Fast and Automated Sensory Analysis: Using Natural Language Processing for Descriptive Lexicon Development

Hamilton, L. M. and Lahne, J.

Virginia Polytechnic Institute and State University, Department of Food Science & Technology

INTRODUCTION

- Humans are the best instruments for analyzing food flavors, but they **do not use standardized vocabulary**. Training is **time-consuming** and often produces **less familiar vocabulary**.
- Consumer descriptions of foods are **readily available online** as reviews[1].
- Natural Language Processing (NLP) methods **automatically** analyze large volumes of text and could **rapidly create a consumer lexicon**.
- Product lexicons are useful **research** and **marketing** tools, and this method can help make them **more accessible**.

Objective:

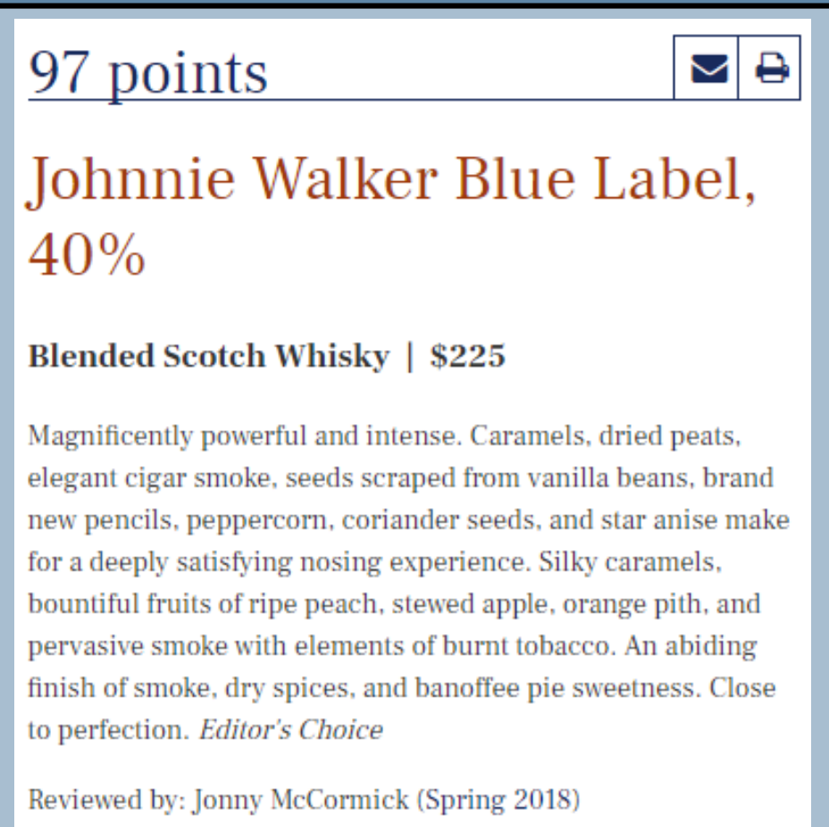
- Develop a novel NLP approach to process **large numbers of reviews** into a **consumer-friendly whisky lexicon**.

