

Homework 4. Fequent Words and Web scraping

Double Click here to edit this cell

- Name: 전기범
- Student ID: 201703091
- Submission date: 2019/04/18

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>
(<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 moby dick and generate WordCloud
 - First, you must split the text into words.
 - To do that, **you find word delimiters** (for example, . or , ... whatever).
 - To split into words, use **re** (regular expression module)
 - Numbers should not be words.
 - Null string is not a word.
 - Any delimiters should not be words.
 - (Upper or lower) Cases does not matter in words.

웹사이트의 데이터 긁어오기, 가장 자주 사용되는 단어 찾아보기

1.1 Print top 50 most common words (5 pts)

```
In [1]: import re, requests
freq={}
'''with open("moby10b.txt", 'r', encoding='utf-8') as f:    # txt 파일 이용시
    #L = re.findall('[a-z]+', f.read().lower())'''

response = requests.get('http://www.gutenberg.org/files/2701/old/moby10b.txt')    # 크롤링 이용시
L = re.findall('[a-z]+', response.text.lower())

for word in L:
    cnt = freq.get(word, 0)
    freq[word] = cnt + 1

print([(key,value) for key,value in sorted(freq.items(), key =lambda x:x[1], reverse=True)[:50])]
```

```
[('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)]
```

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)

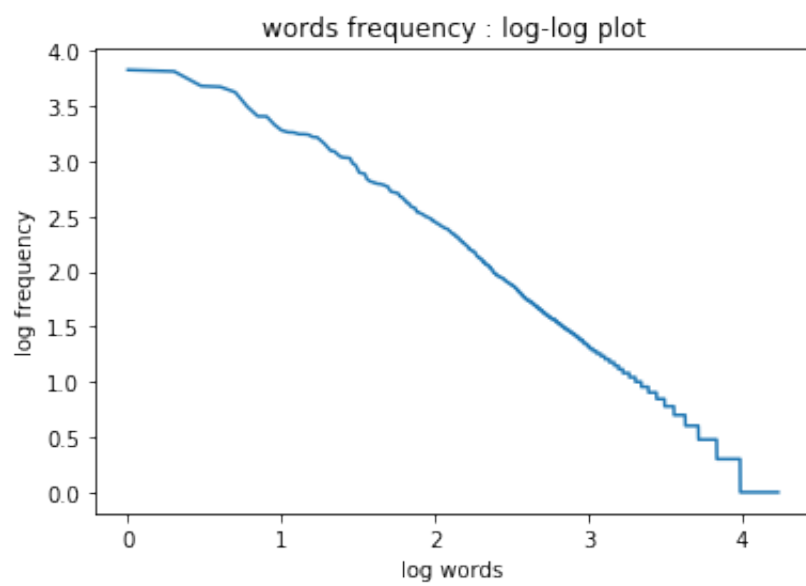
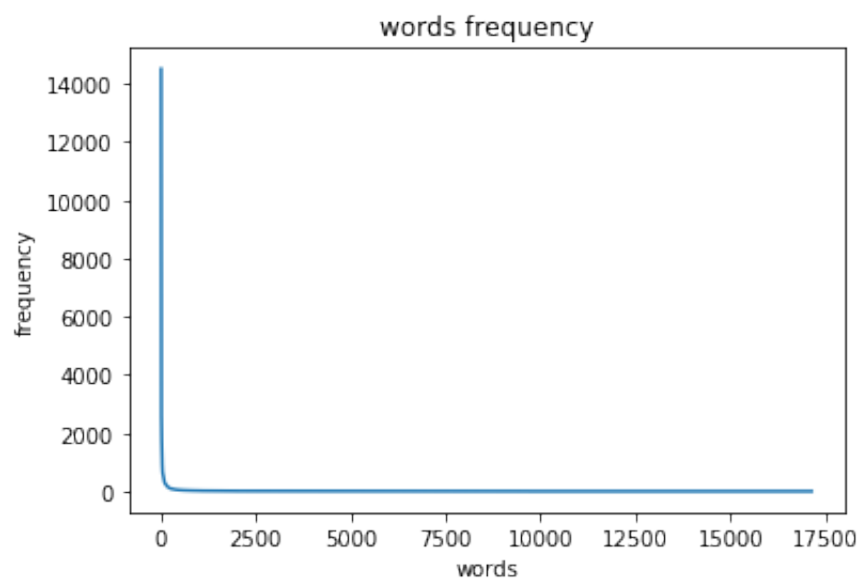
- Sort the word frequency in descending order
- Plot the word frequency
- Plot the word frequency in log-log plot.

```
In [2]: %matplotlib inline
import matplotlib.pyplot as plt
import numpy as np

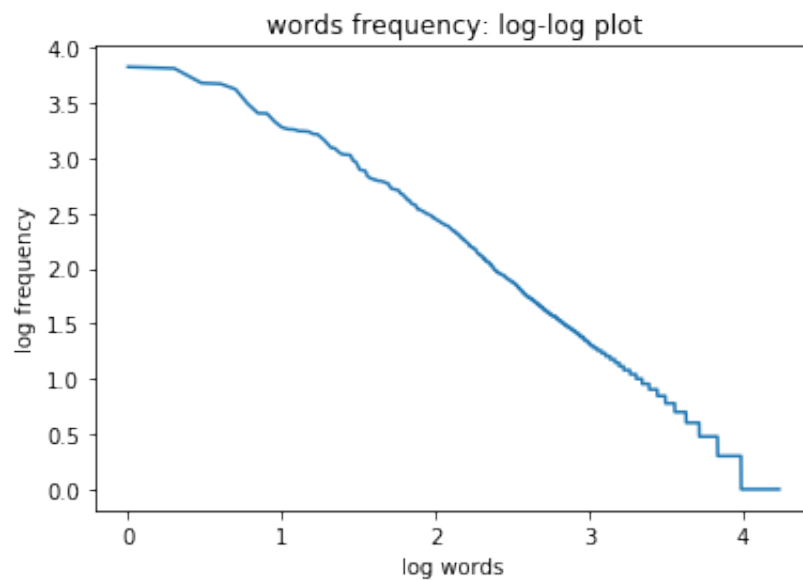
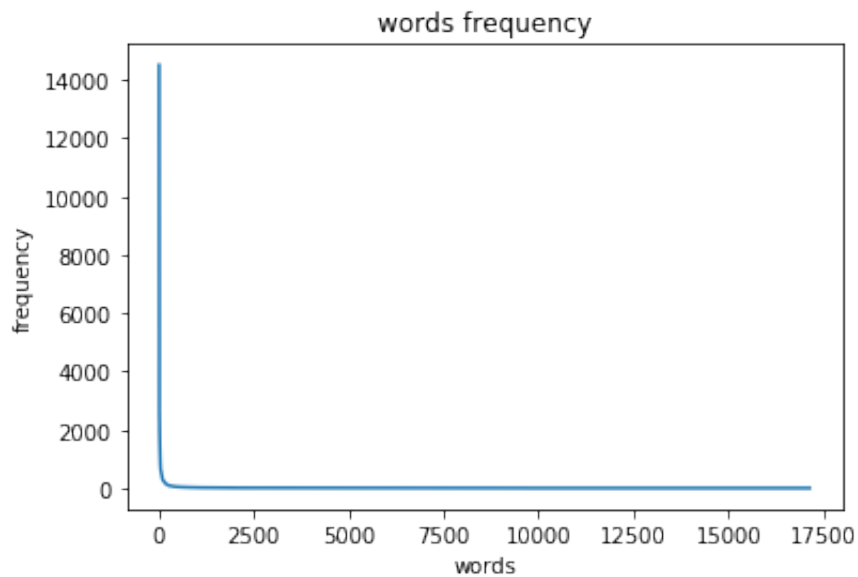
def plot():
    plt.plot(freq_len, word_freq)
    plt.title("words frequency")
    plt.ylabel("frequency")
    plt.xlabel("words")
    plt.show()

def loglog():
    #plt.loglog(freq_len, word_freq)
    np.seterr(divide = 'ignore')
    plt.plot(np.log10(freq_len), np.log10(word_freq))
    plt.title("words frequency : log-log plot")
    plt.ylabel("log frequency")
    plt.xlabel("log words")
    plt.show()

word_freq = np.sort(list(freq.values()))[::-1]
freq_len = np.arange(len(freq))
plot()
loglog()
```



