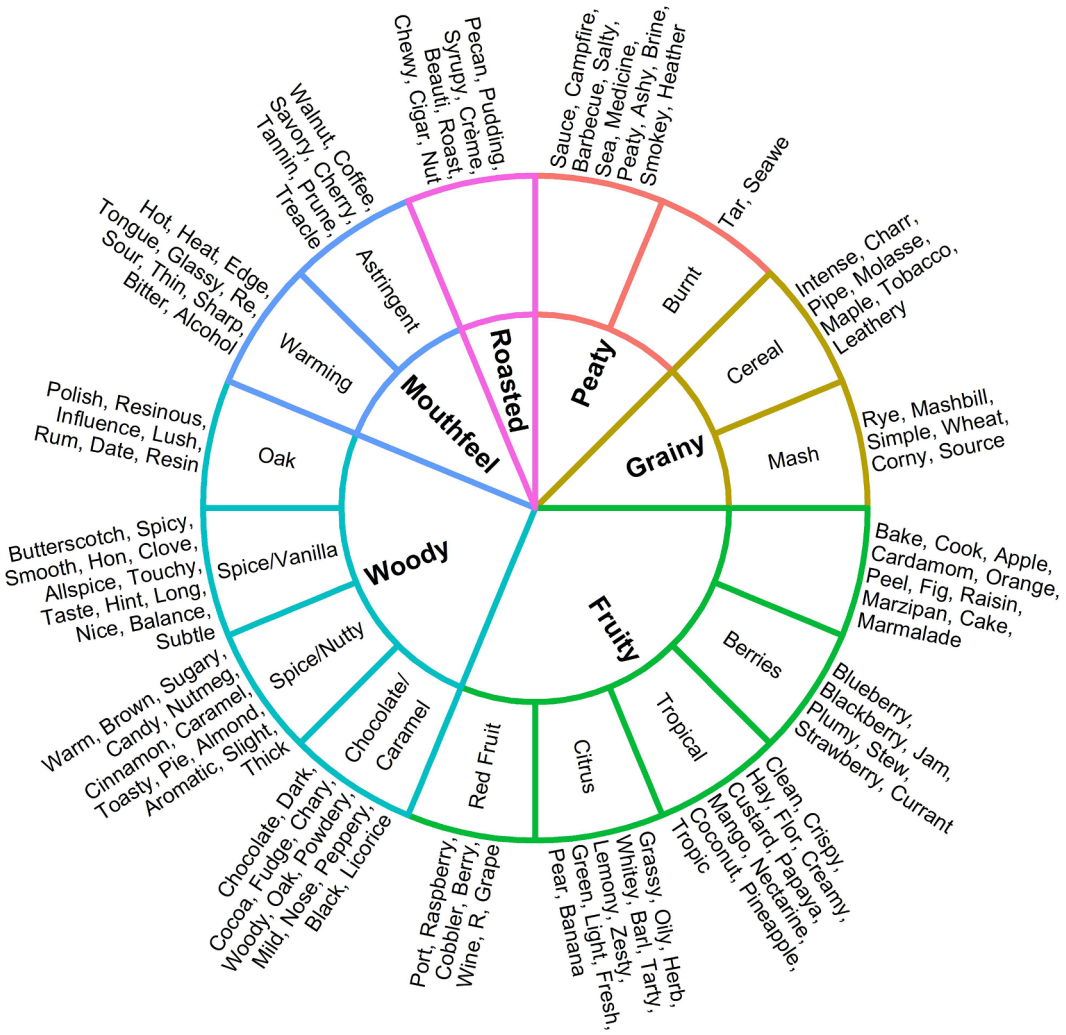

Fig. 6. In (a), the descriptive terms programmatically acquired from online reviews of international whiskies and grouped based on co-occurrence comprise the outside of a flavor wheel. This can be compared to (b), which reproduces the terms and organization from the most prominent published whisky flavor wheel (Lee et al., 2001), with the exception of intermediate-level category labels for the woody attributes. Category labels for (a) were manually assigned and match those used in (b) where possible.

four different values of *k*.

HCA using Ward's clustering criterion (within-cluster variance minimization) was conducted on the 7 × 407 matrix of term factor scores resulting from CA. Another scree plot was created of the total within-cluster sum of squares for each possible number of clusters 1 through 407 (Kassambara, 2017). Based on the elbow of this plot, it was determined that 32 clusters should be kept. Of these categories, 16 were comprised mostly or exclusively of production-related terms or non-descriptors, such as the cluster which contained the words "texture", "aroma", "amber", "color", "gold", and "exotic". While some descriptors remained in these clusters, HCA was able to function as an automatic term selection mechanism with much less work than curating a manual stopword list. The remaining 16 categories are represented as a flavor wheel in Fig. 6a, with the terms comprising each cluster displayed around the edge of the wheel (only the 13 most frequent are shown in the case of large clusters). The inner wheel categories, where possible, are labeled with terms drawn from the Lee et al. (2001) wheel (reproduced in Fig. 6b) for the sake of comparison.

Most of the categories in our wheel contain related or even synonymous terms, with some outliers (e.g. "flor" [floral] in the tropical fruits group, "sour" in the warming group). There are several large clusters of non-synonymous terms which are related through barrel-aging and are similar to the subgroup called "woody-extractive" by Lee et al. (2001)—thus, the alignment between these two wheels is helped by the fact that the 2001 wheel was arranged with more of an emphasis on origin of the flavor note. This kind of organization is likely to emerge from co-occurrence-based word clusters, and—as in the 2001 Scotch wheel—can be useful for producers trying to alter their process to obtain a new flavor profile.

The 2001 Scotch wheel, produced through a literature review and based in a large body of whisky expertise (Lee et al., 2001), shares many categories with the algorithmically-generated term categories produced in this paper (i.e., fruity, woody, grassy, grainy, and peaty). This is in part because of the predominance of Scotch in our sample set of reviews (3578, 54.2% of reviews), but many of these terms are not unique to Scotch.