METHODS

1. Whisky Review Collection

Scraped 6597 full-text reviews from 2 websites

WhiskyAdvocate *(4288 reviews)*



97 points

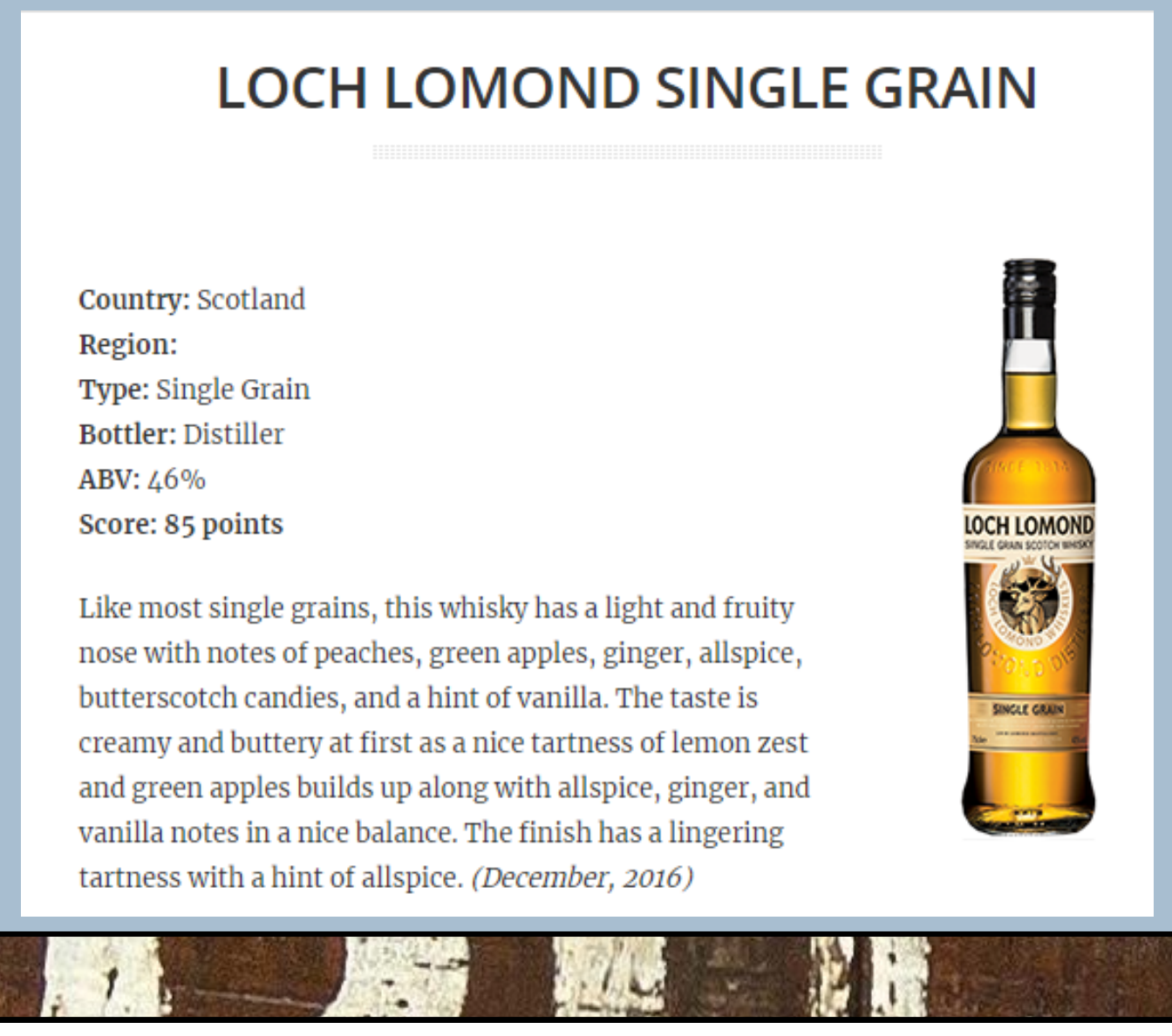
Johnnie Walker Blue Label, 40%

Blended Scotch Whisky | $225

Magnificently powerful and intense. Caramels, dried peats, elegant cigar smoke, seeds scraped from vanilla beans, brand new pencils, peppercorn, coriander seeds, and star anise make for a deeply satisfying nosing experience. Silky caramels, bountiful fruits of ripe peach, stewed apple, orange pith, and pervasive smoke with elements of burnt tobacco. An abiding finish of smoke, dry spices, and banoffee pie sweetness. Close to perfection. *Editor's Choice*

Reviewed by: Jonny McCormick (Spring 2018)