Your output 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%EC%A7%80%ED%94%84%EC%9D%98_%EB%B2%95%EC%B9%
- Discuss what you learned from the distribution.
- WRITE HERE (To edit, double click this cell)

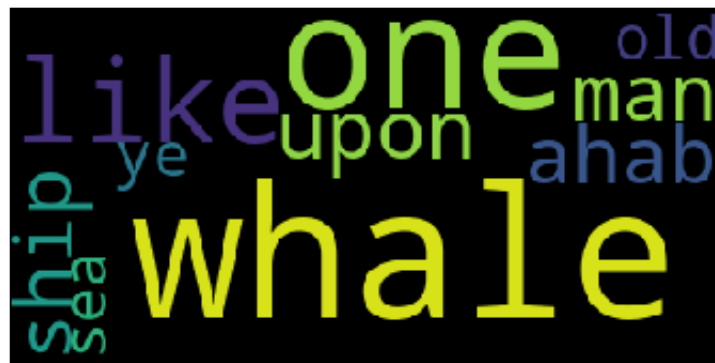
1.3 Word Cloud (5 pts)

- Print top 10 most words except stop words
- Draw word cloud of top 10 most common words

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

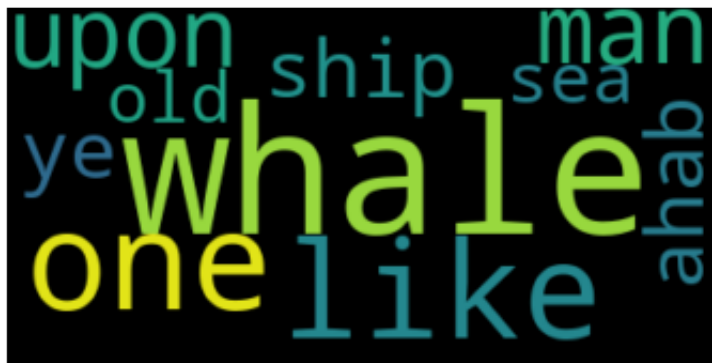
```
In [3]: stopwords = {'it', 'than', 'out', 'an', 'at', 'until', 'wouldn', 't', 'oo', 'each', 'off', 'whom', 'not', 'ain', 'weren', 'you've', 'isn', 've', 'that'll', 'didn', 'couldn', 'with', 'for', 'me', 'shouldn't', 'those', 'once', 'them', 'him', 'again', 'what', 'to', 's', 'don', 'yourselves', 'she's', 'd', 'we', 'so', 'does', 'your', 'is', 'such', 'hasn', 'doesn', 'doesn't', 'no', 'll', 'their', 'before', 'my', 'being', 'and', 'but', 'below', 'won', 'don't', 't', 'myself', 'very', 'why', 'mustn't', 'that', 'been', 'you', 'you'd', 'few', 'ot her', 'ma', 'any', 'having', 'against', 'into', 'on', 'just', 'herself', 'hadn't', 'mightn't', 'aren', 'wouldn't', 'ours', 'about', 't hen', 'mustn', 'i', 'y', 'should', 'all', 'while', 'himself', 'do', 'up', 'were', 'this', 'most', 'when', 'nor', 'from', 'hadn', 'their s', 'she', 'be', 'under', 'or', 'will', 'through', 'our', 'isn't', 'as', 'between', 'had', 'more', 'aren't', 'weren't', 'if', 'the', 'am', 'how', 'both', 'you're', 'yourself', 'couldn't', 'only', 'in', 'itself', 'own', 'it's', 'because', 'some', 'didn't', 'wasn', 'shan', 'hasn't', 'mightn', 'shouldn', 'here', 'he', 'where', 'm', 're', 'was', 'after', 'has', 'same', 'shan't', 'further', 'wasn't', 'down', 'yours', 'should've', 'now', 'needn't', 'above', 'haven', 'its', 'who', 'of', 'ourselves', 'did', 'these', 'there', 'his', 'haven't', 'won't', 'themselves', 'you'll', 'a', 'are', 'which', 'have', 'by', 'during', 'can', 'hers', 'over', 'her', 'doing', 'o', 'needn', 'they'}
```

```
In [4]: # YOUR CODE MUST BE HERE
%matplotlib inline
import matplotlib.pyplot as plt
from wordcloud import WordCloud

for stop in stopwords:
    if stop in freq:
        del freq[stop]
words=([(key,value) for key,value in sorted(freq.items(), key =lambda
da x:x[1], reverse=True)][0:10])
print(words)

text = ([key for key,value in sorted(freq.items(), key =lambda x:x[
1], reverse=True)][0:10])
cloud = WordCloud(max_font_size=100, stopwords=stopwords).generate(
" ".join(text))
plt.figure()
plt.imshow(cloud, interpolation="bilinear")
plt.axis("off")
plt.show()

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



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> (<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 [5]: from bs4 import BeautifulSoup
import requests

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

```
In [6]: print(soup.tbody.prettify())
```

```
<tbody>
  <tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
      <a href="https://profiles.stanford.edu/maneesh-agrawala" target="_blank">
        
      </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
      <a href="https://profiles.stanford.edu/maneesh-agrawala" target="_blank">
        Maneesh Agrawala
      </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
      Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
      Computer Graphics
      <br/>
      Human-Computer Interaction (HCI)
    </td>
  </tr>
  <tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
      <a href="http://theory.stanford.edu/~aiken/" target="_blank">
        
      </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
      <a href="http://theory.stanford.edu/~aiken/" target="_blank">
        Alex Aiken
      </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
      Professor
    </td>
```



```

    <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
        Computer Systems
        <br/>
        Programming Systems and Verification
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://www.bailis.org/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://www.bailis.org/" target="_blank">
            Peter Bailis
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Assistant Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Computer Systems
        <br/>
        Operating/Distributed Systems
        <br/>
        Data Science
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://theory.stanford.edu/~barrett/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://theory.stanford.edu/~barrett/" target="_blank">
            Clark Barrett
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Associate Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Computer Systems
        <br/>
        Architecture

```

```

        <br/>
        Programming Systems and Verification
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://cs.stanford.edu/~serafim" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://cs.stanford.edu/~serafim" target="_blank">
            Serafim Batzoglou
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Associate Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Artificial Intelligence
        <br/>
        Computational Biology
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://bejerano.stanford.edu/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://bejerano.stanford.edu/" target="_blank">
            Gill Bejerano
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Associate Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Artificial Intelligence
        <br/>
        Computational Biology
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">

```

```

        <a href="http://hci.stanford.edu/msb/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://hci.stanford.edu/msb/" target="_blank">
            Michael Bernstein
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Assistant Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Human-Computer Interaction (HCI)
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="https://engineering.stanford.edu/people/jeannette-bohg" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="https://engineering.stanford.edu/people/jeannette-bohg" target="_blank">
            Jeannette Bohg
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Assistant Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Artificial Intelligence
        <br/>
        Robotics
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://www.stanford.edu/~dabo/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">

```

```

        <a href="http://www.stanford.edu/~dabo/" target="_blank">
            Dan Boneh
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
        Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
        Computer Security
        <br/>
        Theory
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
        <a href="https://cs.stanford.edu/people/ebrun/" target="_blank"
>
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="https://cs.stanford.edu/people/ebrun/" target="_blank"
>
            Emma Brunskill
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
        Assistant Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
        Artificial Intelligence
        <br/>
        Machine Learning
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
        <a href="https://profiles.stanford.edu/moses-charikar" target="
_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="https://profiles.stanford.edu/moses-charikar" target="
_blank">
            Moses Charikar
        </a>
    </td>

```

```

        </td>
        <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
            Professor
        </td>
        <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
            Theory
        </td>
    </tr>
    <tr>
        <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
            <a href="http://cva.stanford.edu/billd_webpage_new.html" target
="_blank">
                
            </a>
        </td>
        <td class="views-field views-field-title fac_prof_view_name">
            <a href="http://cva.stanford.edu/billd_webpage_new.html" target
="_blank">
                Bill Dally
            </a>
        </td>
        <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
            Professor
        </td>
        <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
            Computer Systems
            <br/>
            Architecture
        </td>
    </tr>
    <tr>
        <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
            <a href="http://verify.stanford.edu/dill" target="_blank">
                