The category called "roasted" in our wheel contains terms from multiple different categories on the 2001 wheel (namely grainy, woody-extractive, and peaty) which we would argue are tied together by sensory similarity rather than shared origin, making this a category unique to our new wheel. There was also a category seemingly related to the presence of tannins, which is labeled "Mouthfeel/Astringent" in Fig. 6a as this was the closest analog in the 2001 Scotch wheel. It seems likely that both of these categories of terms exist due to the representation of American whiskey descriptions in our dataset (the second-most represented country with 1450 or 22% of the reviews),

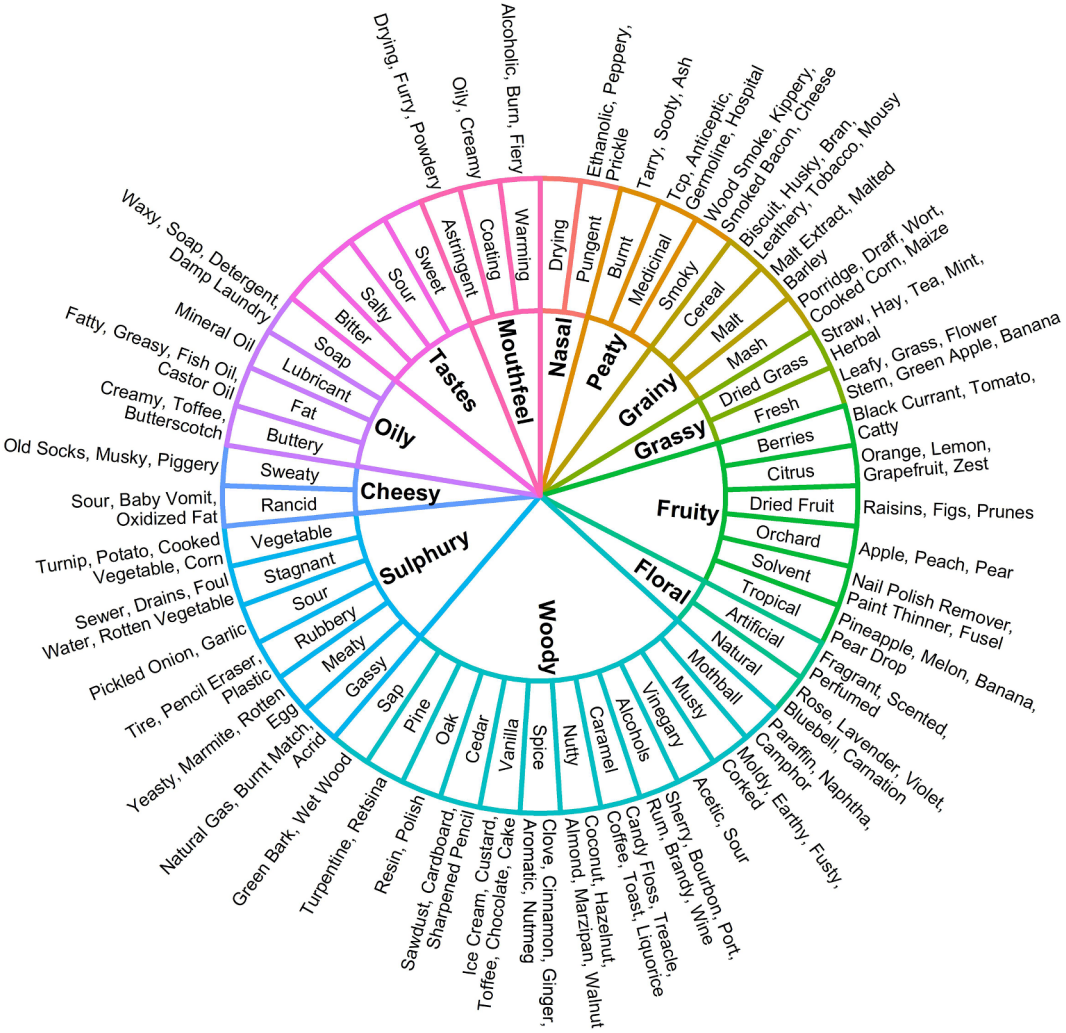

Fig. 6. (*continued*)

which are aged in charred new oak barrels. Furthermore, we suspect that the increased descriptor variance from the heterogeneous products in the sample set probably leads to a less fine-grained wheel than the existing Scotch wheel, which could be addressed through developing more category-specific wheels from our new dataset. This difference between the existing Scotch flavor wheel and our wheel indicates a potential direction for future research into flavor development in American whiskeys.

Notably, our wheel is missing the non-aromatic categories (taste and mouthfeel) and the "off-flavor" categories (oily, cheesy, and sulfury). The basic tastes are miscategorized into other clusters, but the off-flavor terms are entirely missing from the dataset. Many potential explanations for this exist, from the hope that these attributes are not widely present in commercially available whiskies and are more useful for quality control (an idea proposed by Lee et al. (2001) themselves) to the possibility that the reviewers, seeking to preserve their good relationships with producers, will simply not publish scathing reviews with words like "acrid" or "rotten vegetable".

This is a good example of the importance of data quality, irrespective of quantity. A flavor wheel generated from the kind of professional product reviews found in this corpus is sufficient for marketing purposes, but not a quality control program or a researcher interested in off-flavors. Future work desiring more comprehensive flavor wheels could attempt to better-utilize rare descriptors by weighting terms from low-scoring reviews more heavily in frequency calculation or by oversampling (which could also help balance other groups with unequal representation such as production categories). However, rather than trying to recover rare events from an easily-accessible dataset, it would likely be more effective to supplement the primary corpus with other, smaller but more targeted sources of text to cover any gaps (e.g., highly flawed whiskies), such as negative consumer reviews, customer complaints, tasting notes from quality assurance panels, or even free text data solicited in a lab setting.

