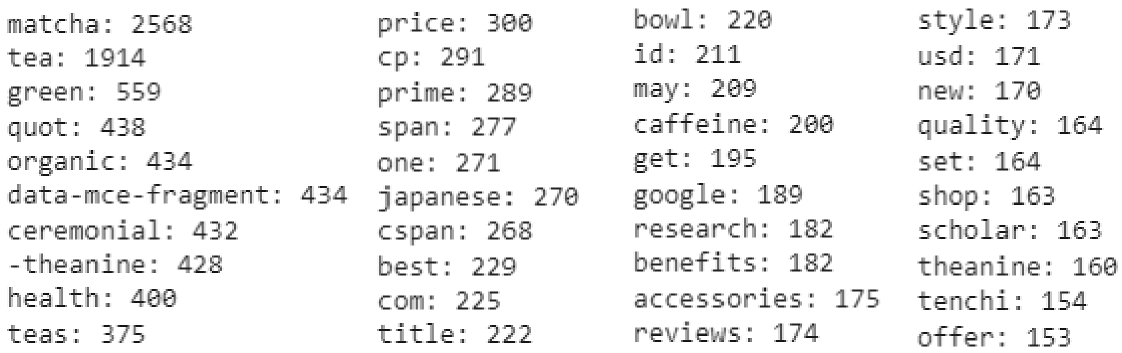
Jimmy Harvin & Charles Wallis
3/11/2023

Finding & Building a Corpus using a WebCrawler

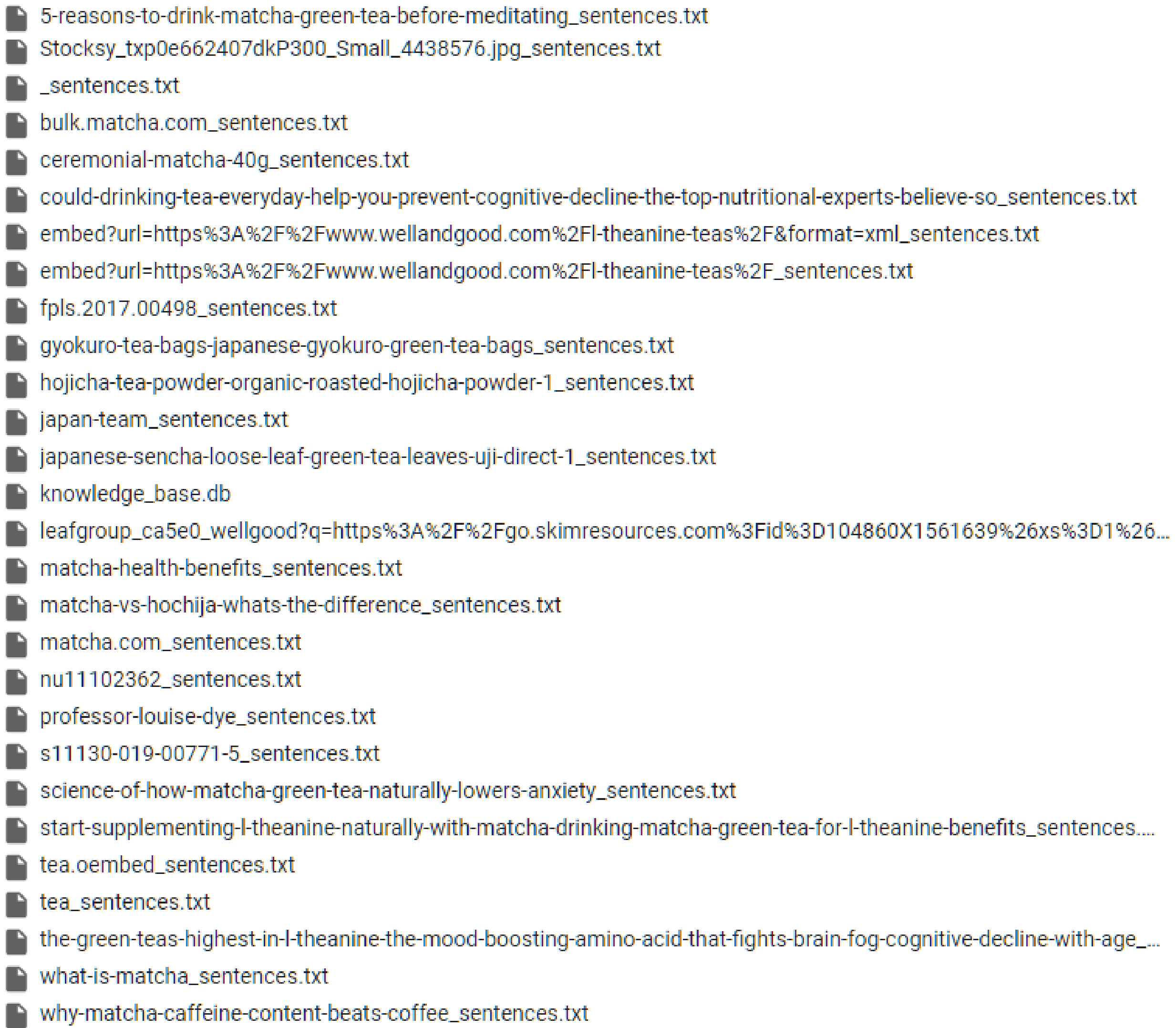
To build a corpus, we developed a web crawler that takes a URL as input and searches for relevant URLs within the original domain. The crawler loops through the collected URLs, scraping text from each page and storing it in a separate file and dictionary. The files are then cleaned to extract only the useful information, and a function is used to extract important terms based on term frequency. We selected the top 10 terms we deemed most useful and important for this domain, which include [tea, theanine, matcha, green tea, black tea, oolong tea, herbal tea, caffeine, health, and flavors]. Finally, we created a knowledge base by identifying sentences that discuss these terms.

