Exercise 4.1 - Group 6

1. reads the contents of "aston.txt" (file can be downloaded via the downloads tab below)

2. finds all unique words from the file

3. prints them to a file called "unique_words.txt", sorted alphabetically

4. finds the longest word in the file and prints it to the screen.

In [3]:
```
'''
1. I noted that aston.txt needs some work in order to achieve the requirements.

1a. There are citation brackets throughout the text in the form of [1], [2], etc.
These need to removed, but becareful not to remove the dates and
other information in normal brackets ()

1b. There may be instances of words with apostrophes like 's or s'.
These need to cleaned, i.e. it's and its to become it,
as it's and its could skew the result.

1c. The entire text needs to be either uppercase or lowercase to avoid us
picking up duplicates. i.e. Major and major are essentially the same word,
but to python, they different words as python it is case senstitve.

'''

# IMPORT REGULAR EXPRESSIONS
import re

# OPEN THE FILE ASTON.TXT
text_file = open('aston.txt', 'r')
# READ THE FILE AND SAVE TO THE TEXT VARIABLE
text = text_file.read()
# LOWERCASE ALL TEXT
text = text.lower()
# RE SUB WILL REPLACE ONE OR MANY MATCHES WITH A STRING
new_words = re.sub("[\(\[].*?[\)\]]", "", text)
# SPLIT THE WORDS
words = new_words.split()
# LETS CLEAN THE TEXT
words = [word.strip('.,!;()[]:') for word in words]
# CLEAN THE IT'S TO ITS
words = [word.replace("'s", '') for word in words]

# CREATE A BLANK UNIQUE LIST
# SAVE ALL UNIQUE WORDS TO A LIST
# BY LOOPING THROUGH THE BODY OF THE TEXT
unique = []
for word in words:
    if word not in unique:
        unique.append(word)
# SORT THE UNIQUE LIST
unique.sort()

# GET THE TOTAL UNIQUE WORDS
total_words = len(unique)
```

```
# GET THE SHORTEST WORD
shortest_string = min(unique, key=len)
# GET THE LONGEST WORD
longest_string = max(unique, key=len)

# PRINT THE OUTPUTS
print(f"The number of words in the list is: {total_words}")
print(f"The shortest word in the list is: {shortest_string}")
print(f"The longest word in the list is: {longest_string}")

# OPEN A FILE HANDLER AND CREATE A NEW TXT FILE
with open('unique_words.txt', 'w') as file_handler:
    # LOOP THROUGH THE UNIQUE LIST
    for list_item in unique:
        # SAVE ALL UNIQUE WORDS TO A LIST
        file_handler.write('%s\n' % list_item)
```

```
The number of words in the list is: 333
The shortest word in the list is: &
The longest word in the list is: apprenticeship
```

In [ ]: