Analysis of Feeding America: the Historic American Cookbook Project

By Austin Hsu

Abstract

Food is one of the most important parts of people's lives. Jobs are created for the sole purpose of developing food further. Food is always evolving and improving. New recipes are being created daily. Cultures are mixing in order to create more creative and different dishes and flavors. To understand a certain culture, taking a look at their diet is a good example of their history and their culture from long ago. This paper seeks to gain a better understanding of the most popular ingredients of specific historical cookbooks to better understand the culture then compared to culture now.

1 Introduction

The purpose of this report is to discover the most popular ingredients throughout the multiple sub-corporas of the “Feeding America: The Historic American Cookbook Dataset.” Of this sub-corpora, we looked at some of the most popular cuisines: the Chinese-Japanese Cookbook, French Cookbook, Italian Cookbook and the Swedish Cookbook to determine what each of the cuisine’s most popular ingredients would be and the ingredients they have in common. These specific sub-corpora gave insight into the most popular cuisines’ ingredients and flavor. Using this insight, we can understand the history of their cuisine and take a look at the evolution of their cuisine through comparing their most popular ingredients to their respective modern cuisines.

In this report, I use the most common tokens and frequency tables for the listed most popular cuisines. I used the wordVectors package to help with discovering additional ingredients in the food groups that I was trying to find more information about.

The rest of the paper is organized as follows. Section 2 covers background and related work. Section 3 outlines the methods used in the report. Section 4 details the results. Section 5 concludes the findings and discoveries made through working on the project.

2 Data

Looking at the dataset more closely, the cookbooks are outdated. Though there are some cuisines that have not changed their traditions and their core values in the cuisine, there are many cuisines that have transformed their cuisines in the past couple of years. Analyzing the dataset, it is hard to confirm accuracy and the credibility of the cookbooks as they have also been edited. The range of the ingredients is also very large. Though the cookbooks may not directly represent what the cuisine is currently like, we can explore the idea that the cuisine may not have changed too much due to cultural reasons and popular ingredients that are grown in that area.

There was minimal preprocessing for the data. Here we used the text2vec package's tokenizer to generate a vocabulary which would then be put into unigrams. From here, we prune this vocabulary to include tokens that appear more than once.

3 Methods

In our analysis of the Chinese-Japanese Cookbook, French Cookbook, Italian Cookbook and the Swedish Cookbook, we start by looking at the individual sizes of each of the cookbooks and taking the most common tokens. From there, we look at the words that are nouns and try to filter out excess words and filler words. Since our main goal is to look at the ingredients, we will be trying to see if the ingredients are very common or not. From looking at these specified cookbooks, we learned that most of the cuisines involved water, butter, sauce, salt, sugar, eggs and peppers. To my surprise, there are no meats or fruits and only one vegetable in the most common Nouns altogether of these cookbooks. As seen in figure 1, there are no meats or fruits and only one vegetable listed. From the figure, we can infer that most of the dishes put an emphasis on ways to cook that do not contain certain proteins or ingredients. Water, butter, sauce, salt and sugar can be placed in a variety of dishes while specific proteins and specific vegetables do not have to be in most dishes.

Most Common Nouns Altogether

feature	frequency	rank	docfreq	group	stringAsFactors
13water_nn	993	13	4	all	FALSE
19butter_nn	816	19	3	all	FALSE
25sauce_nn	695	25	4	all	FALSE
27salt_nn	653	27	4	all	FALSE
32sugar_nn	522	32	4	all	FALSE
33eggs_nns	508	33	4	all	FALSE
35pepper_nn	469	35	4	all	FALSE
38pan_nn	448	38	4	all	FALSE
39minutes_nns	445	39	4	all	FALSE
40flour_nn	439	40	4	all	FALSE
46time_nn	385	46	4	all	FALSE
55preparation_nn	354	55	2	all	FALSE
56no._nn	351	56	2	all	FALSE
59fire_nn	339	59	4	all	FALSE
62hour_nn	320	62	4	all	FALSE
63sauce_nnp	315	63	4	all	FALSE
65pound_nn	309	65	4	all	FALSE
66boil_nn	306	66	4	all	FALSE
67cream_nn	306	66	4	all	FALSE
69de_nnp	301	69	1	all	FALSE

Figure 1: Most Common Nouns from the Chinese-Japanese Cookbook, the French Cookbook, the Italian Cookbook, and the Swedish Cookbook

