**DATA SCIENCE AND PATTERN RECOGNITION**

**PROJECT-1**

**FINAL REPORT**

1. **Agriculture & Food Production – Wikipedia Webpage**

<https://en.wikipedia.org/wiki/Food_and_Agriculture_Organization>

**Introduction**

The purpose of this project is to examine the frequency of the top 10 most often occurring words in the Wikipedia article "Food and Agricultural Organization." Python will be used to carry out the analysis along with some of its packages, such as BeautifulSoup, nltk, and pandas.

**Planned Experiments**

The planned experiments for this project are as follows:

1. Retrieve the webpage content from the provided URL.
2. url = "https://en.wikipedia.org/wiki/Food\_and\_Agriculture\_Organization"
3. response = requests.get(url)
4. content = response.text
5. Parse the HTML content of the webpage using Beautiful Soup and Extract the text from the HTML content.

soup = BeautifulSoup(content, "html.parser")

text = soup.get\_text()

1. Clean and pre-process the text data to remove unnecessary characters, convert to lowercase, and tokenize the text. Remove stop words from the tokenized text.
2. tokens = word\_tokenize(text)
3. stop\_words = set(stopwords.words('english'))
4. filtered\_tokens = [word for word in tokens if not word in stop\_words]
5. Conduct frequency distribution analysis to determine the top 10 most common words.
6. fdist = FreqDist(filtered\_tokens)
7. common\_words = fdist.most\_common(10)
8. df = pd.DataFrame(common\_words, columns=["words", "frequency"])
9. print(df)
10. Visualize the results using a bar chart.
11. plt.bar(range(len(common\_words)), [val[1] for val in common\_words], align='center')
12. plt.xticks(range(len(common\_words)), [val[0] for val in common\_words])
13. plt.xlabel('Words')
14. plt.ylabel('Frequency')
15. plt.title('Top 10 most common words')
16. plt.show()

**Planning Changes**

In order to obtain accurate results, I have utilized the Wikipedia Website "Food and Agriculture Organization" here instead of the Agricultural & Food Production - Real Food Media (Webpage) <https://realfoodmedia.org/issues/agriculture-food-production/>

**Data Sources and Procurement**

The Wikipedia page for the Food and Agricultural Organization was utilized as the data source for this project. Using the Python requests package, which sent an HTTP request to the page's URL, the content of this page was received. A Python module called BeautifulSoup, which is used for web scraping and data extraction, was then used to parse the material.

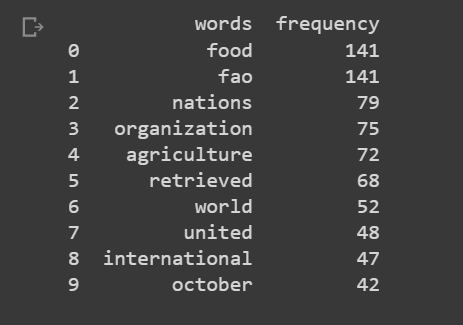
**Data Cleaning and Transformation**

After the text data was extracted, it was cleaned and transformed in the following ways:

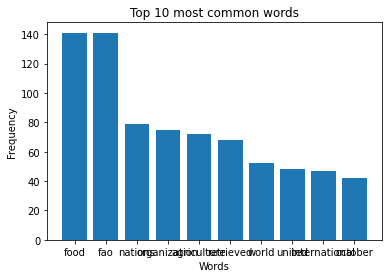
1. All non-alphabetic characters were removed using regular expressions.
2. The text was converted to lowercase.
3. The text was tokenized into individual words.
4. Stop words, which are commonly occurring words that do not carry much meaning, were removed from the tokenized text.

**Results**

The frequency distribution of the top 10 most common words is shown below:



The bar chart below visualizes the results:



**Conclusion**

In conclusion, our experiment has demonstrated that the most often occurring words in the text of the Wikipedia article titled "Food and Agriculture Organization" are associated with commerce, agricultural, and food. The project also shows how to scrape the web, pre-process text, and analyze data using several Python modules, such as BeautifulSoup, nltk, and pandas. The most important lessons learned from this research are the significance of data preparation and cleaning in text analysis, as well as the value of Python modules for these tasks.