Whiskycast *(2309 reviews)*



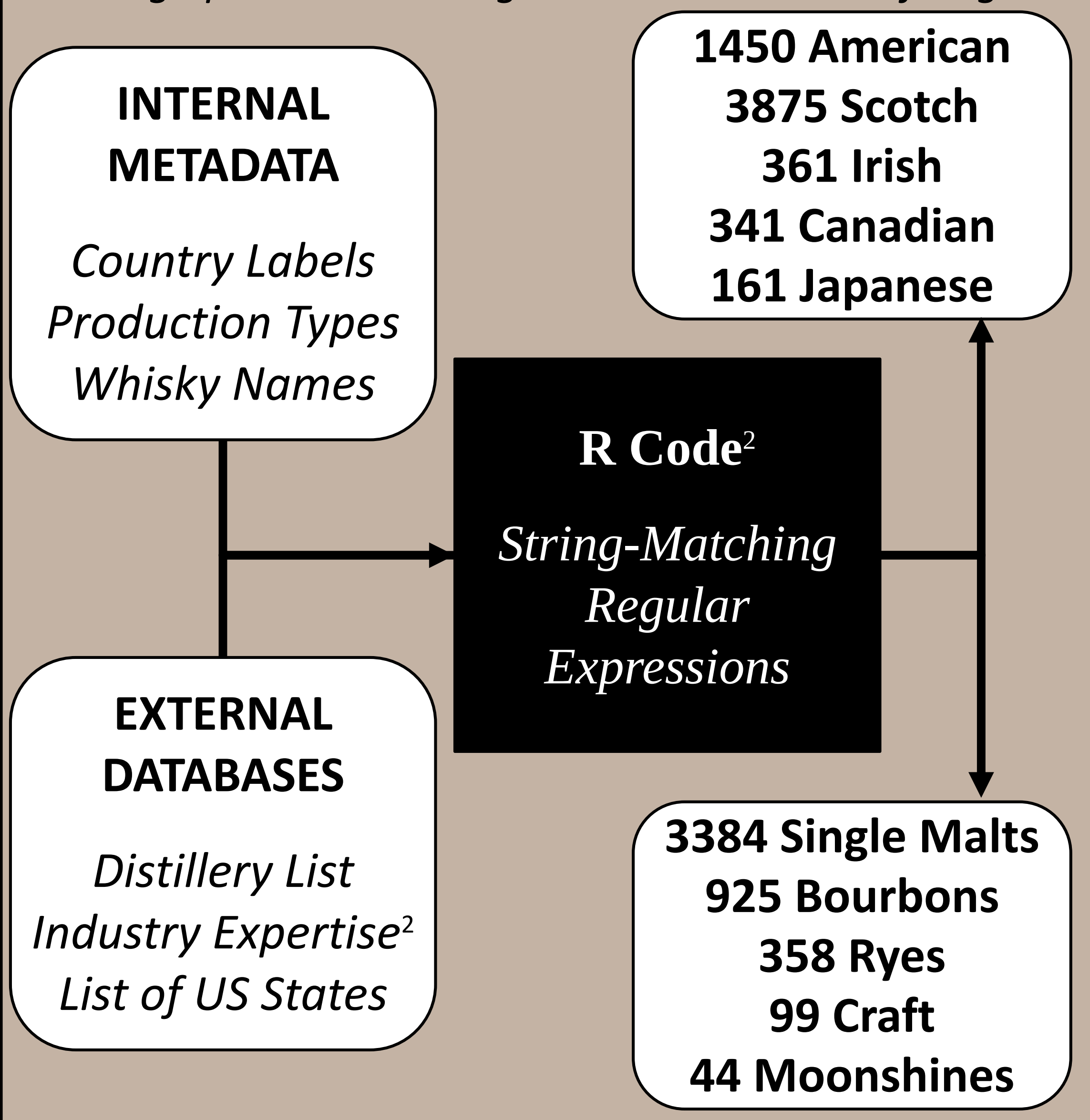
LOCH LOMOND SINGLE GRAIN

Country: Scotland
Region:
Type: Single Grain
Bottler: Distiller
ABV: 46%
Score: 85 points

Like most single grains, this whisky has a light and fruity nose with notes of peaches, green apples, ginger, allspice, butterscotch candies, and a hint of vanilla. The taste is creamy and buttery at first as a nice tartness of lemon zest and green apples builds up along with allspice, ginger, and vanilla notes in a nice balance. The finish has a lingering tartness with a hint of allspice. *(December, 2016)*

2. Whisky Classification

Used both internal review data and external databases to assign production categories and countries of origin



INTERNAL METADATA

Country Labels
Production Types
Whisky Names

EXTERNAL DATABASES

Distillery List
Industry Expertise[2]
List of US States

R Code[2]