Looking at the most similar words in the categories of fish, meat, greens and fruit, we can also find a connection between the most similar words in the cuisines to what the culture mainly ate. Looking at the similar words of fish, the Chinese-Japanese Cookbook and looking at the current Chinese/Japanese culture, we can visually see that there were not many fish that were listed in the similar words chart. In the similar words chart as shown in Figure 2, there does not seem to be that many words that are listed as fish; Only salmon, trout, and shad were listed. Observing the current Japanese culture, the most popular fishes used in their culture are Salmon, Bluefin Tuna, Sanma, Bonito, Willow, Japanese Amber, Snapper, Horsetail, Eel, and Horse Mackerel. Comparing the most popular fishes to what was listed on the most similar words chart, we can see that the Japanese cuisine/culture has changed dramatically. The diet consists of more fish and seafood. Taking a look into their current culture, this is definitely true. Japanese food culture puts a large emphasis on fish and seafood.

word <chr>	similarity to "fish" <dbl>
fish	1.0000000
meat	0.6905519
any	0.5646684
which	0.5575222
kettle	0.5545952
the	0.5492648
of	0.5226115
or	0.5177944
other	0.5116724
bouillon	0.4978851

1-10 of 20 rows Previous 1 2 Next

Figure 2: Similarity Chart to Fish

Looking at the most similar words for meat, we can see that many of the cultures listed have used meat for a long time. As we can see in Figure 3, a similarity chart for “meat”, the most similar meat to “meat” was chicken, beef, veal, fish and pork. It seems that there is a much less emphasis on meat and more emphasis on other things (as we will see later that there is a large emphasis on greens). Looking at this similarity chart, there is not much variety in the meats that are consumed. With fish being its own category, the only meats that could be potentially consumed would be chicken, beef, veal and pork. Looking at the modern Italian Culture, cured pork is one of the staples of Italian cuisine. Each region in Italy has their own custom of curing pork. We can see that even Italian cuisine has changed with the evolution of new ideas for cooking and, in this case, a new way of preserving meats and food. Through discovering a new way of processing, preserving and aging meat, a whole culture around a specific meat was created. Looking at modern French cuisine, in the late 1800s, a new staple dish that the French have been using for centuries are snails. Escargot, the French word for snails, have also impacted French cuisine and culture. It seems that as time continues, cultures expand their knowledge and curiosity in consuming meats/fish.

word <chr>	similarity to model[[c("beef", "fish", "fat", "chicken", "veal", "gravy")]] <dbl>
chicken	0.7727141
beef	0.7155489
veal	0.7143830
fat	0.6788609
meat	0.6671502
fish	0.6541123
pork	0.6383918
or	0.6344483
soup	0.6330915
the	0.5542670

1-10 of 10 rows

Figure 3: Similarity Chart to Meat given a List to Search for Multiple Words

Looking at the generic words of “green” to specifically look at vegetables, it seems that a lot of cultures were built around the vegetables/greens that they could farm. In Figure 4, the similarity data frame for “green”, the most similar words to green were mostly vegetables rather than possible filler words, colors, cooking techniques compared to the other similarity data frames. It seems that vegetables and greens were a staple in some of these cuisines/cultures. Currently, China is the number one consumer for peas. We could infer that China has incorporated peas into many authentic dishes that have stayed as a main ingredient in recipes over time. Unlike meats and fish, vegetables/greens have stayed as a main ingredient not just in China, but across many cultures. The most popular consumers for parsley, as shown as the 6th most similar green in the chart, is consumed across all of Europe and Asia. China is also currently the number one

consumer for carrots. Things like onions are also currently incorporated into many French dishes. It seems that generally vegetables have lasted throughout generations in a Countries' culture and cuisine. Even though there were new ways in discovering how to process, preserve and age meats, vegetables have stayed as a very large part of each culture.

word <chr>	similarity to "green" <dbl>
green	1.0000000
peas	0.7475645
peppers	0.5068135
vegetables	0.4847216
white	0.4674940
parsley	0.4670369
carrots	0.4547466
onion	0.4466577
beans	0.4453454
leaves	0.4311976