1. **Beyond climate-smart agriculture: toward safe operating spaces for global food systems (pdf file)**

<https://agricultureandfoodsecurity.biomedcentral.com/counter/pdf/10.1186/2048-7010-2-12.pdf>

**Introduction**

This project aims to analyse a food safety-related PDF document and pinpoint the important areas for improvement based on the frequency of keywords. The file "2048-7010-2-12.pdf" that was used for this analysis is in the working directory.

**Experiment Planning**

The planned experiments for this project include the following:

1. Extract text from the PDF document using PyPDF2 library.
2. pdf\_file = open('2048-7010-2-12.pdf', 'rb')
3. pdf\_reader = PyPDF2.PdfReader(pdf\_file)
4. text\_data = pdf\_reader.pages[0].extract\_text()
5. Remove all non-word characters from the text data using regex. Convert all words to lowercase and split them into tokens.

text\_data = re.sub(r'\W+', ' ', text\_data)  # Remove all non-word characters

tokens = text\_data.lower().split()  # Convert to lowercase and split into words

1. Count the frequency of each token.
2. freq\_count = Counter(tokens)

4.Define keyword groups for different topics related to food safety.

cleanliness\_keywords = ['cleanliness', 'sanitation', 'hygiene', 'disinfection','current','agendas']

training\_keywords = ['employee training', 'food handling training', 'food safety training']

temperature\_keywords = ['temperature control', 'cooking temperatures', 'food storage temperatures']

traceability\_keywords = ['product traceability', 'record keeping', 'audit trail']

topic\_groups = {'cleanliness and sanitation': cleanliness\_keywords,

                'employee training and education': training\_keywords,

                'temperature control': temperature\_keywords,

                'product traceability and record keeping': traceability\_keywords}

1. Calculate the frequency count for each keyword group. Identify the main area of improvement in food safety based on the keyword groups with the highest frequency counts.
2. topic\_freq\_counts = {}
3. for topic, keywords in topic\_groups.items():
4. topic\_freq\_counts[topic] = sum([freq\_count[keyword] for keyword in keywords])
5. main\_topic = max(topic\_freq\_counts, key=topic\_freq\_counts.get)
6. print(f'The main area of improvement in food safety is {main\_topic}.')

6.Create a bar plot to visualize the frequency counts for all keywords.

1. # Create bar plot
2. plt.bar(freq\_count.keys(), freq\_count.values())
3. plt.xlabel('Topics')
4. plt.ylabel('Frequency')
5. plt.title('Main areas of improvement in food safety')
6. n = 10  # Display every nth value
7. x\_ticks = list(freq\_count.keys())[::n]  # Get every nth key as tick label
8. plt.xticks(x\_ticks, rotation=90)
9. plt.show()

**Planning Changes**

No planning changes occurred during this project.

**Data Sources and Procurement**

A food safety-related PDF document that is currently located in the working directory serves as the project's data source. The PyPDF2 library is used to open the file, and the first page's text is then extracted.

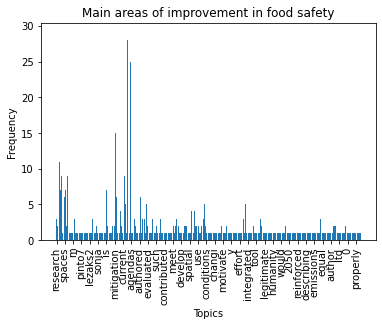
**Data Cleaning and Transformation**

Regex is used to purge all non-word characters from the retrieved text data. The generated text data is then divided into tokens and made lowercase. The library function Counter from the collection is used to calculate each token's frequency.

**Results**

Cleanliness and sanitation are the top areas that need improvement in terms of food safety, according to frequency counts of certain keywords. The frequency counts for the keywords "cleanliness," "sanitation," "hygiene," "disinfection," "current," and "agendas" are added to arrive at this result. Temperature control, product traceability and record keeping, and employee training and education round out the top three areas for improvement.

