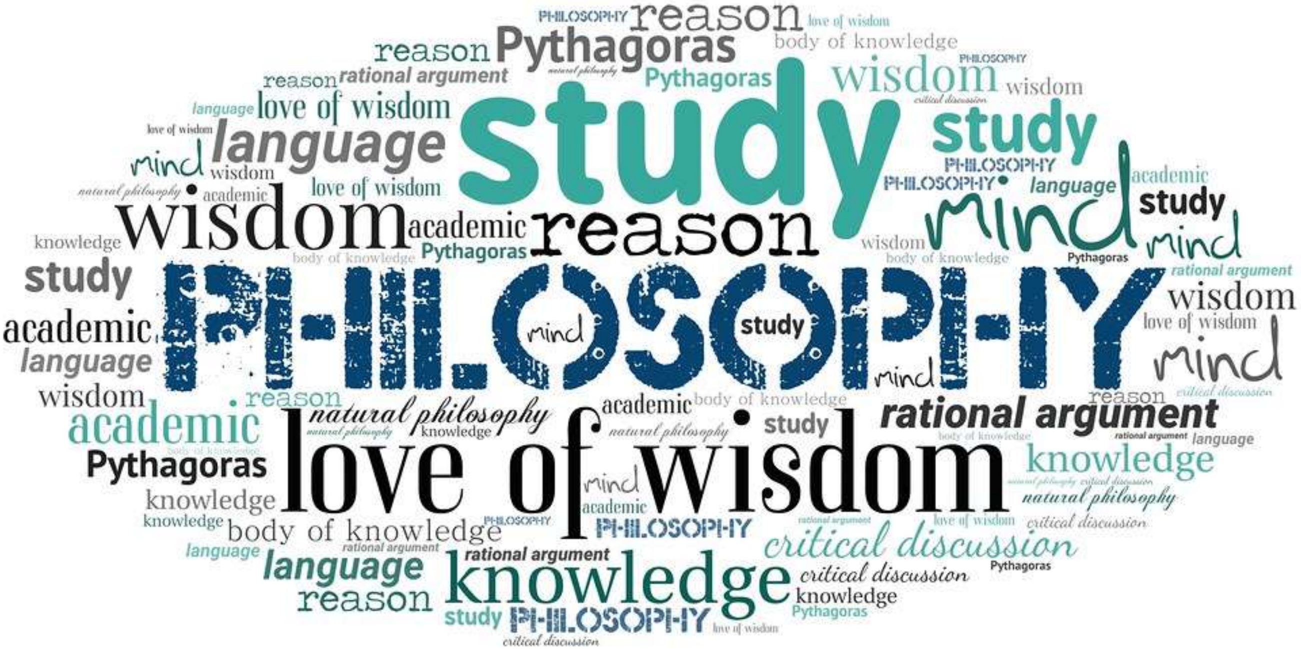
Project: 1 How do philosopher talk?

Introduction and Motivation

In this project I will be investigating how past philosopher talk. Specifically, perform a words frequency analysis to see if I can have some interesting findings.

Philosophers are thinkers of our worlds. They often bring new ideals, instigate reform, or propose hypothesis. An analyzsis of how those people talk can help us understand their ideas more clearly and bring fresh insight of how we look at the world. I was given a dataset of the sentences said by philosophers. Sentences are composes of words. Maybe, if I can do analyze on each word said by philosophers, I can have an idea of what philosopher talk each day. Using this data, I can further interprete what topic do philosophers often talk about. Other than a gross summary of every philosophers. I am also interested in finding the words analysis for each author and school. I can build a search function that can output the words used by an author or school given their name. From this, I can do comparison between each author or school. Philosophers are well know to having debate. Using the word analyze. I might be able to see which schools share similar word choice and which are different.



Information about the dataset

The dataset I used for this project adopted from [here (https://www.kaggle.com/kouroshalizadeh/history-of-philosophy)](https://www.kaggle.com/kouroshalizadeh/history-of-philosophy). This dataset contains 360808 sentences said by 36 different famous philosophers such as Plato and Aristotle as well as the 13 different schools those philosopher belongs to. Each sentence also has a title and the year of publication.