1-10 of 20 rows Previous 1 2 Next

Figure 4: Similarity Chart to Green

Looking at the last Similarity Chart for "fruit" the cuisines in the cookbooks we looked at didn't incorporate many fruits at all in their cuisines. As shown in Figure 5 for the similarities to "fruit", none of the words in the chart were associated with a fruit. We could think of some theories as to why possibly no fruits came up in the similarity chart. Some cultures didn't use fruits in their recipes but mainly for snacking/consuming. There could have been environments where only vegetables were easier to grow than growing fruits. Vegetables could be grown in mass compared to fruit. There could be endless possibilities as to why there were no fruits in recipes before, but looking at modern cuisine of these 4 countries' cuisines/recipes, there is a large presence of different fruits. One of the most famous desserts in the world was created with the idea of taking the tartness and freshness of fruit and adding it to pastry: the fruit tart! The fruit tart is a staple in French cuisine that came in the late 1800s. Another theory that we could infer from looking at the fruit tart was that fruit could mainly be used in desserts, not specifically in cooking. This could be the reason why there is an absence of fruit in the cookbooks. The cookbooks could be solely focused on cooking appetizers and entrees and not specifically desserts and sweet flavors.

word <chr>	similarity to "fruit" <dbl>
fruit	1.0000000
game	0.5514737
breast	0.4777442
leg	0.4695667
can	0.4629237
sausages	0.4585919
saddle	0.4513914
kidney	0.4487888
peck	0.4264544
considerable	0.4226250

1-10 of 20 rows Previous 1 2 Next

Figure 5: Similarity Chart to Fruit

Using PCA (Principal Component Analysis), we can visually see the similarity calculations in a graphic form. Through PCA, we can better see the retaining trends and patterns that can be hard to see in the Similarity charts. In some of these PCA graphs, we can see that some of the words that we were looking for are very close to other words that we saw in the charts. The closer a word is to our "target" word, the more similar the word is.

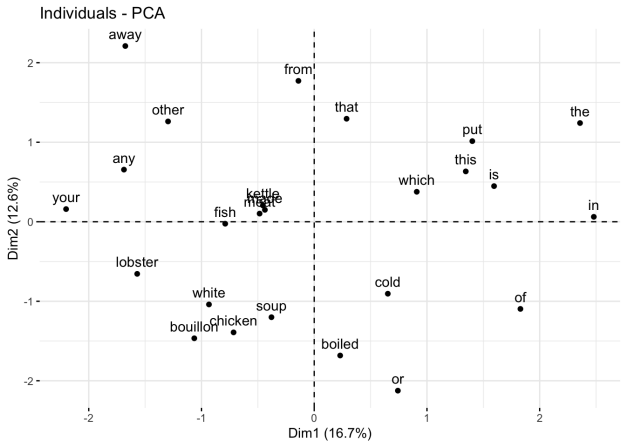


Figure 6: PCA for Fish Data

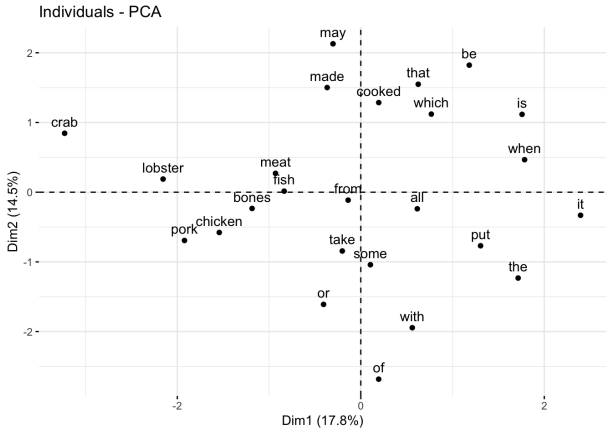
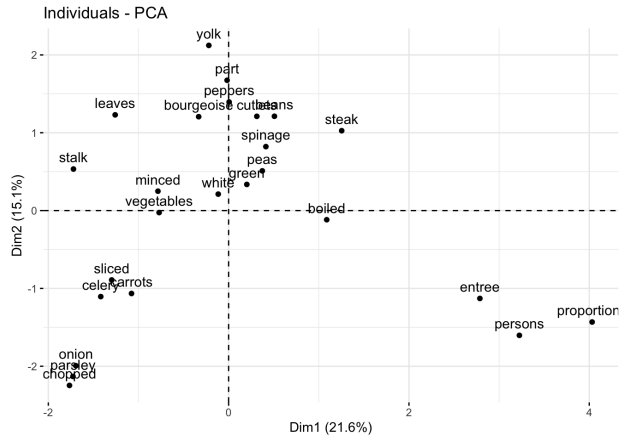


