STAT5342 Project 1

Author: Mingze Xu
UNI: mx2269
Date: 7.8.2023

Introduction:

What is philosophy?
This is a really abstract question, and for most of us, we might not be able to give a proper definition to it. The origin of the term 'philosophy' can be traced back to ancient Greece, around the 6th century BCE. Some people might argue that philosophy is a branch of knowledge that seeks to understand the world and our place in it. This argument is really convincing.
In this **History of Philosophy** project, particularly the dataset **philosophy_data**, 36 world reknowned philosophers from 13 different schools of philosophy and their thounsands of sentences are included. From the dataset, we would manipulate the data and might be able to explore the progress or the development of the subject Philosophy.
More specifically in this project, after researching the history of philosophy and simply oberserving the data, I would focus on how philosophy might have evolved. I categorized these schools into three groups:

- Ancient Greek Philosophy:Stoicism, Aristotle, Plato
- 19th and 20th Century Philosophy: Capitalism, Continental, Feminism, German Idealism, Phenomenology, Communism, Nietzsche
- Modern Philosophy: Analytic, Rationalism, Empiricism

Acient Greek Philosophy and Modern Philosophy are actualy on the two endpoints of this situation, so I will focus on these two groups.
Research question: The research question would be how modern philosophy is different from the Ancient Greek Philosophy, from the aspects of writing language and the central theme.
Hypothesis: Considering the time gap of almost 2700 years, I hypothesize that the writing language would be different because the way people talk and write are evolving. Also, the central themes were also shifted because the ideology and understanding of life are changing in an unpredictable speed.

Setup

importing the packages

In [1]:

```
# All packages used in this project are imported in this cell
import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns
import re
import tqdm
import ast
import nltk
from nltk.corpus import stopwords
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from nltk.probability import FreqDist
```

importing the Dataset and printing the first five rows

In [2]:

```
# The dataset is imported
philosophy_data = pd.read_csv('/Users/xu/Desktop/5243/philosophy_data.csv')

# The first five rows are printed to show an insight into the dataset
philosophy_data.head(5)
```

Out[2]:

	title	author	school	sentence_spacy	sentence_str	original_publication_date	corpus_edition_date	sentence_length	sentence_lowered	tokeni
0	Plato - Complete Works	Plato	plato	What's new, Socrates, to make you leave your ...	What's new, Socrates, to make you leave your ...	-350	1997	125	what's new, socrates, to make you leave your ...	['wha 'socra 'make
1	Plato - Complete Works	Plato	plato	Surely you are not prosecuting anyone before t...	Surely you are not prosecuting anyone before t...	-350	1997	69	surely you are not prosecuting anyone before t...	['surel 'ar 'prosec
2	Plato - Complete Works	Plato	plato	The Athenians do not call this a prosecution b...	The Athenians do not call this a prosecution b...	-350	1997	74	the athenians do not call this a prosecution b...	'ath 'do', 'nc
3	Plato - Complete Works	Plato	plato	What is this you say?	What is this you say?	-350	1997	21	what is this you say?	['w 'thi
4	Plato - Complete Works	Plato	plato	Someone must have indicted you, for you are no...	Someone must have indicted you, for you are no...	-350	1997	101	someone must have indicted you, for you are no...	['so 'must 'i

Data Cleaning and Data Wrangling

In [3]:

```
# The number of distinct authors and their name
author = philosophy_data['author'].unique()

# The number of distinct schools and their name
school = philosophy_data['school'].unique()

print(len(author), author)
print(len(school), school)
```

```
36 ['Plato' 'Aristotle' 'Locke' 'Hume' 'Berkeley' 'Spinoza' 'Leibniz'
 'Descartes' 'Malebranche' 'Russell' 'Moore' 'Wittgenstein' 'Lewis'
 'Quine' 'Popper' 'Kripke' 'Foucault' 'Derrida' 'Deleuze' 'Merleau-Ponty'
 'Husserl' 'Heidegger' 'Kant' 'Fichte' 'Hegel' 'Marx' 'Lenin' 'Smith'
 'Ricardo' 'Keynes' 'Epictetus' 'Marcus Aurelius' 'Nietzsche'
 'Wollstonecraft' 'Beauvoir' 'Davis']
13 ['plato' 'aristotle' 'empiricism' 'rationalism' 'analytic' 'continental'
 'phenomenology' 'german_idealism' 'communism' 'capitalism' 'stoicism'
 'nietzsche' 'feminism']
```