5.3. Limitations and advantages

Grouping similar descriptors is perhaps the most challenging part of any descriptive sensory study. In Descriptive Analysis this requires lengthy discussion and the use of physical references; in rapid descriptive methods like Flash or Free-Choice Profiling this alignment may be statistical (e.g., using General Procrustes Analysis or other alignment techniques). An NLP workflow has more in common with the latter, but the size of the dataset provides more options. This is of course a potential advantage, as many techniques familiar to sensory scientists (e.g., SVD-based approaches, HCA) will function well, as is clearly demonstrated in our example.

However, this flexibility is also a key disadvantage of NLP approaches: the wide array of viable analysis options requires the analyst to make many, potentially subjective decisions which may significantly affect the outcomes (Table 1). For example, as demonstrated in Appendix C, decisions about dimensionality will result in very different decisions about descriptor similarity. Similarly, the choice of

Table 1
Describes the general steps seen in most term clustering workflows, the methods utilized in this paper's case study for each step, and lists some alternatives (non-exhaustive) that could be considered.

General Workflow Step	Method Used in Case Study	Alternatives (non-exhaustive)
Dimensionality reduction	Correspondence Analysis	Latent Semantic Analysis, Principal Components Analysis, Latent Dirichlet Allocation, semantic word embeddings (e.g. Word2Vec)
Determining *k* dimensions to keep	Scree Plot	Significance tests, percentage of inertia, rules of thumb
Dissimilarity metric	Euclidean Distance (on CA factor scores)	Chi-square, Manhattan, Euclidean Squared, Spearman
Clustering	Agglomerative Hierarchical Clustering	K-means clustering
Aggregation	Ward's method	Centroid, single linkage
Number of partitions	Elbow method	Theory-based, Hartigan index, Average Silhouette, Gap statistic

dimensionality reduction tool (PCA vs CA, etc) and of clustering parameters (distance metric, agglomeration heuristic) will each have a significant effect. This "unsupervised learning" problem is an active area of research in NLP and the larger area of statistical machine learning (James et al., 2013), and there are many advances that can be made for the particular area of sensory descriptions, but at this point the analyst should bring both subject-area and statistical expertise to the problem in order to make good decisions.

A second challenge comes from the assumption, in this and other NLP (and even novel sensory) research, that descriptor co-occurrence implies descriptor similarity: that is, if "caramel" and "oak" co-occur in more whisky reviews than "caramel" and "peat", should we in fact assume that "caramel" and "oak" are also more similar sensory descriptors? This assumption deserves more research, as do the identification of alternative metrics for similarity.

6. Conclusions & future work

This paper introduces the methods of Natural Language Processing as a toolbox for sensory scientists to tackle unstructured text datasets, from solicited comments in controlled studies to big datasets obtained without an experimental design. In our case study of the latter, even with a very preliminary workflow containing a number of arbitrary decisions, the results demonstrate the utility of using NLP on existing data for sensory science, particularly for lexicon development. The analysis of large samples of existing data, as in this study, is only feasible when both collection and cleaning of textual data can be streamlined or automated.

We hope that more researchers will begin to use similar techniques to answer their own research questions. Researchers should always carefully consider the source of their text data, and determine whether it is truly representative of the product category they are interested in. More than one corpus may be necessary for sufficient coverage, as can be seen in the lack of off-flavors in our novel flavor wheel. By using tools like web scrapers to rapidly collect the bulk of the data, an analyst can still save time and resources while collecting a small amount of more-intensive supplemental data (e.g., rare product categories or noncommercial samples).

This paper outlines a full workflow of data collection, cleaning, and grouping which can be used to generate a lexicon from web-based product descriptions. Many of the methods outlined in this paper could be used *à la carte* to answer different research questions or be applied to datasets of a different size. The two largest limitations in the current workflow are the lack of an ability to specifically separate hedonic and production-related from descriptive words and the difficulty in obtaining stable, semantically-meaningful ("synonymous") groups of words. The field would benefit greatly from a tagger which could identify words with sensory relevance. For this and improving other tagging schemes (e.g., POS, sentiment), annotated datasets of sensory descriptions are needed, the production of which could be expedited by active learning methods (Kholghi, Sitbon, Zuccon, & Nguyen, 2017).

When it comes to the grouping of words, we would encourage future research to investigate methods which do not rely on dimension reduction and HCA due to their large decision space and the considerable impact of these arbitrary decisions on the final output. While a solution based on "AI"—state-of-the-art machine learning research—would be exceptional, for now methods such as network analysis may show more promise for being able to work with the small- and medium-sized datasets sensory scientists have available (as opposed to the billions of words required to train embeddings). Additionally, it may be some time before any automated synonym-grouping method is as accurate as manual text analysis, although automated methods remain much more feasible for large datasets.

Beyond the methodological questions, the application of NLP to existing text for sensory science is wide open for new research. Different corpora (e.g. consumer reviews, non-English text, various food and consumer products, social media posts, free response data from a controlled sensory test) will present different challenges. For example, in their dataset Valente (2016) found a low average of 1.32 sensory attributes per wine review, which would require a more accurate method of identifying the very small number of genuine sensory descriptors. Social media and other sources of text without attached quality scores or product ratings may necessitate sarcasm detection or sentiment analysis to identify positive or negative descriptions (Kumar, Narapareddy, Aditya Srikanth, Malapati, & Neti, 2020). These and other challenges will provide new opportunities to develop NLP tools and methods suitable for sensory science.

CRediT authorship contribution statement

Leah M. Hamilton: Conceptualization, Software, Methodology, Formal analysis, Writing - original draft. **Jacob Lahne:** Methodology, Validation, Formal analysis, Writing - review & editing, Supervision, Funding acquisition.

Acknowledgments

This work has been partially supported by a grant from the Virginia Tech Institute for Creativity, Art, and Technology.