Figure 7: PCA for Meat Data



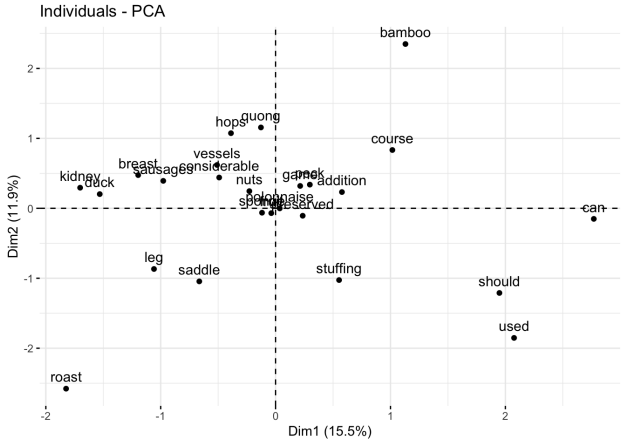


Figure 8: PCA for Green Data

Figure 9: PCA for Fruit Data

Using a PCA Biplot on tastes, we can also see the tastes of the 4 cookbooks using a calculation of Cosine Similarity. In Figure 10, the Biplot of Tastes, we can see that we only see four of the five tastes. Using a specified list for tastes and using these tests to create clusters, we can see that the two main taste categories are salty and bitter. We can also see that the savory taste is missing. The term Umami, or in other areas known as savory, was coined by the Japanese in 1908. The savory taste actually compliments the idea that we saw before with the discovery of more seafood and different ways of making/consuming meat. The savory taste is mainly found in shellfish and cured meats. As we mentioned in the previous Similarity Chart analysis, these ideas were brought into effect later. This could explain the missing portions of savory in the Biplot. Looking more closely at the Biplot, the sour and sweet components are very little. The sweet component could be explained by the lack of fruits in the cookbooks. The sour component is much harder to explain though. Sour food could be potentially not as well liked than other tastes.

