INTRODUCTION

http://quotes.toscrape.com/

This website is a static website that lists several quotes, tags and authors

Our desire as the collaborators on this project is to scrap the website and obtain words, common tags and what authors have the most quotes so that we carry out an analysis on what words were most common, how often words were mentioned, what are most common tags, and what authors have the most quotes.

N.B: robots.txt rules are obeyed

Techniques used in this project were as per requirements which are:

i. Scrapy

ii. Beautiful Soup

iii. Selenium

The above mentioned techniques were used to automate the a script that extracted data associated with quotes from the above mentioned website. As per mention, multiple web pages were scrapped and output results were given in csv format.

The general format for the scrapers is a four step process:

i. Import request libraries

ii. Specify URL of website to be scraped

iii. Send HTTP request to the specified URL

iv. Save response from server in a response object file

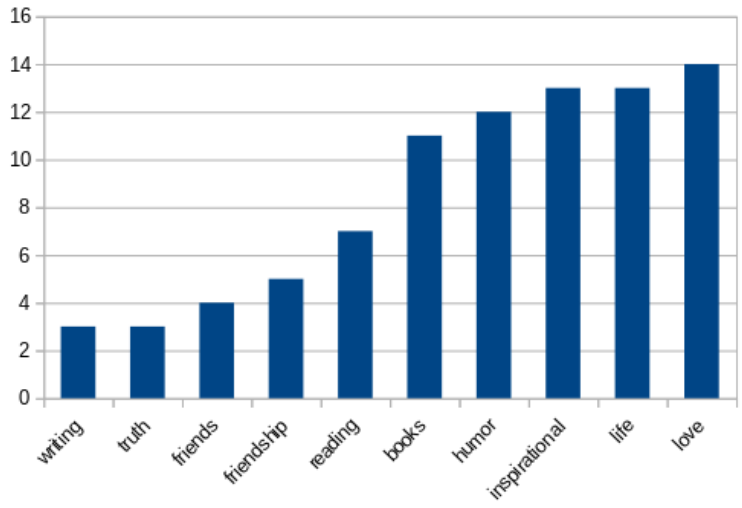
DESCRIPTION OF OUTPUT

The output is 7 csv files

i. authors.csv shows the author names and the number of times the author appears on the website
ii. quotes.csv shows the quote, the name of author and the tags associated with it
iii. tags.csv shows the tags and the number of times the tag appears on the website
iv. top_authors shows the top author of quotes
v. top_tags shows the top common tags
vi. top_words shows the top common words
vii. words.csv shows the frequency of occurrence of words

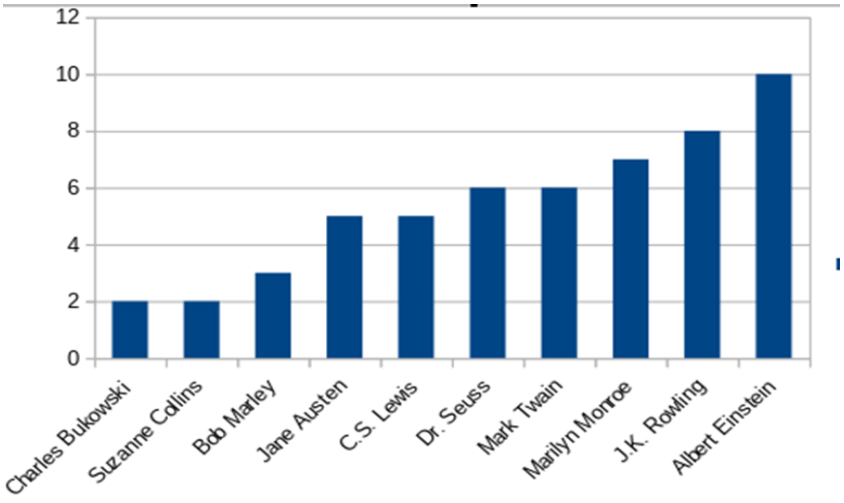
DATA ANALYSIS

Top 10 Tags



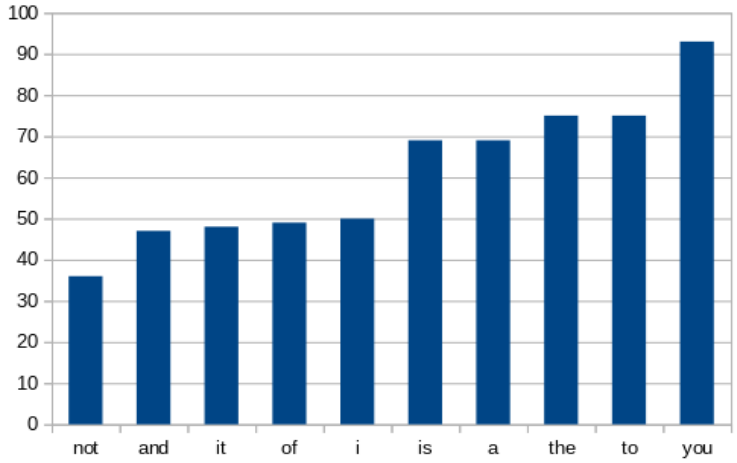
This is a simple bar chart that shows the number of times a word appears in tags