Appendix A

The body of the paper briefly introduces the concept of a web scraper as a tool for data collection in Section 3.1, but most of the text in Section 3 focuses on the workflow used in the case study rather than the theory of scraping. We suspect that some readers interested in scraping data for their own projects will not be familiar with the operations of a scraper or the structure of websites, so this appendix will provide some basic fundamentals of website structure and typical scraper operation to aid the reader in understanding more of the linked resources.

As mentioned in Section 3.1, a web-scraping script takes a list of URLs and loads each page in the list, following the programmer's instructions to

locate and extract information from the pages. This appendix will first discuss how webpages are laid out and how scrapers can find information in any given page, then will turn to the question of how to compile a list of pages. This is structured in this way as finding URLs from an initial page/set of pages is often an iterative scraping process.

A.1. Accessing and Navigating a Web Page

For each web page, a scraper requests the page data through the Hyptertext Transfer Protocol (HTTP), then parses the page information (as Hypertext Markup Language, HTML) into a tree structure (Fig. A.1c), and finally follows the programmer's instructions to find the desired information within that tree. Depending on the website's structure, this process may need to be iterative, following links from an initial list of pages to create a complete list.

The first step in viewing a web page, either for a user or a bot, is to make a request for the page via HTTP (or HTTPS, for websites using secure encryption). The user only needs the URL (e.g., https://www.elsevier.com/) and an internet connection to request a web page. The contents of that web page are returned in HTML, a text-based specification which houses the page content inside of tags that describe the page layout, formatting, and functionality. Most tags have an opening (e.g., < div >) and closing (e.g., < /div >) version, with the tag attributes applying to other content and tags in-between. Some example HTML is shown in Fig. A.1b.

A pair of opening and closing tags with the content in-between are collectively called an element. Elements must be arranged in a nested tree structure (Fig. A.1b), with no two elements partially overlapping. When a human user is viewing a web page, the browser will parse the HTML-tagged text into a tree in order to display the page in a 2D graphical layout (Fig. A.1a), and similarly, the scraper will generate a parse tree (Fig. A.1c) which is then navigated directly to search for specific elements.

The above steps are universal to any bot or user accessing any web page, but the identification and extraction of relevant information from the parse tree will vary from application to application, depending on where the relevant information is located. Most of the work in creating a custom scraper is in finding the HTML elements which contain the relevant information. Usually, HTML opening tags contain additional information in the form of element attributes, and these attributes allow the scraper to find relevant information. The first opening div tag in Fig. A.1b has an attribute called "class" with the value "search-result-body". Class attributes, which are like categorical labels, and ID attributes, which are names unique to a single element on the page (also seen in Fig. A.1b), are most often the pointers to the elements of interest on a page.

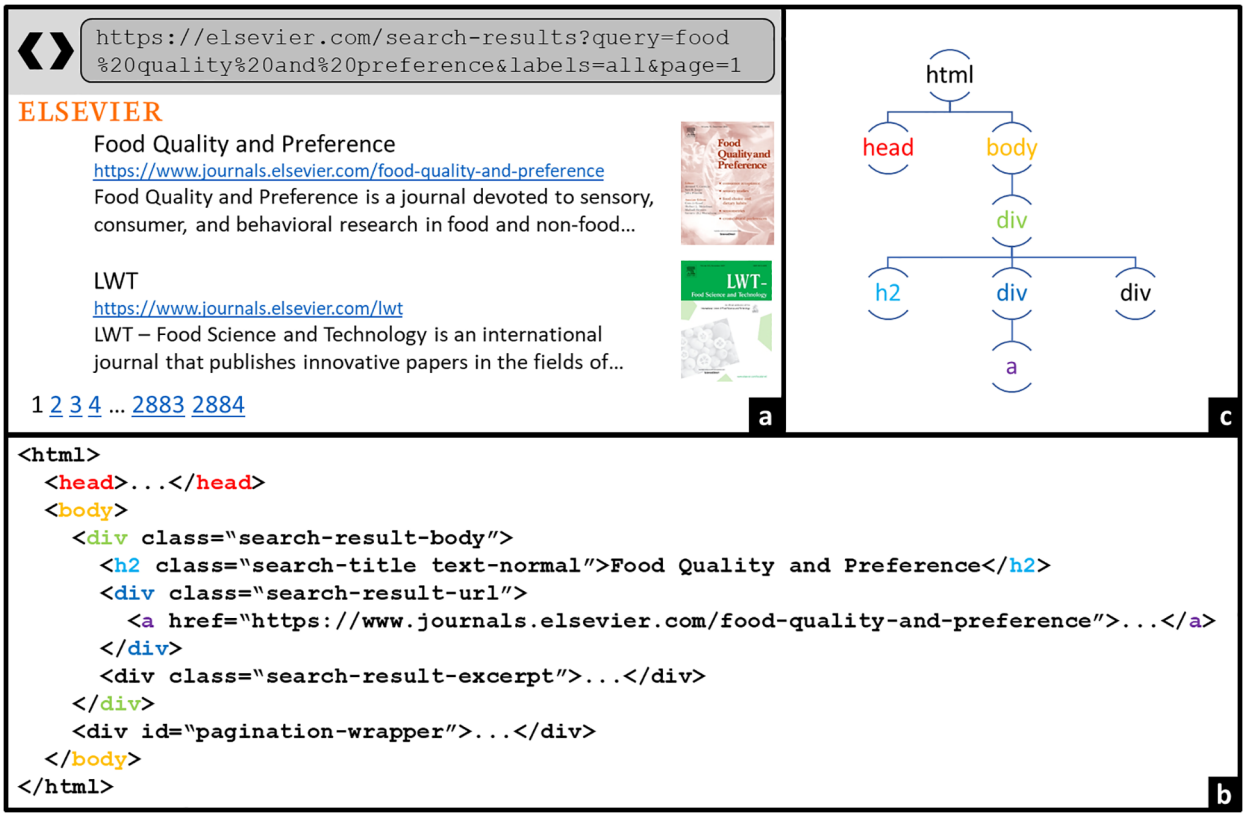

Fig. A.1. An example search page based on the Elsevier journal search. a) A graphical view of the website elements, as a human user would see them in a browser, displaying the first of 2884 pages of search results for the string "food quality and preference". b) A partial view of the HTML used to display this page graphically for the human user. Some pairs of opening and closing tags are color-coded for clarity. c) A parse tree representation of the nested tags shown in the HTML excerpt of b.