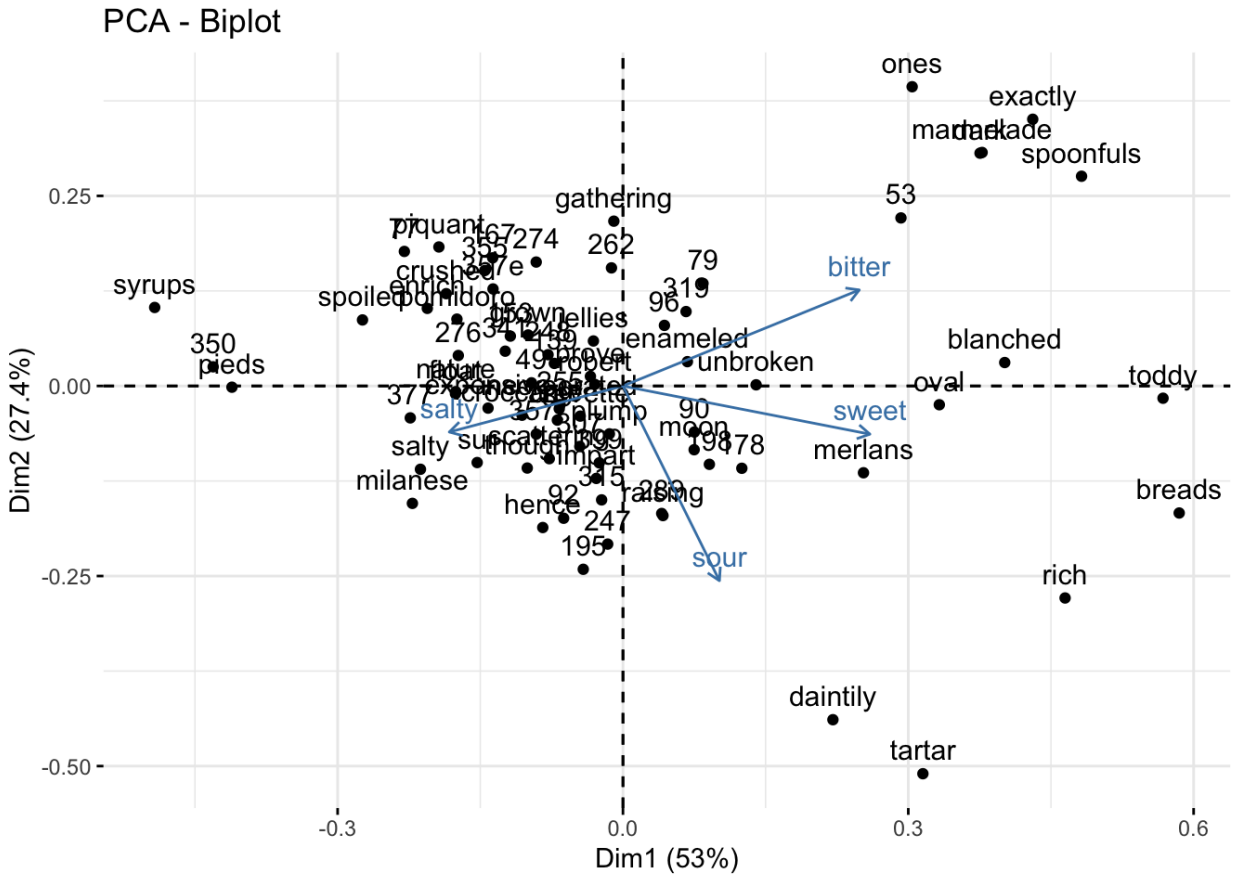


Figure 10:Biplot of Tastes

4 Results

In our analysis, we saw that most of the cuisine/recipes are outdated. For the most common keywords and Similarity Chart in our analysis, we saw that some of the cuisine was outdated comparing the cookbook to modern cooking. We did see that the most common keywords overall in the 4 cookbooks were butter and water. In the Fish/Seafood Similarity Chart, we saw that many of the cuisines didn't find seafood/use fish that much. Looking at the cuisines now, many of the cuisines, especially Japanese cuisine, has seafood/fish as a backbone of their culture and cuisine. In the Meat Similarity Chart, we saw that many of the cuisines didn't have as much variety in their meats as modern cuisine has now. Over time many of these cultures began understanding different ways of processing, preserving and aging meat. This allowed for a more depth of flavor and to allow cuisine to develop. In the Vegetable/Green Similarity Chart, we saw that the chart was filled with vegetables. This tells us that vegetables were and still are the cornerstone of many cuisines and cultures. Having a good agricultural background allows for the growth of cuisine, culture and society. We see that in modern cuisine, many of the vegetables seen in the similarity chart are still highly consumed by these same countries. It seemed that time allowed the vegetables to stay in the culture and could possibly be used in different ways. In the Fruit Similarity Chart, we saw that the chart wasn't filled with many fruits. This led me to believe that there weren't many fruits used in cooking in the past. Most of the fruits that were used in cuisine were mainly for desserts. These findings also agree with the Biplot of tastes. Most of the tastes from each culture were salty and bitter, while there weren't many sour or sweet foods. Overall, it seemed that cuisines continued to evolve through curiosity and building a very good foundation of food.

5 Discussion

Through our analysis, we discovered a little bit of the differences between the historical cookbooks versus the modern cuisine of those cultures. We saw that China/Japan changed their use of fish tremendously. We saw that China also continued to use the same vegetables as a staple of their cuisine until modern times. We saw that the French cuisine had different flavors. Though the French cuisine didn't have the most similar cuisine to other cuisines, we see in modern day that French cuisine has included some of the things that were missing in the charts. French cuisine contains more variety of meats, like escargot, and utilizes the sweetness of fruits, like fruit tarts. The Swedish cuisine did not shine as the other Cuisines did. However, through more research, the use of pickling and different seafood were apparent. Swedish cuisine also contained a considerable amount of potatoes and flour. Italian cuisine as we saw changed a ton throughout the centuries. Now having different ways of processing, preserving and aging meat in different local regions, Italian cuisine has a lot more variety now then it did before. It seemed

like over time, cultures and cuisines developed. A possible question we could ask is by how much has each cuisine developed. Through more in depth research we can see when, for example, the aging process for meats started in Italy which probably gave way to a huge culture shift. Or maybe looking at when Japan started to put a large emphasis on fish/seafood. These are both culture shifts worth looking into.