In [4]:

```
# Create dataframe 'schools' sorted the author by their schools
schools = philosophy_data.groupby('school')['author'].agg(lambda x: list(set(x)))
schools = schools.reset_index()
schools.columns = ['school', 'name']
schools['number of philosopher'] = schools['name'].apply(lambda x: len(x))
schools = schools.sort_values(by='number of philosopher', ascending=False)
schools
```

Out[4]:

	school	name	number of philosopher
0	analytic	[Wittgenstein, Moore, Russell, Kripke, Quine, ...	7
11	rationalism	[Leibniz, Descartes, Spinoza, Malebranche]	4
2	capitalism	[Keynes, Smith, Ricardo]	3
4	continental	[Deleuze, Derrida, Foucault]	3
5	empiricism	[Berkeley, Hume, Locke]	3
6	feminism	[Davis, Wollstonecraft, Beauvoir]	3
7	german_idealism	[Fichte, Kant, Hegel]	3
9	phenomenology	[Merleau-Ponty, Heidegger, Husserl]	3
3	communism	[Lenin, Marx]	2
12	stoicism	[Marcus Aurelius, Epictetus]	2
1	aristotle	[Aristotle]	1
8	nietzsche	[Nietzsche]	1
10	plato	[Plato]	1

In [5]:

```
# Ignoring the possible space at the beginning of the sentence
philosophy_data['sentence_str'] = philosophy_data['sentence_str'].str.strip()

# Create a new column 'sentence_split' storing the list of strings that are the words in the sentence
philosophy_data['sentence_split'] = philosophy_data['sentence_str'].str.split(r"[\s']+")

# Create a new column 'word_count' storing the number of words in the sentence
philosophy_data['word_count'] = philosophy_data['sentence_split'].str.len()
```

In [6]:

```
# Dataframe 'author_sentence_length' shows the average length of the sentences of each author
author_sentence_length = philosophy_data.groupby('author')['sentence_length'].mean().sort_values(ascending=False).to_frame()
author_sentence_length
```

Out[6]:

author	sentence_length
Descartes	247.381625
Locke	200.395836
Kant	198.159400
Keynes	196.654060
Wollstonecraft	190.957796
Foucault	189.637467
Ricardo	186.252751
Husserl	185.473703
Smith	185.277944
Lenin	181.423137
Hume	180.192372
Hegel	175.720088
Merleau-Ponty	170.934009
Moore	167.254907
Malebranche	164.434023
Deleuze	163.671850
Leibniz	157.085140
Aristotle	153.224953
Fichte	151.964582
Beauvoir	148.790351
Spinoza	146.544424
Russell	146.296669
Derrida	143.431239
Marx	143.253466
Marcus Aurelius	139.776221
Davis	139.671134
Berkeley	139.653987
Popper	139.545105
Quine	121.643429
Kripke	119.025082
Heidegger	118.541965
Epictetus	118.430341
Nietzsche	116.599867
Plato	114.938018
Lewis	109.717607
Wittgenstein	84.883772

In [7]:

```
# Dataframe 'author_word_count' shows the average number of words in the sentences of each author
author_word_count = philosophy_data.groupby('author')['word_count'].mean().sort_values(ascending=False).to_frame()
author_word_count
```

Out[7]:

author	word_count
Descartes	45.372792
Locke	36.275295
Kant	33.823613
Keynes	33.538259
Wollstonecraft	33.033607
Ricardo	33.007120
Foucault	32.162073
Smith	32.135295
Moore	31.169029
Hume	31.056785
Husserl	30.261930
Merleau-Ponty	29.989726
Hegel	29.740837
Lenin	29.681808
Malebranche	29.571978
Aristotle	28.227639
Leibniz	28.006763
Deleuze	27.135486
Fichte	26.967408
Spinoza	26.394411
Beauvoir	26.183376
Russell	25.788094
Marcus Aurelius	25.745027
Berkeley	25.141185
Marx	24.520498
Derrida	24.146858
Popper	23.370671
Davis	23.267735
Epictetus	22.030960
Plato	21.735052
Nietzsche	20.854296
Quine	20.757087
Kripke	20.728263
Heidegger	20.287683
Lewis	19.110213
Wittgenstein	16.142905

In [8]:

```
# Analytic
analytics = philosophy_data[philosophy_data['school'] == 'analytic']

# Rationalism
rationalisms = philosophy_data[philosophy_data['school'] == 'rationalism']

# Empiricism
empiricisms = philosophy_data[philosophy_data['school'] == 'empiricism']

# Modern
modern = pd.concat([analytics, rationalisms, empiricisms], axis=0)
modern.head()
```

Out[8]:

	title	author	school	sentence_spacy	sentence_str	original_publication_date	corpus_edition_date	sentence_length	sentence_lowered	t
130025	The Analysis Of Mind	Russell	analytic	This book has grown out of an attempt to harmo...	This book has grown out of an attempt to harmo...	1921	2008	217	this book has grown out of an attempt to harmo...	
130026	The Analysis Of Mind	Russell	analytic	On the one hand, many psychologists, especiall...	On the one hand, many psychologists, especiall...	1921	2008	186	on the one hand, many psychologists, especiall...	
130027	The Analysis Of Mind	Russell	analytic	They make psychology increasingly dependent on...	They make psychology increasingly dependent on...	1921	2008	167	they make psychology increasingly dependent on...	'i
130028	The Analysis Of Mind	Russell	analytic	Meanwhile the physicists, especially Einstein ...	Meanwhile the physicists, especially Einstein ...	1921	2008	142	meanwhile the physicists, especially einstein ...	
130029	The Analysis Of Mind	Russell	analytic	Their world consists of events, from which mat...	Their world consists of events, from which mat...	1921	2008	87	their world consists of events, from which mat...	

In [9]:

```
# Stoicism
stoicisms = philosophy_data[philosophy_data['school'] == 'stoicism']

# Aristotle
aristotles = philosophy_data[philosophy_data['school'] == 'aristotle']

# Plato
platos = philosophy_data[philosophy_data['school'] == 'plato']

# Ancient Greek
ancientgreek = pd.concat([stoicisms, aristotles, platos], axis=0)
ancientgreek.head()
```

Out[9]:

	title	author	school	sentence_spacy	sentence_str	original_publication_date	corpus_edition_date	sentence_length	sentence_lowere
326090	Enchiridion	Epictetus	stoicism	There are things which are within our power, a...	There are things which are within our power, a...	125	2014	93	there are thing which are with our power, a
326091	Enchiridion	Epictetus	stoicism	Within our power are opinion, aim, desire, ave...	Within our power are opinion, aim, desire, ave...	125	2014	100	within our powe are opinion, ain desire, ave
326092	Enchiridion	Epictetus	stoicism	Beyond our power are body, property, reputatio...	Beyond our power are body, property, reputatio...	125	2014	117	beyond our powe are body, propert reputatio
326093	Enchiridion	Epictetus	stoicism	Now the things within our power are by nature ...	Now the things within our power are by nature ...	125	2014	144	now the thing within our powe are by nature
326094	Enchiridion	Epictetus	stoicism	Remember, then, that if you attribute freedom ...	Remember, then, that if you attribute freedom ...	125	2014	142	remember, ther that if you attribut freedom

Summary of data cleaning and wrangling

1. Create dataframe **schools** sorted the author by their schools
2. Created column **word_count** storing the number of words in the sentence
3. Created column **sentence_split**, which splits the sentence with **space** and **single quotation mark**, because **'s** standing for **is** should also be considered as a word
4. In the original dataset, there is a column called **sentence_length**. This variable stores the number of chracters in the sentence. However, in my opnion, this is not really informative, because the length of a sentence should be the number of words in it. As a result, I calculated the average number of words in the sentence for each author
5. There are some new sub-dataframes created
6. The first three rows of the modified dataframe **philosophy_data** will be printed below to provide an insight