A full exploration of the structure of HTML is beyond the scope of this manuscript. There are many, excellent, in-depth resources for understanding HTML tags, such as the beginner-friendly W3Schools web design tutorial (W3Schools, 2019) or the `BeautifulSoup4` documentation on extracting information from HTML trees (Richardson, 2018). There are many tools in modern web browsers, such as Google Chrome's DevTools (https://developers.google.com/web/tools/chrome-devtools) or Mozilla Firefox's Developer Tools (https://developer.mozilla.org/en-US/docs/Tools), which can aid in viewing website HTML to identify the relevant element identifiers before scraping or crawling. The reader interested in web scraping is encouraged to use developer tools to familiarize themselves with websites of interest.

A.2. Navigating a Website Structure with a Scraper

The above scraping process will usually be repeated across multiple pages to get the desired quantity of information from a website. The bot needs the URL for each page the programmer wants information from, but the programmer may not have a pre-existing list. Luckily, the hyperlinks that human users can use to navigate around website are also accessible by bots. Hyperlinks are created with < a > elements, and the URLs are stored in the href attribute (Fig. A.1b). Again, the process of gathering relevant links from < a > tags will be specific to the website and project goal. Sometimes, a single index page or site map will contain links to each page containing desired information, in which case a bot or user could collect URLs from links on this page. Lawson (2015) discusses how to search for and gather URLs from a scraper-oriented site map.

Often, however, there are too many pages to practically list in any one location of a website. There are several methods a website designer can employ to store only the information that is different between a large number of related pages and/or pages which may be generated programmatically like search pages. A simple example is the GET protocol, which transmits additional information about the content to display after the web page is specified in the URL. As is shown in Fig. A.1a, the search page URL contains a question mark with more parameters after it, separated by ampersands. In this case, the variables are "query", "labels", and "page", with "query" containing the search string "food quality and preference" (note that "%20" is used to replace a space in URLs), "labels" being metadata about the types of results we want, and "page" being the page number. All of these define which results to display on the search page (Fig. A.1b).

GET is the method of dynamically defining page content that is most easily manipulated by a scraper, as the information is all present in the URL and usually easily interpreted and manipulated. Depending on the goal of the scraper, we could manipulate the values of the GET variables to loop through page numbers, different queries, etc. Two examples of different GET manipulations are shown in Fig. 2 and discussed in Section 3.2.

Some websites, however, will use other HTTP methods for communication of this information, or will use tools like Javascript to manipulate or load information after the initial HTTP request with which a simple scraper interacts. Such websites may require bots which can "click" and interact with Javascript, and still others may require more complex crawlers that follow every link on every page in order to index the website. Such structures are outside of the scope of this paper and interested readers can consult Lawson (2015) for a more in-depth review of scrapers and crawlers using various Python libraries.

The process of scraping relevant data is highly website-specific, but in most cases the amount of programming work required for the customization is minimal. For more practice with navigating page HTML and CSS selectors (the basis of packages like `rvest` and `BeautifulSoup4`), the CSS Diner tutorial is a very beginner-friendly resource (Pacholski, 2018).

Appendix B

This appendix includes several lists of words/word parts/word endings that were used in the whisky case study text cleaning workflow described in Section 4.2.

B.1. Adjective Endings

A custom stemmer was written for this application which removed adjective endings from all tokens to combine noun and adjective forms of words where possible (Section 4.2.2). The following sequences of letters were removed when they were located at the end of a lemmatized token:

- ful
- full
- ness
- iness
- al
- ish
- like
- ied
- ed
- ive
- y
- ey

B.2. Whitelist of Verb-Tagged Tokens to Keep

The POS as tagged by `SpaCy` was used to filter out a large proportion of nondescriptive text. In addition to keeping all adjectives and nouns, the following tokens (shown in their lemmatized and stemmed forms) were kept even when they were tagged as verbs:

- almond
- aniseed
- bake
- blacken
- blueberry
- buckwheat
- burn
- candy
- char
- clove
- damp
- dry
- flax
- fruit
- fry
- glaze
- ice
- malt
- molass
- nectarine
- nut
- oak
- oxidize
- papaya
- pepper
- roast
- sherri
- smoke
- spice
- spike
- stew
- strawberri
- sweeter
- toast
- wax
- wheat

This list was manually curated and used as described in Section 4.2.3.

Appendix C

In Section 5 of the paper, we mention the “curse of dimensionality” and the problems associated with the number of arbitrary decisions required in applying a traditional clustering workflow to such sizable data. This appendix outlines a particular instance of this issue in the whisky lexicon case study through the example of the number of dimensions kept from CA for clustering (one of several such instances).

While 7 dimensions (Fig. C.1b) were ultimately retained, Fig. C.1 demonstrates an interesting effect in choosing this cutoff (varying k). As the number of CA dimensions kept for distance calculation is increased, the words do not all get farther apart from each other uniformly. Large, relatively homogenous groups of terms are slow to be separated by the addition of higher dimensions, compared to a few individual words or small clusters of words (e.g. “barbecue” and “sauce” which almost exclusively appear together) which become more isolated. The use of more dimensions for distance calculations tends towards making a few large clusters and many distinct 1–3 word clusters until eventually there is no meaningful way to cluster the terms. We used this justification for picking a smaller number of dimensions than the apparent “elbow” on the scree plot. Since the 2D distances place some clearly unrelated words (e.g. “licorice” and “smokey”; “allspice” and “flowery”) very close together (Fig. C.1a), it is apparent that choosing the correct number of dimensions is a balancing act between making sure that real clusters of distinct words are separated in their own group without over-subdividing groups based on essentially random variation.