Here is what our top 40 most frequent terms look like:



```
matcha: 2568              price: 300      bowl: 220          style: 173
tea: 1914                 cp: 291         id: 211            usd: 171
green: 559                prime: 289      may: 209           new: 170
quot: 438                 span: 277       caffeine: 200      quality: 164
organic: 434              one: 271        get: 195           set: 164
data-mce-fragment: 434    japanese: 270   google: 189        shop: 163
ceremonial: 432           cspan: 268      research: 182      scholar: 163
-theanine: 428            best: 229       benefits: 182      theanine: 160
health: 400               com: 225        accessories: 175   tenchi: 154
teas: 375                 title: 222      reviews: 174       offer: 153
```

Here is what our knowledge base looks like:



- 5-reasons-to-drink-matcha-green-tea-before-meditating_sentences.txt
- Stocksy_txp0e662407dkP300_Small_4438576.jpg_sentences.txt
- _sentences.txt
- bulk.matcha.com_sentences.txt
- ceremonial-matcha-40g_sentences.txt
- could-drinking-tea-everyday-help-you-prevent-cognitive-decline-the-top-nutritional-experts-believe-so_sentences.txt
- embed?url=https%3A%2F%2Fwww.wellandgood.com%2Fl-theanine-teas%2F&format=xml_sentences.txt
- embed?url=https%3A%2F%2Fwww.wellandgood.com%2Fl-theanine-teas%2F_sentences.txt
- fpls.2017.00498_sentences.txt
- gyokuro-tea-bags-japanese-gyokuro-green-tea-bags_sentences.txt
- hojicha-tea-powder-organic-roasted-hojicha-powder-1_sentences.txt
- japan-team_sentences.txt
- japanese-sencha-loose-leaf-green-tea-leaves-uji-direct-1_sentences.txt
- knowledge_base.db
- leafgroup_ca5e0_wellgood?q=https%3A%2F%2Fgo.skimresources.com%3Fid%3D104860X1561639%26xs%3D1%26...
- matcha-health-benefits_sentences.txt
- matcha-vs-hochija-whats-the-difference_sentences.txt
- matcha.com_sentences.txt
- nu11102362_sentences.txt
- professor-louise-dye_sentences.txt
- s11130-019-00771-5_sentences.txt
- science-of-how-matcha-green-tea-naturally-lowers-anxiety_sentences.txt
- start-supplementing-l-theanine-naturally-with-matcha-drinking-matcha-green-tea-for-l-theanine-benefits_sentences....
- tea.oembed_sentences.txt
- tea_sentences.txt
- the-green-teas-highest-in-l-theanine-the-mood-boosting-amino-acid-that-fights-brain-fog-cognitive-decline-with-age_...
- what-is-matcha_sentences.txt
- why-matcha-caffeine-content-beats-coffee_sentences.txt

Here is a sample dialog that we'd like to create with a chatbot based on our knowledge base

Case 1

User: Hello, could you tell me something interesting about tea?
Bot: Hi, what topics of tea are you interested in?
User: What are some health benefits of tea?
Bot: Drinking tea has a variety of health benefits, such as reducing the risk of heart disease, stroke, and certain types of cancer. It also contains antioxidants and other compounds that can boost the immune system and improve brain function.

Case 2

User: Can you tell me more about the different types of tea?
Bot: There are many different types of tea, such as green tea, black tea, oolong tea, and herbal tea. Each type has its own unique flavor and health benefits. Would you like to learn more?
User: Yes please
Bot:

- Green tea is known for its light, fresh taste and high levels of antioxidants.
- Black tea has a bolder, more robust flavor and contains more caffeine than other types of tea.
- Oolong tea has a floral, fruity flavor and contains antioxidants that may help prevent cell damage.
- Herbal tea is not technically tea, as it is made from herbs and not tea leaves. It comes in a variety of flavors and has different health benefits depending on the herbs used