Flaw in the dataset

The data itself is mostly clean. For one author or school, it appears the same through out the dataset making it easy for later to search or group. The sentence is also very clean. There is typo or words incorrectly joined together. The creator of this database also provide a lowered case sentence as well as tokenized and lemmatized text. Unfortunatly those text cannot be used for our analysis.

Another problem with this dataset is with the original publication data. Mainly the publication date of the author Plato and his respective school plato is negatvie. There are also some publication data does not correspond to an actual year. By elimination them, the dataset would loss almost 25% content. Hence, I decide those data must be kept for sake of integrity of the dateset.

Idea of Approch

Since there is a flaw in the year column of the data. I decide to not include any time analysis on the dataset and focus just on the words analysis. Under this circumstance, I believe **words frequency analysis** would be a good approch. Basically, this analysis will give us the number of appearence of each words. During this process, words such as pronouns and preposition will be removed since they appear too often and does not give us much information about the sentence.

Process the data

In order to do word frequency analysis. I have to clean each sentance to remove any pronouns, preposition, punctuation ect as discussed before. Then, split each word into tokens. I also lemmatized the words, this step can reduce each words to its simplist form.

For example:

" What's new, Socrates, to make you leave your usual haunts in the Lyceum and spend your time here by the king archon's court?"

Would become:

'new','socrates','make','leave','usual','haunt','lyceum','spend','time','king','archon','court'

After this step, all I have to do is to count the total number of each words and see what does the data tell us.

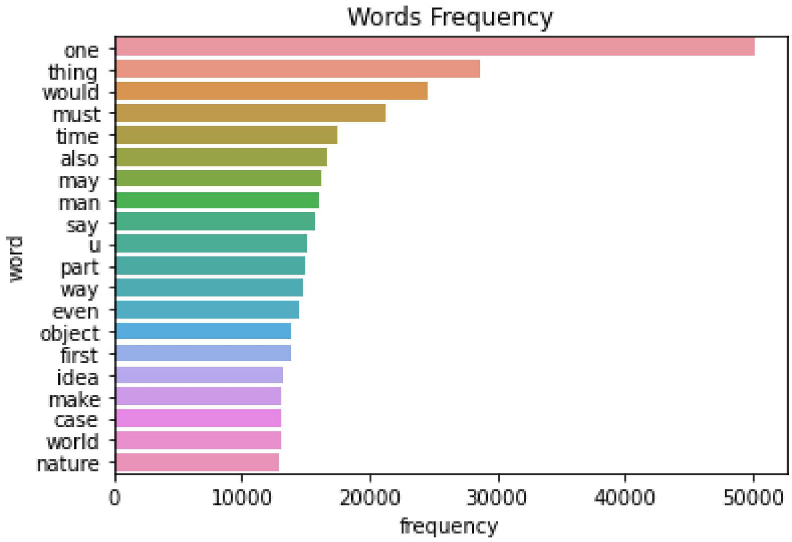
Which is the Most Famous Word Used by Philosophiers

After processing the data. I can anwser the first question: Which is the Most Famous Word Used by Philosophiers? I present the top three words used by philosophier and their respect frequency.

Out[10]:

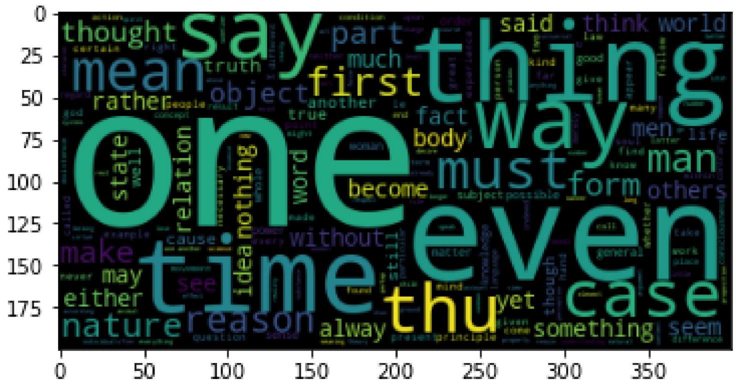
	word	frequency
48	one	50186
51	thing	28692
183	would	24614

I also make a bar chart to show the top 20 words used by philosophiers.Using this result, I can demonstrate the result visually by ploting a word cloud as shown below. This would gives us some idea on the word choice of philosopher and give us insight on interpreting how philosopher talk. Apart from this, I also build a search function that allows to search the dataset given a name or a word.



```
Out[12]:

<matplotlib.image.AxesImage at 0x13c28e78d60>
```



For example, the word "one" is used most frequently. In fact 11.3% sentance contains the word "one". I can search the dataset to see some example of sentence containing the word "one".

Here are some result of my search:

- A not ignoble one I think
- One whom I am thought crazy to prosecute.
- One of you might perhaps interrupt me and say: 'But Socrates, what is your occupation?
- One need not worry about them, but meet them head on.

Here, I can see some trend. Many sentence would start with the word "one" and "one" is usually refers to "person".

We are also interested in analyzing the words frequency of each author. Here, I create a search function that can generate a list of all the words and frequency used by that author and plot the word cloud of that author.

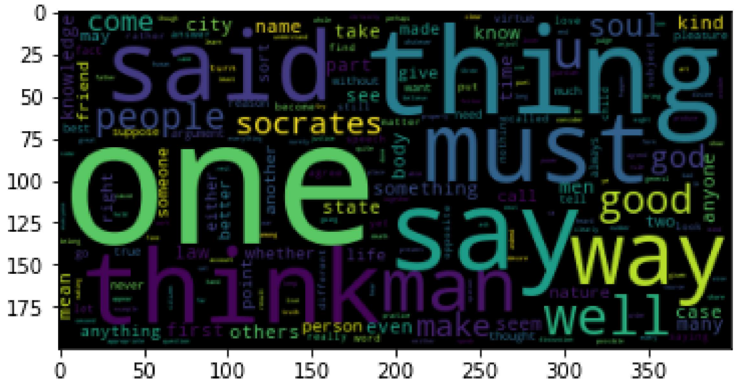
By apply this function on the philosopher "Plato", I have some interesting findings. First, the word choice of Plato follow the general pattern I found when analyzing the whole dataset. Also, there are some interesing words in the word cloud like "socrates". This is in fact a name of yet another philosopher and Plato's teacher. Here are some example of the sentences:

- What's new, Socrates, to make you leave your usual haunts in the Lyceum and spend your time here by the king archon's court?
- I don't know him, Socrates.
- I could wish this were true, Socrates, but I fear the opposite may happen.
- I understand, Socrates.

It is not hard to find that those things Plato said to Socrates. A quick google search would tell us that Socrates is Plato's teacher.Given this context, I can infer that many of the sentence are from the daliy conversation with Socrates even for those sentence that does not have the key word "socrates". Knowing this background story, we can learn more information with the sentence said by Plato.

Out[15]:

	word	frequency
48	one	5556
51	thing	4609
183	would	3302



A similar analyze could be conducted for each school. Here, I used "empiricism" as an example. As shown in the word cloud. "Idea" and "mind" has been used very often in this school which is different from what we obtained from the general trend. Here are some example of sentences contain "idea" in "empiricism" school:

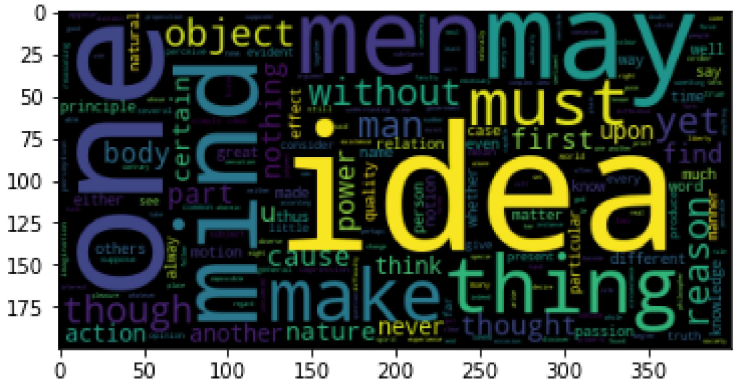
- What Idea stands for.
- Assent to supposed innate truths depends on having clear and distinct ideas of what their terms mean, and not on their innateness.
- But, since no proposition can be innate unless the ideas about which it is be innate, this will be to suppose all our ideas of colours, sounds, tastes, figure, andc.
- For I would gladly have any one name that proposition whose terms or ideas were either of them innate.

From this, I can see that empiricism philosopher often connect innate with idea. They often discuss whether the idea is innate or not. From this, one can learn some general topic a school can discuss.

However, just by looking at the graphs and number is not accurate. In the next section, I will be performing hypothesis testing to see if each school talks differently.

Out[17]:

	word	frequency
3924	idea	5871
33	one	3375
58	may	2696



Does Each School Talk similarly?

After analyzing each author and school, I am also interested in find the similarities or difference between every group.

In this section, I will use ANOVA test to see if the word choice of each school is similar. This test can give us information on weather the choice of words is different when the school is different.

The words that are tested is the top ten words used from all sentences (one,thing,would,must,time,also,many,say,u). The first step is to get the sum of each word in each school as shown below.

Out[27]:

word	one	thing	would	must	time	also	may	man	say	u	school
0	5556	4609	3302	2374	1333	1165	690	2162	2971	2047	plato
1	9698	7365	2027	4481	2654	4344	2444	4100	2440	1168	aristotle
2	3375	2041	1258	1488	789	530	2696	1416	677	1980	empiricism
3	3099	3184	2103	1718	807	1002	790	1319	1137	2595	rationalism
4	6437	2457	4606	2004	2083	2055	3784	852	4255	1894	analytic
5	4214	1152	1967	1529	1781	1130	743	1359	886	1020	continental
6	2863	2242	819	1535	1916	998	520	581	719	1497	phenomenology
7	7302	3072	3149	2752	2359	3332	1483	309	971	1685	german_idealism
8	1814	309	711	696	1225	556	608	334	493	333	communism
9	1610	408	2960	1257	1309	242	1613	381	310	238	capitalism
10	322	687	82	188	145	165	147	341	86	57	stoicism
11	2124	781	605	633	478	648	404	1063	437	408	nietzsche
12	1772	385	1025	582	677	597	369	1863	339	216	feminism

Now, I will perform ANOVA test on the chart we just obtained. This test will tell us does the words choice differ by each school.

Out[59]:

```
F_onewayResult(statistic=7.619955343937744, pvalue=2.8318483451450726e-10)
```

Here the p-value is 2.8318483451450726e-10, which is less than the significant level. Hence, we can conclude that not all school have similar word choice. This is easy to understand since there are 13 different school each has their own idea. It is reasonable that some school talk differently than others. Then my question become, are there some schools have similar word choice? To investigate this, I performed pairwised ANOVA test between every school. Here is a list of which the test is significant between the two schools.

```
plato aristotle
plato empiricism
plato rationalism
plato analytic
plato continental
plato german_idealism
aristotle analytic
aristotle german_idealism
empiricism rationalism
empiricism continental
empiricism phenomenology
empiricism german_idealism
empiricism capitalism
rationalism analytic
rationalism continental
rationalism phenomenology
rationalism german_idealism
rationalism capitalism
analytic german_idealism
continental phenomenology
continental german_idealism
continental capitalism
phenomenology german_idealism
phenomenology capitalism
phenomenology nietzsche
phenomenology feminism
communism capitalism
communism nietzsche
communism feminism
capitalism nietzsche
capitalism feminism
nietzsche feminism
```

This can give us insight on which schools have similar word choice. Based on this information, further inference can be made.

What to do next?

Hopefully, after reading this, you will have an idea of the way philosopher talk. However, this is not the end of this dataset. There are other things data scientists can do using this dataset. For example, one can build a chat AI that talk like a true philosopher based on this word frequency analyze. The sky is not limited.