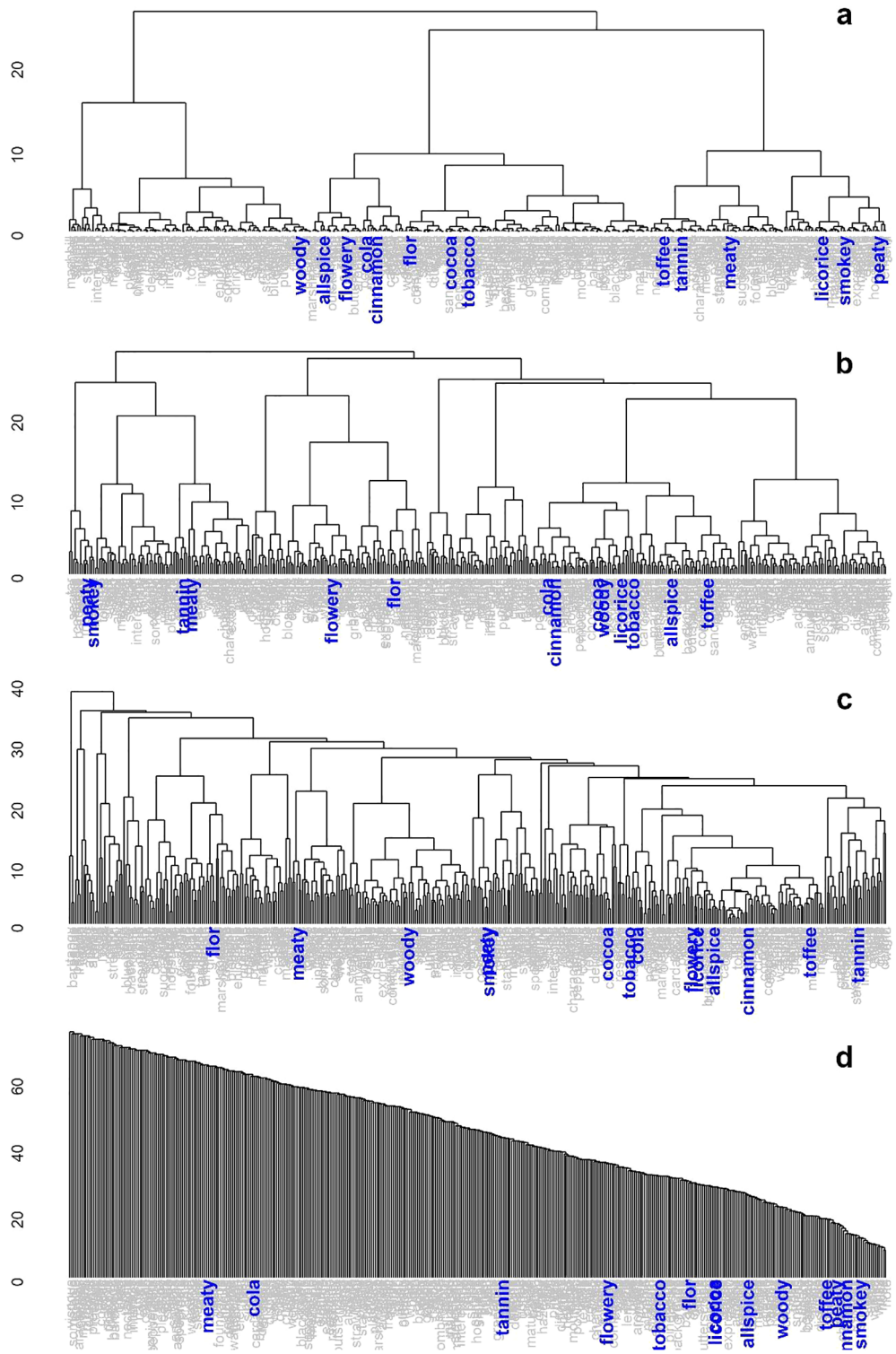

Fig. C.1. Dendrograms produced by conducting the same HCA procedure on all 407 terms, produced by keeping different numbers of dimensions: a) 2 dimensions, b) 7 dimensions, c) 25 dimensions, d) all 406 dimensions (no dimensionality reduction). The same words are highlighted in all 4 plots to demonstrate the differing word similarities when keeping different numbers of dimensions, and the heights on the y axis are the increase in within-groups sums of squares at each agglomeration step.

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodqual.2020.103926.

References

Ahn, Y., Ahnert, S. E., Bagrow, J. P., & Barabasi, A.-L. (2011). Flavor network and the principles of food pairing. *Scientific Reports, 1*.

Bates, M. (1995). Models of natural language understanding. *Proceedings of the National Academy of Sciences of the United States of America, 92*(22), 9977–9982. https://doi.org/10.1073/pnas.92.22.9977.

Bécue-Bertaut, Mónica (2014). Tracking verbal-based methods beyond conventional descriptive analysis in food science bibliography. A statistical approach. *Food Quality and Preference, 32*, 2–15. https://doi.org/10.1016/j.foodqual.2013.08.010.

Bécue-Bertaut, Mónica, Álvarez-Esteban, R., & Pagès, J. (2008). Rating of products through scores and free-text assertions: Comparing and combining both. *Food Quality and Preference, 19*(1), 122–134. https://doi.org/10.1016/j.foodqual.2007.07.006.

Bécue-Bertaut, Monica, & Lê, S. (2011). Analysis Of Multilingual Labeled Sorting Tasks: Application To A Cross-Cultural Study In Wine Industry. *Journal of Sensory Studies, 26*(5), 299–310. https://doi.org/10.1111/j.1745-459X.2011.00345.x.

