Homework 4. Frequent Words and Web scraping

Double Click here to edit this cell

- Name:
- Student ID:
- Submission date:

Problem 1 (15 pts)

- Project Gutenberg is a volunteer effort to digitize and archive cultural works.
- Moby-Dick is an 1851 novel by American writer Herman Melville.
- You can find Moby-Dick in an ordinary text format at https://www.gutenberg.org/files/2701/old/moby10b.txt
- Use **requests** module to get the text.

- We want to compute word frequency of words appearing in mobydick and generate WordCloud
  - First, you must split the text into words.
  - **Any symbols(!, ., ?, ,, +, -, *, ...)** are delimeters
  - Numbers should not be words.
  - Null string is not a word.
  - Any delimiters should not be words.
  - To split into words, use **re** (regular expression module)
  - (Upper or lower) Cases does not matter in words

1.1 Print top 50 most common words (5 pts)

In [ ]:

```
# YOUR CODE MUST BE HERE
```

Your output should be like the following:

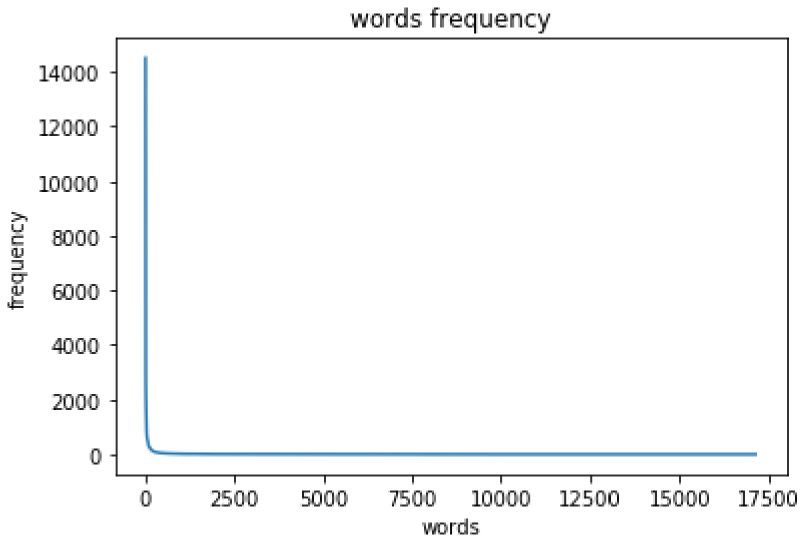
```
[('the', 14512), ('of', 6676), ('and', 6471), ('a', 4774), ('to', 4690),
('in', 4190), ('that', 3095), ('it', 2542), ('his', 2530), ('i', 2128),
('he', 1896), ('but', 1823), ('s', 1811), ('as', 1750), ('is', 1748),
('with', 1729), ('was', 1647), ('for', 1643), ('all', 1537), ('this',
1437), ('at', 1332), ('by', 1232), ('whale', 1228), ('not', 1162), ('from',
1103), ('on', 1077), ('so', 1073), ('him', 1067), ('be', 1058), ('you',
949), ('one', 934), ('there', 870), ('now', 787), ('had', 779), ('have',
773), ('or', 761), ('were', 685), ('they', 669), ('which', 650), ('like',
648), ('me', 634), ('then', 632), ('some', 621), ('what', 620), ('their',
620), ('are', 611), ('when', 608), ('an', 600), ('no', 592), ('my', 589)]
```

1.2 Plot word frequency (5 pts)

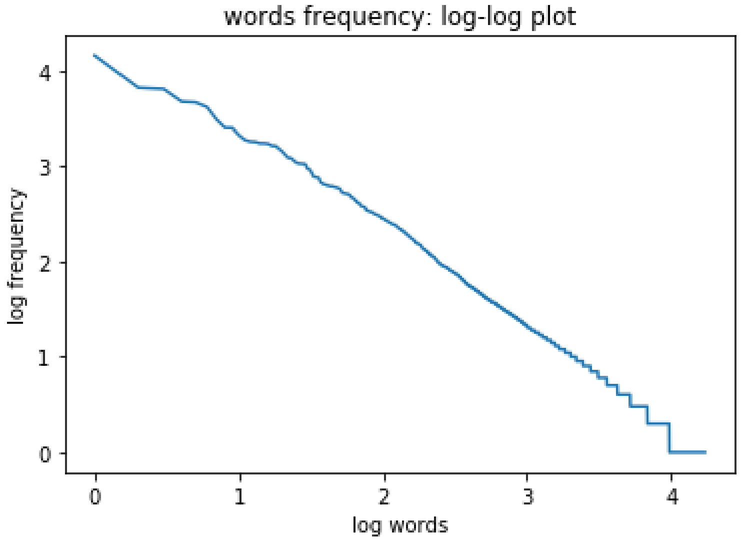
- Sort the word frequency in descending order
- Plot the word frequency
- Plot the word frequency in log-log plot.

In [ ]:

```
# YOUR CODE MUST BE HERE
```



Your output should be like:



Discussion

- Read this wikipedia article : https://ko.wikipedia.org/wiki/%EC%A7%80%ED%94%84%EC%9D%98_%EB%B2%95%EC%B9%99
- Discuss what you learned from the distribution.