References

Most popular fish consumed in Japan -
https://skdesu.com/en/types-of-fish-most-consumed-in-japan/

Most popular meat products in Italy -
https://www.tasteatlas.com/most-popular-meat-products-in-italy

Feeding America: the Historic American Cookbook Project -
https://d.lib.msu.edu/islandora/search?collection_page=fa&type=dismax&f%5B0%5D=RELS_EXT_isMemberOfCollection_uri_ms%3A%22info%3Afedora/fa%3Aroot%22&f%5B1%5D=collection%3A%22Feeding%5C%20America%5C%3A%5C%20the%5C%20Historic%5C%20American%5C%20Cookbook%5C%20Project%22&sort=fgs_label_s%20asc&islandora_solr_search_navigation=0

Code Appendix

````
---
title: 'Coffee Break Experiment #2'
author: "Austin Hsu"
output:
  html_document:
    df_print: paged
  pdf_document:
    fig_caption: yes
    number_sections: yes
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(
  collapse = TRUE,
  comment = "#>",
  tidy.opts=list(width.cutoff=70), # this last bit auto-wraps code and comments so the don't run off the page, but you need to have formatR installed
  tidy=TRUE
)
```

```{r libraries, message = FALSE, error=FALSE, warning=FALSE, include = FALSE}
library(cmu.textstat)
library(tidyverse)
library(quanteda)
library(quanteda.textstats)
library(udpipe)
library(future.apply)
```

```{r}
cookbook_files <- list.files("cookbook_corpus",
                      full.names = T, pattern = "*.txt", recursive = T)

cookbooks <- cookbook_files[str_detect(cookbook_files, "chin|ital|fran|swed")]

cookbooks_txt <- readtext::readtext(cookbooks)
cookbooks_split <- split(cookbooks_txt, seq(1, nrow(cookbooks_txt), by = 4))

```

```{r}
cookbooks_corpus <- corpus(cookbooks_txt)
knitr::kable(head(cookbooks_corpus %>% summary()), caption = "Summaries of Chinese, French, Italian and Swedish Cookbooks")
```

```{r}
ncores <- 4L
plan(multisession, workers = ncores)
````
````

```
annotate_splits <- function(corpus_text) {
  ud_model <- udpipe_load_model("english-ewt-ud-2.5-191206.udpipe")
  x <- data.table::as.data.table(udpipe_annotate(ud_model, x = corpus_text$text,
                                       doc_id = corpus_text$doc_id))
  return(x)
}
```

```{r annotate}
annotation <- future_lapply(cookbooks_split, annotate_splits, future.seed = T)
```

```{r to_tokens, warning=FALSE, message=FALSE}
annotation <- data.table::rbindlist(annotation)

annotation <- annotation %>%
  select(doc_id, sentence_id, token_id, token, lemma, upos, xpos, head_token_id, dep_rel) %>%
  rename(pos = upos, tag = xpos)

annotation <- structure(annotation, class = c("spacyr_parsed", "data.frame"))

cookbook_tkns <- as.tokens(annotation, include_pos = "tag", concatenator = "_")
```

```{r}
doc_categories <- names(cookbook_tkns) %>%
  data.frame(cookbook = . ) %>% mutate(cookbook = str_extract(cookbook, "^[a-z]+"))
docvars(cookbook_tkns) <- doc_categories
```

```{r dfm}
cookbook_dfm <- cookbook_tkns %>%
  tokens_select("^.*[a-zA-Z0-9]+.*_[a-z]", selection = "keep", valuetype = "regex", case_insensitive = T) %>%
  dfm()

chin_dfm <- dfm_subset(cookbook_dfm, cookbook == "chin") %>%
  dfm_trim(min_temfreq = 1)
fran_dfm <- dfm_subset(cookbook_dfm, cookbook == "fran") %>%
  dfm_trim(min_temfreq = 1)
ital_dfm <- dfm_subset(cookbook_dfm, cookbook == "ital") %>%
  dfm_trim(min_temfreq = 1)
swed_dfm <- dfm_subset(cookbook_dfm, cookbook == "swed") %>%
  dfm_trim(min_temfreq = 1)
```

```{r count_table}
cb_corpus_comp <- ntoken(cookbook_dfm) %>%
                data.frame(Tokens = .) %>%
                rownames_to_column("Cookbook") %>%
                mutate(Cookbook = str_extract (Cookbook, "^[a-z]+")) %>%
                group_by(Cookbook) %>%
                summarize(Tokens = sum(Tokens)) %>%
                janitor::adorn_totals()
```