Bird, S., Klein, E., & Loper, E. (2019a). 2. Accessing Text Corpora and Lexical Resources. In Natural Language Processing with Python (2nd Ed). http://www.nltk.org/book/ch02.html.

Bird, S., Klein, E., & Loper, E. (2019b). 3. Processing Raw Text. In Natural Language Processing with Python (2nd Ed). http://www.nltk.org/book/ch03.html.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bruce, P., & Bruce, A. (2017). Data and Sampling Distributions. In Practical Statistics for Data Scientists (pp. 43–77). O'Reilly.

Bryson, L. (2014). *Tasting Whiskey.* Storey Publishing.

Chaulagain, R. S., Pandey, S., Basnet, S. R., & Shakya, S. (2017). Cloud Based Web Scraping for Big Data Applications. *IEEE International Conference on Smart Cloud (SmartCloud), 2017*, 138–143. https://doi.org/10.1109/SmartCloud.2017.28.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing By Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*, 391–407.

Drake, M. A., & Civille, G. V. (2003). Flavor Lexicons. *Comprehensive Reviews in Food Science and Food Safety, 2*(1), 33–40. https://doi.org/10.1111/j.1541-4337.2003.tb00013.x.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E. H., & Smith, N. A. (2014). Retrofitting Word Vectors to Semantic Lexicons. CoRR, abs/1411.4. http://arxiv.org/abs/1411.4166.

Fellbaum, C. (1999). WordNet : An Electronic Lexical Database.: Vol. 2nd printi. A Bradford Book. http://login.ezproxy.lib.vt.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=48571&site=eds-live&scope=site.

Flood, B. J. (1999). Historical note: The Start of a Stop List at Biological Abstracts. Journal of the American Society for Information Science, 50(12), 1066–1066. https://doi.org/10.1002/(SICI)1097-4571(1999)50:12 < 1066::AID-ASI5 > 3.0.CO;2-A.

Gillespie, M. (2018). WhiskyCast. CaskStrength Media. https://whiskycast.com/.

Gold, Z., & Latonero, M. (2018). Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping. *Washington Journal of Law, Technology, & Arts, 13*(3), 275–312.

Greenacre, M. (2017). *Correspondence Analysis in Practice (3rd Ed).* CRC Press.

Heisserer, D. M., & Chambers, E., IV (1993). Determination of the Sensory Flavor Attributes of Aged Natural Cheese. *Journal of Sensory Studies, 8*(2), 121–132. https://doi.org/10.1111/j.1745-459X.1993.tb00207.x.

Hennion, A. (2015). Paying Attention: What Is Tasting Wine About? In A. B. Antal, M. Hutter, & D. Stark (Eds.). *Moments of Valuation: Exploring Sites of Dissonance* (pp. 37–56). Oxford University Press.

Heymann, H., King, E. S., & Hopfer, H. (2014). Classical Descriptive Analysis. In P. Varela & G. Ares (Eds.), Novel Techniques in Sensory Characterization and Consumer Profiling (pp. 9–40). CRC Press.

Honnibal, M., & Montani, I. (2018). *SpaCy (Version 2.0.16) [Computer software].* ExplosionAI GmbH.

Ickes, C. M., Lee, S.-Y., & Cadwallader, K. R. (2017). Novel creation of a rum flavor lexicon through the use of web-based material. *Journal of Food Science, 82*(5), 1216–1223.

Jackson, M. (2017). *Whisky: The definitive world guide.* Dorling Kindersley Ltd.

Jaeger, S. R., Beresford, M. K., Paisley, A. G., Antúnez, L., Vidal, L., Cadena, R. S., … Ares, G. (2015). Check-all-that-apply (CATA) questions for sensory product characterization by consumers: Investigations into the number of terms used in CATA questions. *Food Quality and Preference, 42*, 154–164. https://doi.org/10.1016/j.foodqual.2015.02.003.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Unsupervised Learning. In G. James, D. Witten, T. Hastie, & R. Tibshirani (Eds.). *An Introduction to Statistical Learning: with Applications in R* (pp. 373–418). New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7_10.

Kassambara, A. (2017). *Determining the Optimal Number of Clusters. Practical Guide to Cluster Analysis in R.* Unsupervised Machine Learning. STHDA.

Kholghi, M., Sitbon, L., Zuccon, G., & Nguyen, A. (2017). Active learning reduces annotation time for clinical concept extraction. *International Journal of Medical Informatics, 106*, 25–31. https://doi.org/10.1016/j.ijmedinf.2017.08.001.

Koki, A. (2018). Whisky Data Scraping (64dcdb144ca287fc4e3080050721465ffcf20b9d) [Computer software]. https://github.com/koki25ando/Whisky-Data-Scraping.

Kostov, B., Bécue-Bertaut, M., & Husson, F. (2014). An original methodology for the analysis and interpretation of word-count based methods: Multiple factor analysis for contingency tables complemented by consensual words. *Food Quality and Preference, 32*, 35–40. https://doi.org/10.1016/j.foodqual.2013.06.009.

Kumar, A., Narapareddy, V. T., Aditya Srikanth, V., Malapati, A., & Neti, L. B. M. (2020). Sarcasm Detection Using Multi-Head Attention Based Bidirectional LSTM. *IEEE Access, 8*, 6388–6397. https://doi.org/10.1109/ACCESS.2019.2963630.

Lahne, J., Abdi, H., Collins, T., … Heymann, H. (2019). Bourbon and Rye Whiskeys Are Legally Distinct but Are Not Discriminated by Sensory Descriptive Analysis: Descriptive analysis of American whiskey…. *Journal of Food Science, 84*(3), 629–639. https://doi.org/10.1111/1750-3841.14468.

Lahne, J., Trubek, A. B., & Pelchat, M. L. (2014). Consumer sensory perception of cheese depends on context: A study using comment analysis and linear mixed models. *Food Quality and Preference, 32*, 184–197. https://doi.org/10.1016/j.foodqual.2013.10.007.