Top 10 Authors



This shows that the top 10 authors of quotes

Top 10 Words



This shows the top 10 most common words in all quotes

FURTHER ANALYSIS

The below codes are evaluations in R for the requested elementary statistical data analysis for scraped authors, top_authors, top_tags, and top_words csv files. Solutions from the console are given below. We calculated the following:

i. Mean
ii. trimmed mean
iii. midrange
iv. mode
v. median
vi. quantiles
vii. range
viii. variance and standard deviation

AUTHORS ANALYSIS

```
> authors <- read.csv2(file = "authors.csv", head = T, sep = ",")

> #means

> mean(authors$Freq)

[1] 1.960784
```

```
> mean(authors$Freq, trim=0.1)
[1] 1.414634
> mean(authors$Freq, trim=0.2)
[1] 1.16129
> #midrange
> (min(authors$Freq)+max(authors$Freq))/2
[1] 5.5
> #mode
 names(sort(-table(authors$Freq)))[1]
[1] "1"
> #median
> median(authors$Freq)
[1] 1
> #quantiles
> quantile(authors$Freq, probs=c(0.25, 0.5, 0.75))
25% 50% 75%
 1   1   2
 #range
> range(authors$Freq)
[1]  1 10

 #variance and standard deviation
> var(authors$Freq)
[1] 4.238431
> sd(authors$Freq)
[1] 2.058745
```

TOP AUTHORS ANALYSIS

```
> top_authors <- read.csv2(file = "top_authors.csv", head = T, sep = ",")
> #means
> mean(top_authors$Freq)
[1] 5.4
```

```
> mean(top_authors$Freq, trim=0.1)
[1] 5.25
> mean(top_authors$Freq, trim=0.2)
[1] 5.333333
> #midrange
> (min(top_authors$Freq)+max(top_authors$Freq))/2
[1] 6
> #mode
> names(sort(-table(top_authors$Freq)))[1]
[1] "2"
> #median
> median(top_authors$Freq)
[1] 5.5
> #quantiles
> quantile(top_authors$Freq, probs=c(0.25, 0.5, 0.75))
 25%  50%  75%
3.50 5.50 6.75
> #range
> range(top_authors$Freq)
[1]  2 10
> #variance and standard deviation
> var(top_authors$Freq)
[1] 6.711111
> sd(top_authors$Freq)
[1] 2.590581
```

TOP TAGS ANALYSIS

```
> top_tags <- read.csv2(file = "top_tags.csv", head = T, sep = ",")
> #means
> mean(top_tags$Freq)
[1] 8.5
> mean(top_tags$Freq, trim=0.1)
```

```
[1] 8.5
> mean(top_tags$Freq, trim=0.2)
[1] 8.666667
> #midrange
> (min(top_tags$Freq)+max(top_tags$Freq))/2
[1] 8.5
> #mode
> names(sort(-table(top_tags$Freq)))[1]
[1] "3"
> #median
> median(top_tags$Freq)
[1] 9
> #quantiles
> quantile(top_tags$Freq, probs=c(0.25, 0.5, 0.75))
  25%   50%   75%
 4.25  9.00 12.75
> #range
> range(top_tags$Freq)
[1]  3 14
> #variance and standard deviation
> var(top_tags$Freq)
[1] 20.5
> sd(top_tags$Freq)
[1] 4.527693
```

TOP WORDS ANALYSIS

```
> top_words <- read.csv2(file = "top_words.csv", head = T, sep = ",")
> #means
> mean(top_words$Freq)
[1] 61.1
> mean(top_words$Freq, trim=0.1)
[1] 60.25
```

```
> mean(top_words$Freq, trim=0.2)
[1] 60
> #midrange
> (min(top_words$Freq)+max(top_words$Freq))/2
[1] 64.5
> #mode
> names(sort(-table(top_words$Freq)))[1]
[1] "69"
> #median
> median(top_words$Freq)
[1] 59.5
> #quantiles
> quantile(top_words$Freq, probs=c(0.25, 0.5, 0.75))
  25%   50%   75%
48.25 59.50 73.50
> #range
> range(top_words$Freq)
[1] 36 93
> #variance and standard deviation
> var(top_words$Freq)
[1] 310.9889
> sd(top_words$Freq)
[1] 17.63488
```

PART ALLOCATION

i. Beautiful Soup - Sandile Kholisani Nhlalo – Sibanda
ii. Selenium - Sandile Kholisani Nhlalo – Sibanda
iii. Scrapy - Nomthunzi Moyo
iv. Analysis - Nomthunzi Moyo
v. Project file compilation – Nomthunzi Moyo