In [10]:

```
# The original data 'philosophy' is manipulated in many ways.
# So this cell prints out the new version of the dataframe which has a few more new columns
philosophy_data.head(3)
```

Out[10]:

	title	author	school	sentence_spacy	sentence_str	original_publication_date	corpus_edition_date	sentence_length	sentence_lowered	tokeni
0	Plato - Complete Works	Plato	plato	What's new, Socrates, to make you leave your ...	What's new, Socrates, to make you leave your u...	-350	1997	125	what's new, socrates, to make you leave your ...	['wha 'socra 'make
1	Plato - Complete Works	Plato	plato	Surely you are not prosecuting anyone before t...	Surely you are not prosecuting anyone before t...	-350	1997	69	surely you are not prosecuting anyone before t...	['surel 'ar 'prosec
2	Plato - Complete Works	Plato	plato	The Athenians do not call this a prosecution b...	The Athenians do not call this a prosecution b...	-350	1997	74	the athenians do not call this a prosecution b...	'ath 'do', 'nc

Exploratory Data Analysis

In [11]:

```
# Use the explode method to convert each list of strings into multiple rows
modern_exploded = modern['sentence_split'].apply(pd.Series).stack().reset_index(level=1, drop=True)

# Get the frequency of each string in the exploded column
modern_freq = modern_exploded.value_counts().sort_values(ascending=False)


# Use the explode method to convert each list of strings into multiple rows
ancientgreek_exploded = ancientgreek['sentence_split'].apply(pd.Series).stack().reset_index(level=1, drop=True)

# Get the frequency of each string in the exploded column
ancientgreek_freq = ancientgreek_exploded.value_counts().sort_values(ascending=False)
```

In [12]:

```
print('modern_freq:', + modern_freq)

```

```
modern_freq: the                133031
of                  100106
to                   70337
and                  60363
that                 55040
                      ...
pretended)               1
idolaters                1
conservationem.          1
modif                    1
Presence                 1
Length: 79105, dtype: int64
```

In [13]:

```
print('ancientgreek_freq:', + ancientgreek_freq)
```

```
ancientgreek_freq: the              139379
of                  74213
and                 71175
is                  63426
to                  62340
                    ...
Distinguishing          1
tales;                  1
wrest                   1
about!                  1
Cynosarges,             1
Length: 65194, dtype: int64
```

In [14]:

```
# commonly used preposition - place
place_preposition = ['in', 'at', 'on', 'by', 'next to', 'beside', 'under', 'below', 'over', 'above', 'across', 'through', 'to', '

# commonly used preposition - time
time_preposition = ['on', 'in', 'at', 'since', 'for', 'ago', 'before', 'to', 'past', 'till', 'untill', 'by']
```

In [15]:

```
# Calculating the frequencies of the place preposition and the place preposition
modern_place_frequency = 0
for index, row in modern.iterrows():
    for word in row['sentence_split']:
        if word in place_preposition:
            modern_place_frequency += 1

modern_time_frequency = 0
for index, row in modern.iterrows():
    for word in row['sentence_split']:
        if word in time_preposition:
            modern_time_frequency += 1

ancient_place_frequency = 0
for index, row in ancientgreek.iterrows():
    for word in row['sentence_split']:
        if word in place_preposition:
            ancient_place_frequency += 1

ancient_time_frequency = 0
for index, row in ancientgreek.iterrows():
    for word in row['sentence_split']:
        if word in time_preposition:
            ancient_time_frequency += 1
```

In [16]:

```
print('total words in modern:', modern_freq.sum())
print('modern_place_frequency:', modern_place_frequency)
print('modern_time_frequency:', modern_time_frequency)
```

```
total words in modern: 2488549
modern_place_frequency: 270343
modern_time_frequency: 169267
```

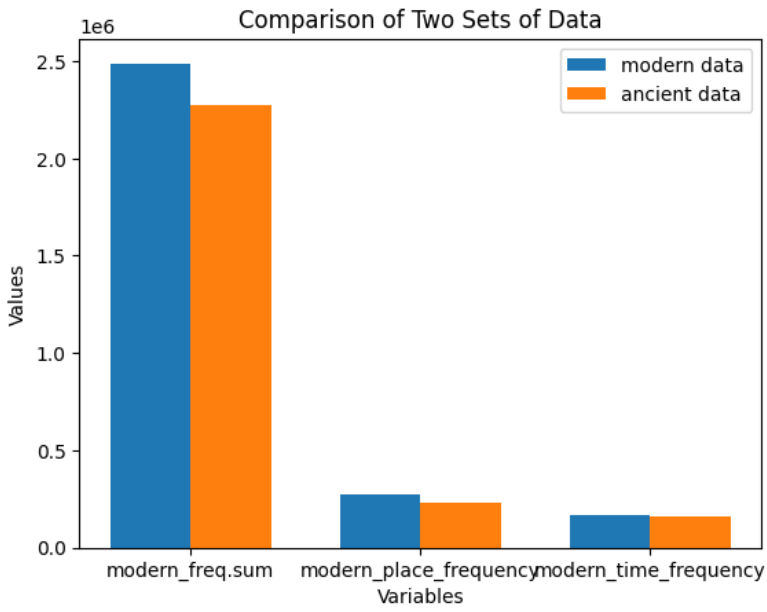
In [17]:

```
print('total words in ancient greek:', ancientgreek_freq.sum())
print('ancient_place_frequency:', ancient_place_frequency)
print('ancient_time_frequency:', ancient_time_frequency)
```

```
total words in ancient greek: 2274867
ancient_place_frequency: 231442
ancient_time_frequency: 161901
```

In [18]:

```
# Plot of the preposition in both Acient Greek and Modern Philosophy

modern_data = [modern_freq.sum(), modern_place_frequency, modern_time_frequency]

ancient_data = [ancientgreek_freq.sum(), ancient_place_frequency, ancient_time_frequency]

# Bar chart data
x = range(len(modern_data))

fig, ax = plt.subplots()
bar_width = 0.35

bar1 = ax.bar(x, modern_data, bar_width, label='modern data')
bar2 = ax.bar([i + bar_width for i in x], ancient_data, bar_width, label='ancient data')

# Add labels and title
ax.set_xlabel('Variables')
ax.set_ylabel('Values')
ax.set_title('Comparison of Two Sets of Data')

# Add X-axis labels
ax.set_xticks([i + bar_width/2 for i in x])
ax.set_xticklabels(['modern_freq.sum', 'modern_place_frequency', 'modern_time_frequency'])

plt.legend()
plt.show()

```



In [19]:

```
# modern_stop = modern_freq[1:150].index.tolist()
# ancient_stop = ancientgreek_freq[1:150].index.tolist()
#nltk.download('stopwords')

# The stop_words that would be removed
stop_words = set(stopwords.words("english"))
stop_words_ob = {'one', 'thing', 'good', 'others', 'would'}
stop_words = stop_words | stop_words_ob

```

In [20]:

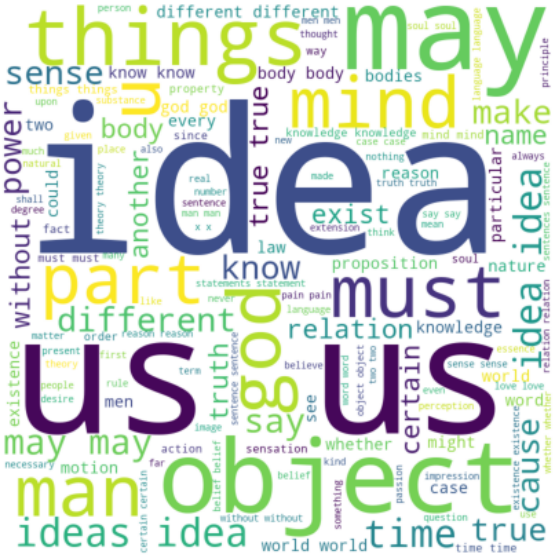
```
# Define two functions

# This function removes the stop words from the data
def remove_stop_words(text):
    words = nltk.word_tokenize(text)
    words = [word for word in words if word.lower() not in stop_words]
    return " ".join(words)

# This function removes the words that appear with low frequency
def remove_low_frequency_words(text):
    words = nltk.word_tokenize(text)
    fdist = FreqDist(words)
    words = [word for word in words if fdist[word] >= 2]
    return " ".join(words)
```