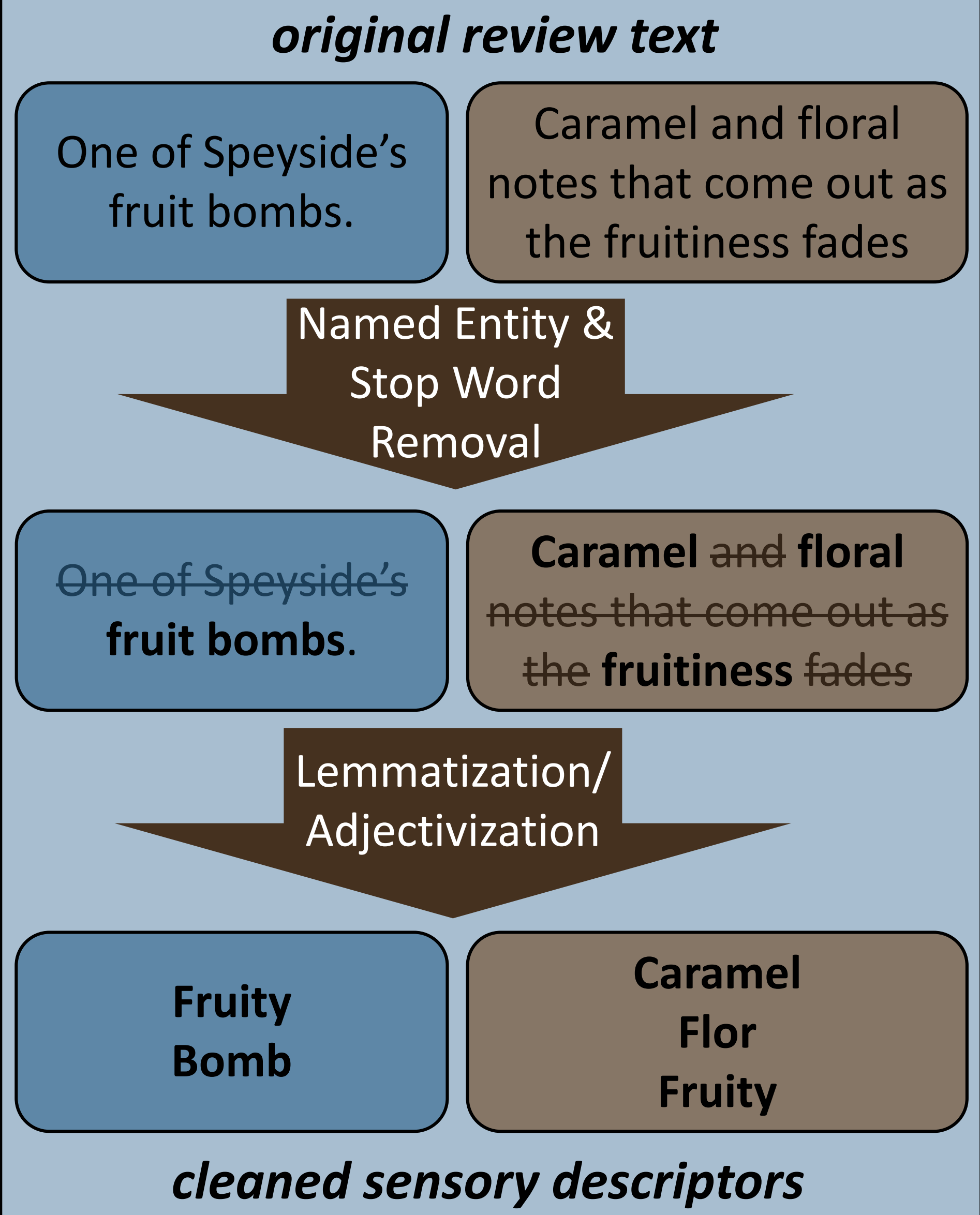
String-Matching
Regular Expressions

1450 American
3875 Scotch
361 Irish
341 Canadian
161 Japanese

3384 Single Malts
925 Bourbons
358 Ryes
99 Craft
44 Moonshines

3. Natural Language Processing

Processed raw review text into individual, relevant terms



original review text

One of Speyside's fruit bombs.

Caramel and floral notes that come out as the fruitiness fades

Named Entity & Stop Word Removal

~~One of Speyside's~~ **fruit bombs.**

Caramel ~~and~~ **floral** ~~notes that come out as the~~ **fruitiness** ~~fades~~

Lemmatization/ Adjectivization

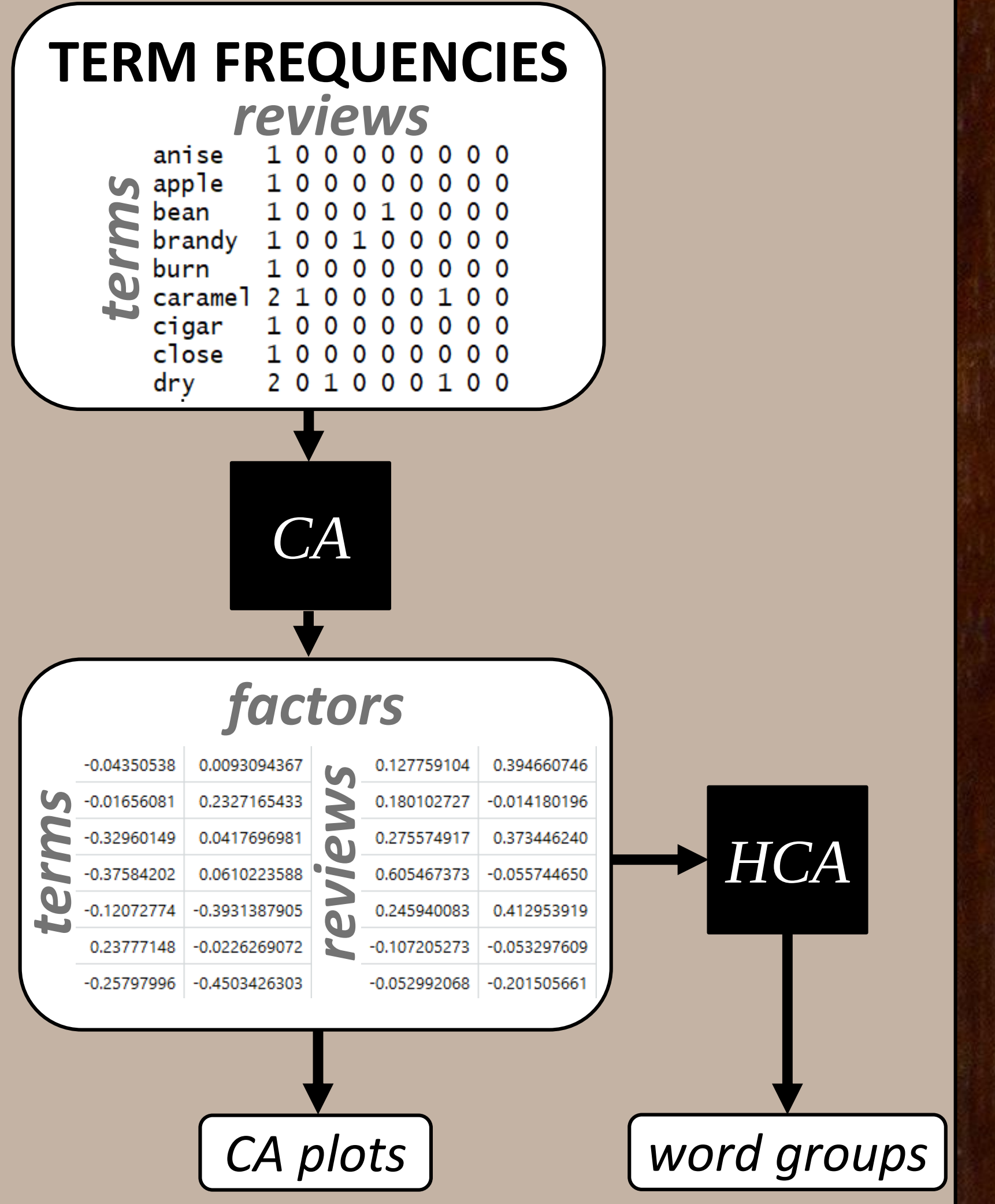
Fruity Bomb

Caramel Flor Fruity

cleaned sensory descriptors

4. Term Clustering

Selected few main axes of variation with Correspondence Analysis (CA) and used Hierarchical Clustering Analysis (HCA)[3] to group related terms



TERM FREQUENCIES

reviews / *terms*

anise	1	0	0	0	0	0	0	0	0
apple	1	0	0	0	0	0	0	0	0
bean	1	0	0	0	1	0	0	0	0
brandy	1	0	0	1	0	0	0	0	0
burn	1	0	0	0	0	0	0	0	0
caramel	2	1	0	0	0	0	1	0	0
cigar	1	0	0	0	0	0	0	0	0
close	1	0	0	0	0	0	0	0	0
dry	2	0	1	0	0	0	1	0	0

CA

factors

terms

-0.04350538	0.0093094367
-0.01656081	0.2327165433
-0.32960149	0.0417696981
-0.37584202	0.0610223588
-0.12072774	-0.3931387905
0.23777148	-0.0226269072
-0.25797996	-0.4503426303

reviews

0.127759104	0.394660746
0.180102727	-0.014180196
0.275574917	0.373446240
0.605467373	-0.055744650
0.245940083	0.412953919
-0.107205273	-0.053297609
-0.052992068	-0.201505661

HCA

CA plots

word groups

RESULTS

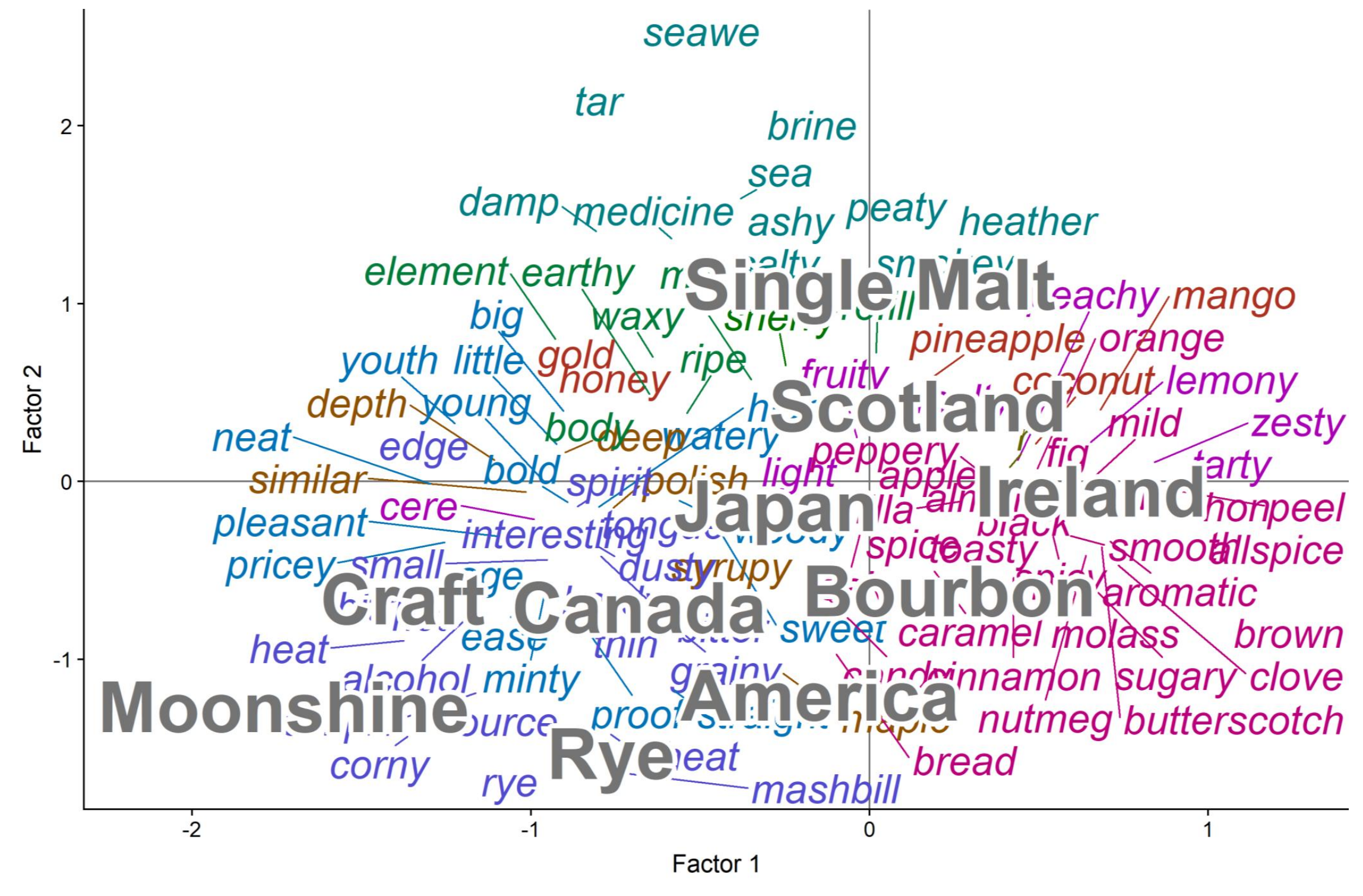


Figure 1. A CA plot of term frequencies across all 6597 reviews. The 104 terms which have the largest contributions to the first two axes are shown. Selected product categories and whisky-producing countries are shown as supplementary points.

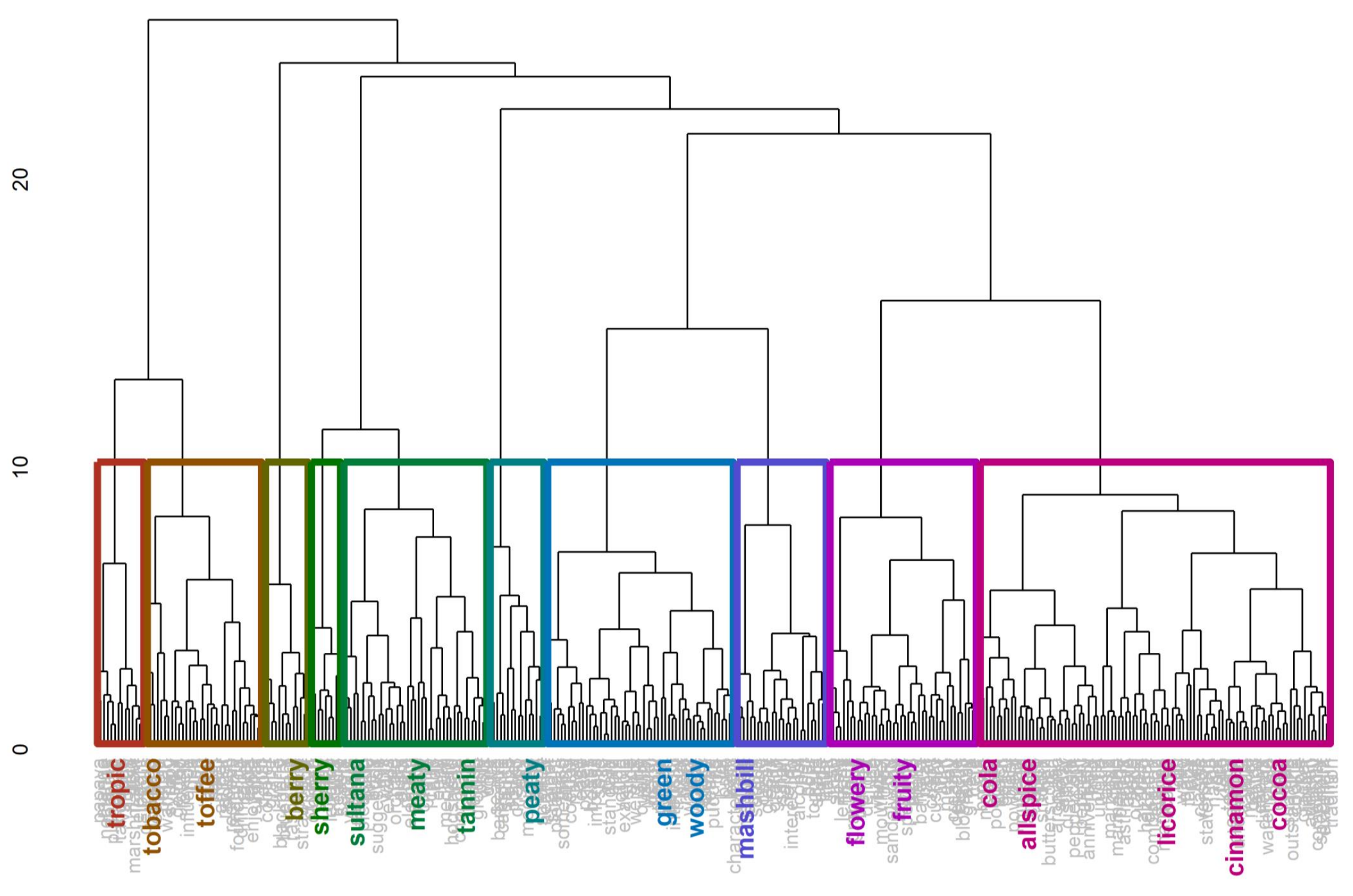


Figure 2. A dendrogram showing the distances between terms, based on a 6D CA plot (first two dimensions in **Fig 1**). Clusters and selected terms representative of each have been highlighted.

- The two most important factors driving differences in word use (**Fig 1**) are related to **whisky age** and **Scottish peat smoke**.
- Ideally, groups are comprised of **synonyms**. One cluster contains ***peaty***, ***tar***, ***smokey***, and ***medicine***, terms for peat smoke flavor[4].
- Some clusters, such as the one containing ***cola***, ***allspice***, ***licorice***, ***cinnamon***, and ***cocoa***, are not true groups of synonyms.

- The clusters (**Fig 4**) are **similar to an existing flavor wheel** (**Fig 3**) but there are fewer categories due to combination categories (e.g. ***Fruity/Floral***) and a lack of off-flavors (e.g. ***Sulfury***, ***Cheesy***).
- Some groups of (not necessarily synonymous) terms are highly associated with particular categories (**Fig 1**), which is likely driving some broad term groups like **cluster 10**.