WRITE HERE (To edit, double click this cell)

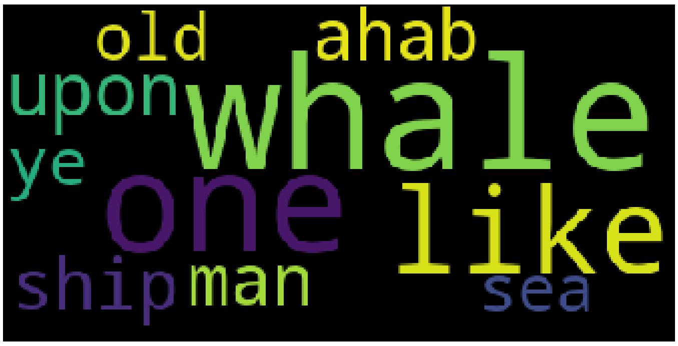
1.3 Word Cloud (5 pts)

- Print top 10 most words except stop words
- Draw word cloud of top 10 most common words
- Googling for how to draw word clouds

Your output should be like:

```
[('whale', 1228), ('one', 934), ('like', 648), ('upon', 566), ('man', 527),
('ship', 518), ('ahab', 511), ('ye', 472), ('sea', 455), ('old', 450)]
```

Your output should be like this (but NOT exactly the the same):



- The following is English stop words list

In [ ]:
```
stopwords = {'it', 'than', 'out', 'an', 'at', 'until', 'wouldn', 'too', 'each', 'c
```

In [ ]:
```
# YOUR CODE MUST BE HERE
```

Problem 2 (20 pts)

- We want to find how many CS faculty members at CS department of Stanford Univ work on CS research areas.
- First, visit https://cs.stanford.edu/research
- Take a look at the source html of the web page.
- We want to scrape data on all the faculty members
- Run the following two cells and see what happens
- If necessary, install html5lib

In [ ]:
```
from bs4 import BeautifulSoup
import requests

url = "https://cs.stanford.edu/research?items_per_page=All&field_faculty_status_va
soup = BeautifulSoup(requests.get(url).text, 'html5lib')

# f = open("stanford_cs.txt", "r")
# text = f.read()
# soup = BeautifulSoup(text, 'html5lib')
```

Remark

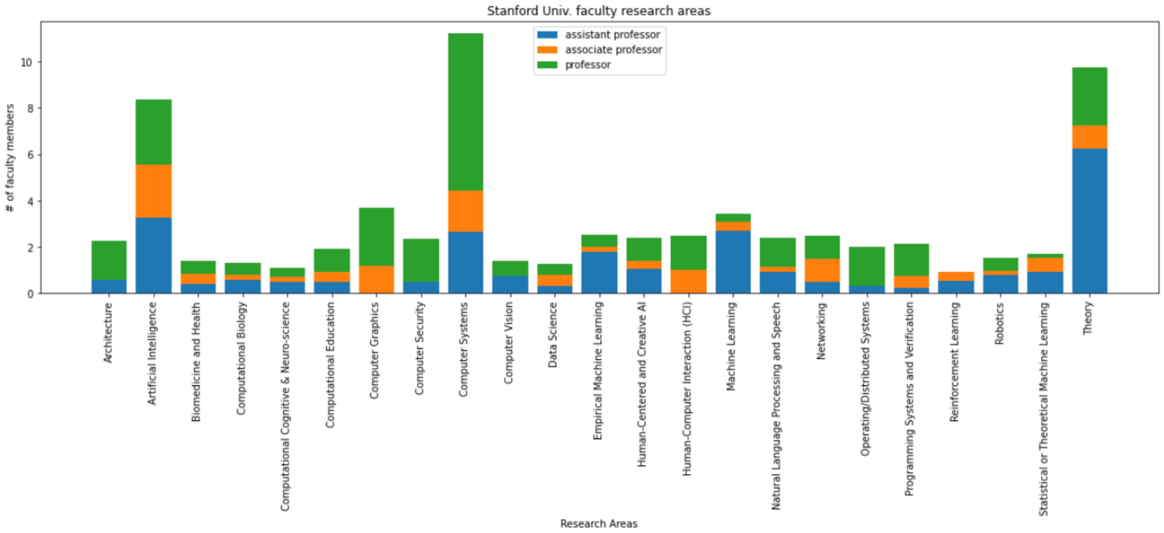
- Stanford Univ에 너무 많이 접속해서 (DDOS처럼 여겨져서) 접속이 막힐 수도 있음
- 해당 웹페이지를 처음 접속해서 파일로 저장한 다음,
- 파일로 부터 읽어서 숙제를 테스트하는 게 필요함.

숙제 제출시 아래 cell은 절대 실행하지(출력에 포함하지) 말 것!!!

In [ ]:
```
print(soup.tbody.prettify())
```

Draw bar charts on research area contributions of Stanford CS faculty

- For each research area, we want to compute how many professors works on that area.
- If one professor works on n research fields, the contribution to one research field is 1/n.
- The colors for professor ranks (assistant, associate, full professors) may be your own choice.
- Your output should be like:



In [ ]:

```
# YOUR CODE MUST BE HERE
```

Ethics:

If you cheat, you will get negatgive of the total points. If the homework total is 22 and you cheat, you get -22.

What to submit

- Run **all cells** after restarting the kernel
- Goto "File -> Print Preview"
- Print the page as pdf
- Pdf file name must be in a form of: homework_4_홍길동_202300001.pdf
- Submit the pdf file in google classroom
- No late homeworks will be accepted
- Your homework will be graded on the basis of correctness, performance, and programming skills
- Your homework will be graded on the basis of correctness and programming skills