Lawless, L. J. R., & Civille, G. V. (2013). Developing Lexicons: A Review: Lexicon Review. *Journal of Sensory Studies, 28*(4), 270–281. https://doi.org/10.1111/joss.12050.

Lawson, R. (2015). *Web Scraping with Python.* Packt Publishing.

Lee, K.-Y. M., Paterson, A., Piggott, J. R., & Richardson, G. D. (2001). Origins of Flavour in Whiskies and a Revised Flavour Wheel: A Review. *Journal of the Institute of Brewing, 107*(5), 287–313. https://doi.org/10.1002/j.2050-0416.2001.tb00099.x.

Lestringant, P., Delarue, J., & Heymann, H. (2019). 2010–2015: How have conventional descriptive analysis methods really been used? A systematic review of publications. *Food Quality and Preference, 71*, 1–7. https://doi.org/10.1016/j.foodqual.2018.05.011.

Manning, C., Raghavan, P., & Schütze, H. (2008). Stemming and Lemmatization. In Introduction to Information Retrieval. Cambridge University Press. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

Martinez, A. R. (2012). Part-of-speech tagging. *WIREs. Computational Statistics, 4*, 107–113. https://doi.org/10.1002/wics.195.

Meyners, M., & Castura, J. C. (2014). Check-All-That-Apply Questions. In Paula Varela, & G. Ares (Eds.). *Novel Techniques in Sensory Characterization and Consumer Profiling* (pp. 271–306). CRC Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [Cs]. http://arxiv.org/abs/1301.3781.

Miller, G. (2019). *Whisky Science—A Condensed Distllation.* Springer Nature.

Mullen, L. (2018). Introduction to the tokenizers Package. https://cran.r-project.org/web/packages/tokenizers/vignettes/introduction-to-tokenizers.html.

Noble, A. C., Arnold, R. A., Buechsenstein, J., Leach, E. J., Schmidt, J. O., & Stern, P. M. (1987). Modification of a Standardized System of Wine Aroma Terminology. *American Journal of Enology and Viticulture, 38*(2), 143–146.

Pacholski, L. (2018, November 27). CSS Diner. http://flukeout.github.io/.

Perrin, L., Symoneaux, R., Maître, I., Asselin, C., Jourjon, F., & Pagès, J. (2008). Comparison of three sensory methods for use with the Napping® procedure: Case of ten wines from Loire valley. *Food Quality and Preference, 19*(1), 1–11. https://doi.org/10.1016/j.foodqual.2007.06.005.

Phetxumphou, K., Miller, G., Ashmore, P., Collins, T., & Lahne, J. (In press). Mashbill and barrel aging effects on the sensory and chemical profiles of American whiskey. Journal of the Institute of Brewing, XX, XX.

R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.r-project.org.

Richardson, L. (2018). Beautiful Soup (Version 4.6.0) [Computer software]. https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

Romero, G. (2019, July 6). Rethinking rule-based lemmatization. spaCy IRL 2019. https://www.youtube.com/watch?v=88zcQODyuko.

Shanken, M., Lindenmuth, J., Schwenk, M., Barton, S. S., Simmons, T., & Kostro, Z. (Eds.). (2018). WhiskyAdvocate. M Shanken Communications. http://whiskyadvocate.com/.

Shapin, S. (2016). A taste of science: Making the subjective objective in the California wine world. *Social Studies of Science, 46*(3), 436–460. https://doi.org/10.1177/0306312716651346.

Singh, A., Shukla, N., & Mishra, N. (2018). Social media data analytics to improve supply chain management in food industries. *Transportation Research Part E: Logistics and Transportation Review, 114*, 398–415. https://doi.org/10.1016/j.tre.2017.05.008.

Spencer, M., Sage, E., Velez, M., … Guinard, J.-X. (2016). Using Single Free Sorting and Multivariate Exploratory Methods to Design a New Coffee Taster's Flavor Wheel: Design of coffee taster's flavor wheel…. *Journal of Food Science, 81*(12), S2997–S3005. https://doi.org/10.1111/1750-3841.13555.

Stevenson, M., & Agirre, E. (2018). Word Sense Disambiguation. In R. Mitkov (Ed.), The Oxford Handbook of Computational Linguistics (2nd Ed). Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199573691.001.0001/oxfordhb-9780199573691-e-28.

Symoneaux, R., & Galmarini, M. V. (2014). Open-Ended Questions. In P. Varela, & G. Ares (Eds.). *Novel Techniques in Sensory Characterization and Consumer Profiling* (pp. 307–333). CRC Press.

Symoneaux, R., Galmarini, M. V., & Mehinagic, E. (2012). Comment analysis of

consumer's likes and dislikes as an alternative tool to preference mapping. A case study on apples. *Food Quality and Preference, 24*(1), 59–66. https://doi.org/10.1016/j.foodqual.2011.08.013.

Teil, G., & Hennion, A. (2004). Discovering quality or performing taste? A sociology of the amateur. In M. Harvey, A. McMeekin, & A. Warde (Eds.). *Qualities of food* (pp. 19–37). Manchester University Press.

ten Kleij, F., & Musters, P. A. D. (2003). Text analysis of open-ended survey responses: A complementary method to preference mapping. *Food Quality and Preference, 14*(1), 43–52. https://doi.org/10.1016/S0950-3293(02)00011-3.

Valente, C. C. (2016). Understanding South African Chenin Blanc wine by using data mining techniques applied to published sensory data [Stellebosch University]. http://hdl.handle.net/10019.1/98866.

Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: A review of new descriptive methods in food science. *International Journal of Food Science & Technology, 47*(8), 1563–1578. https://doi.org/10.1111/j.1365-2621.2012.03022.x.

W3Schools. (2019). HTML5 Tutorial. Refsnes Data. https://www.w3schools.com/html/default.asp.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., … Houston, A. (2013). OntoNotes Release 5.0. *Linguistic Data Consortium; Linguistic Data Consortium.*.

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science, 2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1.