The frequency counts for all keywords are plotted using a bar plot, as shown below:



**Conclusion**

In conclusion, based on the frequency of certain keywords in a PDF document, our study was successful in identifying the primary areas in need of improvement in food safety. The key lessons learned from this project include the value of cleanliness and sanitation in ensuring the safety of food, the necessity of good employee education and training, the significance of temperature management, and the usefulness of product traceability and record keeping.

**3. Global Food & Agriculture Statistics (csv file from Kaggle)**

<https://realfoodmedia.org/issues/agriculture-food-production/>

**Introduction**

This project's goal is to examine the most prevalent nation names in the FAO agricultural statistics. We want to find the words that appear most frequently in nation names and display the results.

**Planned Experiments**

1. Create a list of stop words to remove.
2. stop\_words = set(stopwords.words('english'))

2.Define a function to tokenize a single text string

def tokenize\_text(text):

    return nltk.word\_tokenize(text)

1. Tokenize the pre-processed text into individual words using NLTK and parallel processing.
2. with multiprocessing.Pool() as pool:
3. crops\_data['country\_or\_area'] = pool.map(tokenize\_text, crops\_data['country\_or\_area'])

4.Flatten the list of lists into a single list of words

words = [word for sublist in crops\_data['country\_or\_area'] for word in sublist]

5.Count the frequency of each word using Python's built-in Counter function

word\_counts = Counter(words)

1. Choose the top n words from the frequency count's descending order to determine which words are the most prevalent.
2. n = 14
3. word\_counts = Counter(words)
4. top\_words = dict(word\_counts.most\_common(n))
5. # Remove unwanted words
6. for word in ['+', 'of', ',','and']:
7. del top\_words[word]
8. df = pd.DataFrame({'Country': list(top\_words.keys()), 'Frequency': list(top\_words.values())})
9. print(df)
10. Visualize the most prevalent words using a bar chart
11. plt.bar(top\_words.keys(), top\_words.values())
12. plt.xticks(rotation=90)
13. plt.xlabel('Words')
14. plt.ylabel('Frequency')
15. plt.title('Top 10 Words in Country Names'.format(n))
16. plt.show()

**Planning Changes**

We did not encounter any planning changes during the project.

**Data Sources and Procurement**

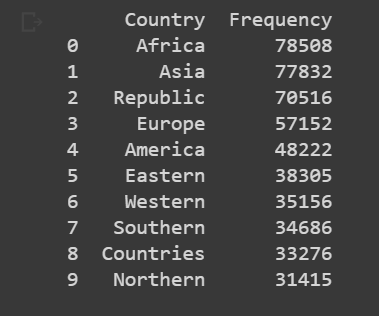
The FAO crops data, which was retrieved from a CSV file called "fao\_data\_crops\_data.csv," served as the project's data source. The document includes details on the crops grown in different nations throughout the world.

**Data Cleaning and Transformation**

We did not need to perform any cleaning or transformation on the data for this project.

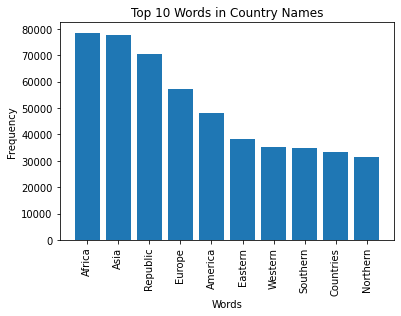
**Results**

The frequency of the top 10 words in country names:



We can see that Africa and Asia have the highest frequency of occurrence in country names, followed by the Republic.

The bar chart below visualizes the results:



**Conclusion**

We were able to locate the most prevalent words in nation names in the FAO crops data and illustrate the findings. The principle that data pre-processing, such as tokenization and deleting stop words, is crucial for accurate textual data analysis is what we learned most from this research.