In [21]:

```
# Apply the functions defined to both sub-dataframe so that the plot would not be a mess
modern['nostop'] = modern['sentence_lowered'].apply(remove_stop_words).apply(remove_low_frequency_words)
ancientgreek['nostop'] = ancientgreek['sentence_lowered'].apply(remove_stop_words).apply(remove_low_frequency_words)
```

In [22]:

```
# Word Cloud for mordern data

# concatenate the list of words in the column
text = " ".join(str(word) for word in modern["nostop"])

# create a word cloud object
wordcloud = WordCloud(width = 800, height = 800,
                background_color ='white',
                stopwords = set(stop_words),
                min_font_size = 10).generate(text)

# plot the word cloud
plt.figure(figsize = (4, 4), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```



In [23]:

```
# concatenate the list of words in the column
text = " ".join(str(word) for word in ancientgreek["nostop"])

# create a word cloud object
wordcloud = WordCloud(width = 800, height = 800,
                background_color ='white',
                stopwords = set(stop_words),
                min_font_size = 10).generate(text)

# plot the word cloud
plt.figure(figsize = (4, 4), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```

Summary of the entire project

Procedures

1. At the very beginning of the project, even before data cleaning, I observed the entire dataset, and looked for the similarities and differences among the variables. So I acquired the relationship between the **author** and **school**.
2. After that, I did some background research on the **author** and **school**. I asked the research question of this project and proposed my hypothesis.
3. During data clearning, I found, in the original dataset, a column called **sentence_length**. I believed that this length given by the characters is not informative, so I created a new column storing the number of words in the sentence. Also, I created another column to store the split of the sentence, which is a list of strings.
4. In EDA, I analyzed the frequency of time and place prepositions in both Modern and Acient Greek philosophy, and plot the relationship in bar plot. Also, I defined two functions **remove_stop_words** and **remove_low_frequency_words** to remove the **stop_words** and the low frequency words, so that I will be able to draw the WordCloud.

Interpretation of the results in EDA

1. In the EDA, I acquired the frequency of the commonly seen time and place preposition in both Modern and Ancient Greek Philosophy sentences. The Result is that:
   total words in modern: 2488549
   modern_place_frequency: 270343
   modern_time_frequency: 169267

   total words in ancient greek: 2274867
   ancient_place_frequency: 231442
   ancient_time_frequency: 161901

   Also the relationship between these two periods of philosophy is plotted in the bar graph in the section of EDA.
   By looking at the data and the plot, I could conclude that there is not much difference in the uses of prepositions during these two periods of time, and the existing difference might be the error. However, we could also observe that the use of place preposition is more than that of the time preposition. There might be two possible explanations. First, the time preposition list contains less words than that of the place preposition. Second, the philosopher indeed tends to use place preposition more than time preposition.
2. After removing the **stop_words** and the low frequency words, the two **Word Clouds** could support my hypothesis.
   In Modern Philosophy, words philosopher used frequently are "idea", "us", "man", "things", "sense power", "mind", and etc.
   In Ancient Greek philosophy, words philosopher used frequently are "man", "part", "things", "must", "animal", "motion", "like", and etc.

Even though there is a few words overlapping, we can still observe that words tend to be close to the reality of the time. For example, ancient people would certainly be concerned about "motion", "animal", "man", because these are what support their livings. In Modern Philosophy, people are being physically satisifed, so philosophers would have more imagination, innovation or idea. As a result, "idea", "us", "mind", and "sense power" are what being used the most in the sentences.

Conclusion

The **research question** was how modern philosophy is different from the Ancient Greek Philosophy, from the aspects of writing language and the central theme, and I **hypothesized** that the writing language would be different because the way people talk and write are evolving. Also, the central themes were also shifted because the ideology and understanding of life are changing in an unpredictable speed.
By the previous parts, I did not prove whether philosophers' writing language or habit are changed. However, by the EDA, I showed how the central themes were shifted based on the reality of time.