```
kableExtra::kbl(cb_corpus_comp,
            caption = "Corpus Composition with Chinese-Japanese, French, Italian and Swedish Cookbooks",
          booktabs = T, linesep = "") %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

```{r}

cb_freq <- textstat_frequency(cookbook_dfm) %>%
  data.frame(stringAsFactors = F) %>%
  filter(str_detect(feature, '_nn'))

knitr::kable(cb_freq[1:20,], caption = "Most Common Nouns Altogether")

chin_freq <- textstat_frequency(chin_dfm) %>%
  data.frame(stringAsFactors = F) %>%
  filter(str_detect(feature, '_nn'))
knitr::kable(chin_freq[1:20,], caption = "Most Common Nouns in the Chinese-Japanese Cookbook")

fran_freq <- textstat_frequency(fran_dfm) %>%
  data.frame(stringAsFactors = F) %>%
  filter(str_detect(feature, '_nn'))
knitr::kable(fran_freq[1:20,], caption = "Most Common Nouns in the French Cookbook")

ital_freq <- textstat_frequency(ital_dfm) %>%
  data.frame(stringAsFactors = F) %>%
  filter(str_detect(feature, '_nn'))
knitr::kable(ital_freq[1:20,], caption = "Most Common Nouns in the Italian Cookbook")

swed_freq <- textstat_frequency(swed_dfm) %>%
  data.frame(stringAsFactors = F) %>%
  filter(str_detect(feature, '_nn'))
knitr::kable(swed_freq[1:20,], caption = "Most Common Nouns in the Swedish Cookbook")

```

```{r keywords}
ci_kt <- keyness_table(chin_dfm, ital_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")

kableExtra::kbl(head(ci_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with Chinese-Japanese versus Italian
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()

cf_kt <- keyness_table(chin_dfm, fran_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")

kableExtra::kbl(head(cf_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with Chinese-Japanese versus French
                  as the Target", booktabs = T, linesep = "",
```

```
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()

cs_kt <- keyness_table(chin_dfm, swed_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")

kableExtra::kbl(head(cs_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with Chinese-Japanese versus Swedish
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()

```

```{r}
ic_kt <- keyness_table(ital_dfm, chin_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")
kableExtra::kbl(head(ic_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with Italian vs Chinese-Japanese
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()

if_kt <- keyness_table(ital_dfm, fran_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")
kableExtra::kbl(head(if_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with Italian versus French
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()

is_kt <- keyness_table(ital_dfm, chin_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")
kableExtra::kbl(head(is_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with Italian versus Swedish
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```



```{r}
fc_kt <- keyness_table(fran_dfm, chin_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")
kableExtra::kbl(head(fc_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with French vs Chinese-Japanese
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
```

```
  kableExtra::kable_classic()

fi_kt <- keyness_table(fran_dfm, ital_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")
kableExtra::kbl(head(fi_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with French versus Italian
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()

fs_kt <- keyness_table(fran_dfm, swed_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")
kableExtra::kbl(head(fs_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with French versus Swedish
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

```{r}
sc_kt <- keyness_table(swed_dfm, chin_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")
kableExtra::kbl(head(sc_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with Swedish vs Chinese-Japanese
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()

sf_kt <- keyness_table(swed_dfm, fran_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")
kableExtra::kbl(head(sf_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with Swedish versus French
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()

si_kt <- keyness_table(swed_dfm, ital_dfm) %>%
  separate(col = Token, into = c("Token", "Tag"), sep = "_")
kableExtra::kbl(head(si_kt %>% filter(Tag == c("nn", "nns")), n = 10),
            caption = "Highest Keyness Tokens in Cookbooks, with Swedish versus Italian
                  as the Target", booktabs = T, linesep = "",
            digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```
```{r}
library(text2vec)
library(tidyverse)
library(wordVectors)
library(factoextra)
```

```
library(tsne)
cook_txt <- list.files("cookbook_corpus", full.names = T)

cookbooks <- cookbook_files[str_detect(cook_txt, "chin|ital|fran|swed")]

cookbooks_txt <- readtext::readtext(cookbooks)

```

```{r}
cook_filtered <- cookbooks_txt %>% filter(doc_id == "fran.txt" | doc_id == "chin.txt" | doc_id == "ital.txt" | doc_id == "swed.txt")
cook_filtered
cook_tks <- itoken(cook_filtered$text,
                   preprocessor = tolower,
                   tokenizer = word_tokenizer,
                   ids = cook_filtered$doc_id,
                   progressbar = TRUE)