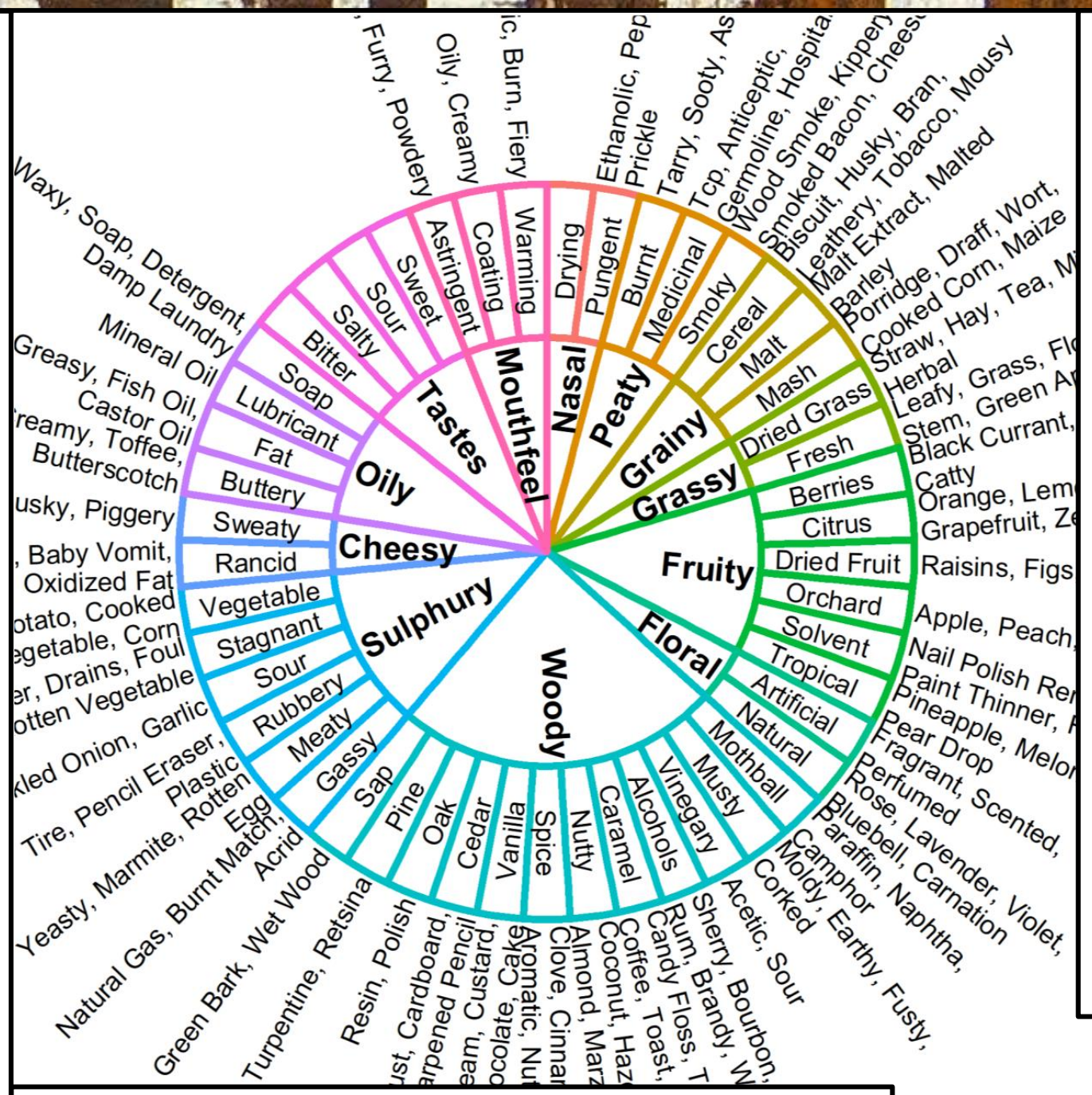
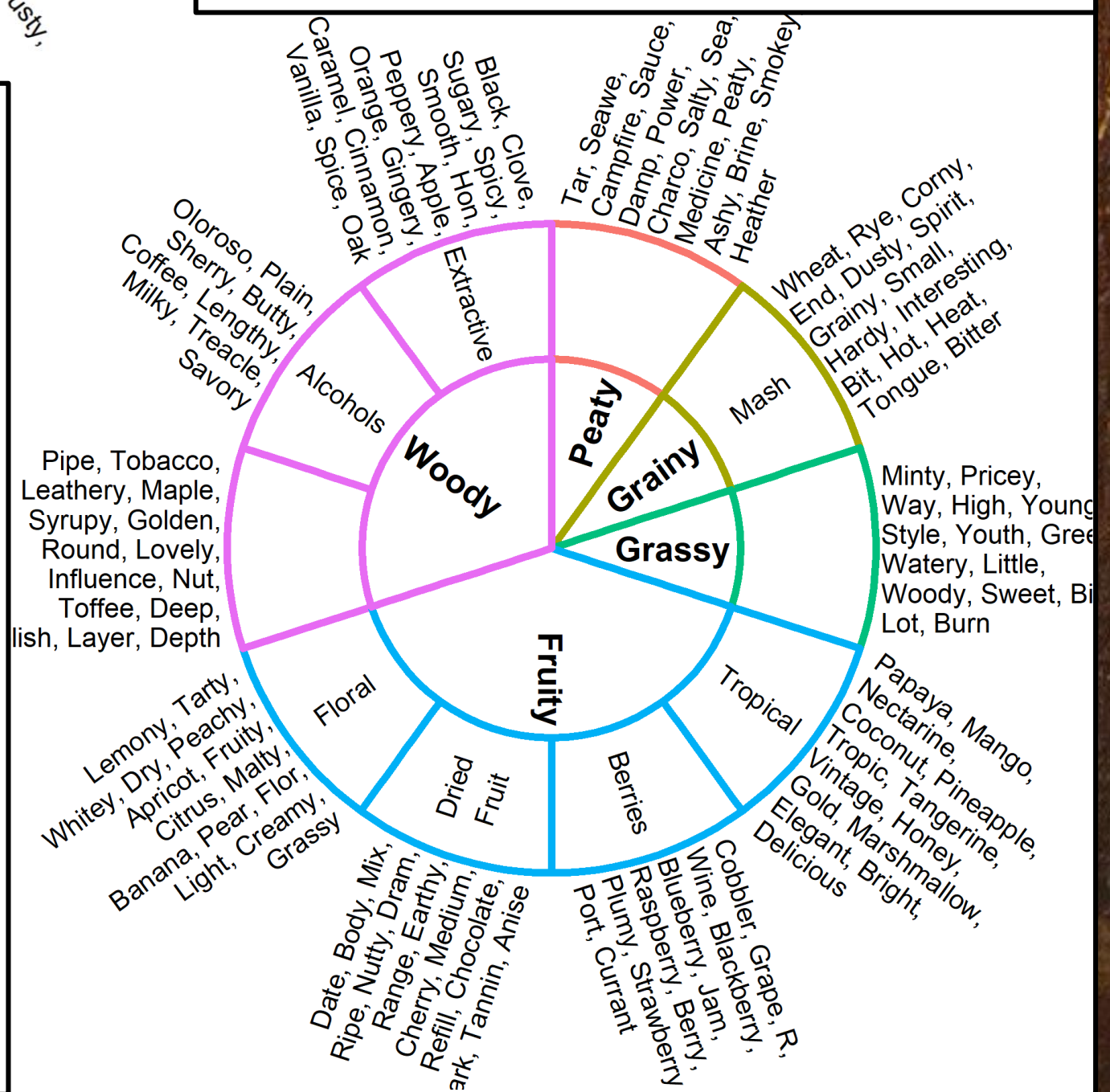


Figure 3. An existing flavor wheel for Scotch whiskies[4]. *Secondary* attributes (outer circle) are grouped into *primary* categories (inner circle). *Tertiary* terms (outer edge) are specific descriptors.

Note that a previous whisky flavor wheel grouped ***Fruity*** and ***Floral*** under the primary category ***Estery***.

Figure 4. A flavor wheel produced with an **NLP and CA-HCA workflow**. The *outer* circle contains the 10 clusters from **Fig 2** and the 15 most frequent terms in each cluster are *tertiary* terms around the edge. The structure, ordering, and *primary/secondary* labels are based on **Fig 3**, when possible.


Woody
Peaty
Grainy
Grassy
Malt
Fruity
Floral
Dried
Fruit
Berries
Tropical
Alcoholic
Extractive
Spicy
Pipe, Tobacco, Leathery, Maple, Syrupy, Golden, Round, Lovely, Influence, Nut, Toffee, Deep, Rich, Layer, Depth
Minty, Pricey, Way, High, Young, Style, Youth, Green, Watery, Little, Woody, Sweet, Big, Lot, Burn
Papaya, Mango, Nectarine, Coconut, Pineapple, Tropical, Tangerine, Melon, Grapefruit, Peach, Juicy, Bright

CONCLUSIONS

- The most prominent patterns of word usage in reviews are related to production categories, such as **moonshine/craft** or **peated** whiskies.
- Some expected groups of terms, such as categories of **fruits** (**tropical**, **berry**, and **dried**), are **easily separated into clusters**.
- Some expected groups, such as distinct **fruity** and **floral** categories, are difficult to separate due to frequent **co-occurrence in products** from a category.
- NLP and CA-HCA are capable of **isolating descriptive terms** and **grouping related terms** into categories similar to the interior of a **flavor wheel**.

FUTURE WORK

Lexicon Fine-Tuning:

- Find **sub-groups** within overall clusters (*i.e.* secondary **flavor wheel attributes**) using additional clustering steps.
- Investigate alternative methods of word-grouping such as a **neural network** model (e.g. *LDA2vec*).

Application:

- Use the developed lexicon for a consumer **check-all-that-apply** test.
- Assess whether certain **attributes** or **descriptors correlate** with **price** or **quality score** so that whisky producers can **decide what types of product to make**.

Future Projects:

- Develop the workflow into **general-use tools** for producers and researchers.
- Apply this workflow to **more reviews** of whisky or **other products** like tea or beer.
- Associate intensity terms like **not** and **slight** with the descriptors they modify.

References

[1] Ickes CM, Lee SY, Cadwallader KR (2017) Novel Creation of a Rum Flavor Lexicon Through the Use of Web-Based Material. J Food Sci 82:1216–1223. doi: 10.1111/1750-3841.13707

[2] R Core Team (2017) R: A Language and Environment for Statistical Computing

[3] Greenacre M (2017) Correspondence Analysis in Practice, 3rd Ed. CRC Press, Boca Raton, FL

[4] Lee KYM, Paterson A, Piggott JR, Richardson GD (2001) Origins of Flavour in Whiskies and a Revised Flavour Wheel: a Review. J Inst Brew 107:287–313. doi: 10.1002/j.2050-0416.2001.tb00099.x