##cook_tks <- itoken(cookbooks_txt$text,
#                   preprocessor = tolower,
#                   tokenizer = word_tokenizer,
#                   ids = cook_txt$doc_id,
#                   progressbar = TRUE)

cook_vocab <- create_vocabulary(cook_tks)
cook_vocab <- prune_vocabulary(cook_vocab, term_count_min = 2L)
cook_vocab %>% head()
vectorizer <- vocab_vectorizer(cook_vocab)
tcm <- create_tcm(cook_tks, vectorizer, skip_grams_window = 5L)
glove <- GlobalVectors$new(rank = 50, x_max = 10)
```

```{r}
cook_main <- glove$fit_transform(tcm, n_iter = 30, convergence_tol = 0.001)
cook_context <- glove$components
cook_vectors <-  cook_main + t(cook_context)
```

```{r}
fishy <- cook_vectors["fish", , drop = FALSE]
fishy <- sim2(x = cook_vectors, y = fishy, method = "cosine", norm = "l2")
head(sort(fishy[,1], decreasing = TRUE), 5)

cook_vectors <- as.matrix(cook_main + t(cook_context)) %>%
  as.VectorSpaceModel()

model <- as.matrix(cook_main + t(cook_context)) %>%
  as.VectorSpaceModel()
```

```{r}
cook_vectors %>% closest_to("fish", 20)
cook_vectors %>% closest_to("meat", 20)
```

```
cook_vectors %>% closest_to("green", 20)
cook_vectors %>% closest_to("fruit", 20)
```
```{r}
fish_neighbors <- model %>% closest_to("fish", 25)
fishy <- model[[fish_neighbors$word, average=F]]
prcomp(fishy@.Data) %>% fviz_pca_ind()

model %>%
  closest_to(model[[c("fish","salmon","trout","shad","flounder","carp","roe","eels")]],10) %>%
  as_tibble()

```
```{r}
meat_neighbors <- model %>% closest_to("meat", 25)
meaty <- model[[meat_neighbors$word, average=F]]
prcomp(meaty@.Data) %>% fviz_pca_ind()

model %>%
  closest_to(model[[c("beef","fish","fat","chicken","veal","gravy")]],10) %>%
  as_tibble()

```
```{r}
green_neighbors <- model %>% closest_to("green", 25)
greeny <- model[[green_neighbors$word, average=F]]
prcomp(greeny@.Data) %>% fviz_pca_ind()

green_neighbors

model %>%
  closest_to(model[[c("radish","leaves","cucumbers","beans","onions")]],10) %>%
  as_tibble()
```

```{r}
fruit_neighbors <- model %>% closest_to("fruit", 25)
fruity <- model[[fruit_neighbors$word, average=F]]
prcomp(fruity@.Data) %>% fviz_pca_ind()

fruit_neighbors

model %>%
  closest_to(model[[c("plums","damsons","marmalade","preserves","cherries")]],10) %>%
  as_tibble()
```
```{r}
dairy_neighbors <- model %>% closest_to("dairy", 25)
dairy <- model[[dairy_neighbors$word, average=F]]
prcomp(dairy@.Data) %>% fviz_pca_ind()

dairy_neighbors

model %>%
```

```
  closest_to(model[[c("beef","fish","fat","chicken","veal","gravy")]],10) %>%
  as_tibble()
```

```{r}
tastes <- model[[c("butter","water"), average=F]]
butter_and_water <- model[1:3000,] %>% cosineSimilarity(tastes)
butter_and_water <- butter_and_water[
   rank(-butter_and_water[,1]) < 20 |
   rank(-butter_and_water[,2]) < 20,] %>% data.frame()
ggplot(butter_and_water, aes(x = butter, y = water)) +
  geom_text(label = rownames(butter_and_water)) +
  theme_classic()


tastes <- model[[c("sugar","salt"), average=F]]
sweet_and_saltiness <- model[1:3000,] %>% cosineSimilarity(tastes)
sweet_and_saltiness <- sweet_and_saltiness[
   rank(-sweet_and_saltiness[,1]) < 20 |
   rank(-sweet_and_saltiness[,2]) < 20,] %>% data.frame()
ggplot(sweet_and_saltiness, aes(x = salt, y = sugar)) +
  geom_text(label = rownames(sweet_and_saltiness)) +
  theme_classic()

```

```{r}
tastes <- model[[c("sweet","salty","savory","bitter","sour"), average=F]]
common_similarities_tastes <- model[1:3000,] %>% cosineSimilarity(tastes)

high_similarities_to_tastes <- common_similarities_tastes[rank(-apply(common_similarities_tastes,1, max)) < 75,]

high_similarities_to_tastes %>% prcomp %>% fviz_pca_biplot()
```