            </a>
        </td>
        <td class="views-field views-field-title fac_prof_view_name">
            <a href="http://verify.stanford.edu/dill" target="_blank">
                David Dill
            </a>
        </td>
        <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
            Professor
        </td>
        <td class="views-field views-field-field-research-focus fac_prof
_view_focus">

```

```

        Computational Biology
        <br/>
        Computer Systems
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://cs.stanford.edu/people/rondror/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://cs.stanford.edu/people/rondror/" target="_blank">
            Ron Dror
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Associate Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Artificial Intelligence
        <br/>
        Computational Biology
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="https://zakird.com/index.html" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="https://zakird.com/index.html" target="_blank">
            Zakir Durumeric
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Assistant Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Computer Systems
        <br/>
        Computer Security
    </td>
</tr>

```

```

<tr>
  <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
    <a href="http://web.stanford.edu/~engler/" target="_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="http://web.stanford.edu/~engler/" target="_blank">
      Dawson Engler
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_view_title">
    Associate Professor
  </td>
  <td class="views-field views-field-field-research-focus fac_prof_view_focus">
    Computer Systems
    <br/>
    Programming Systems and Verification
  </td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
    <a href="http://cs.stanford.edu/~ermon/" target="_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="http://cs.stanford.edu/~ermon/" target="_blank">
      Stefano Ermon
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_view_title">
    Assistant Professor
  </td>
  <td class="views-field views-field-field-research-focus fac_prof_view_focus">
    Artificial Intelligence
    <br/>
    Machine Learning
  </td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
    <a href="http://graphics.stanford.edu/~kayvonf/" target="_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="http://graphics.stanford.edu/~kayvonf/" target="_blank"
">
      Kayvon Fatahalian
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
    Assistant Professor
  </td>
  <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
    Computer Graphics
  </td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
    <a href="http://physbam.stanford.edu/~fedkiw/" target="_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="http://physbam.stanford.edu/~fedkiw/" target="_blank">
      Ron Fedkiw
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
    Professor
  </td>
  <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
    Computer Graphics
  </td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
    <a href="http://infolab.stanford.edu/people/hector.html" target
="_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="http://infolab.stanford.edu/people/hector.html" target
="_blank">
      Hector Garcia-Molina
    </a>
  </td>

```



```

<td class="views-field views-field-field-faculty-title fac_prof_
view_title">
  Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof_
_view_focus">
  Computer Systems
<br/>
  Data Science
</td>
</tr>
<tr>
<td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
  <a href="http://logic.stanford.edu/people/genesereth/genesereth
.html" target="_blank">
    
  </a>
</td>
<td class="views-field views-field-title fac_prof_view_name">
  <a href="http://logic.stanford.edu/people/genesereth/genesereth
.html" target="_blank">
    Mike Genesereth
  </a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_
view_title">
  Associate Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof_
_view_focus">
  Artificial Intelligence
</td>
</tr>
<tr>
<td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
  <a href="http://geometry.stanford.edu/member/guibas/index.html"
target="_blank">
    
  </a>
</td>
<td class="views-field views-field-title fac_prof_view_name">
  <a href="http://geometry.stanford.edu/member/guibas/index.html"
target="_blank">
    Leonidas Guibas
  </a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_
view_title">
  Professor
</td>
<td class="views-field views-field-field-research-focus fac prof

```

```

_view_focus">
    Artificial Intelligence
    <br/>
    Computer Graphics
</td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="https://graphics.stanford.edu/~hanrahan/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="https://graphics.stanford.edu/~hanrahan/" target="_blank">
            Patrick Hanrahan
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Computer Graphics
        <br/>
        Computer Systems
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://web.stanford.edu/~hennessy/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://web.stanford.edu/~hennessy/" target="_blank">
            John Hennessy
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Computer Systems
    </td>
</tr>
<tr>

```

```
  |
```

```

</td>
<td class="views-field views-field-title fac_prof_view_name">
  <a href="http://web.stanford.edu/~jurafsky/" target="_blank">
    Dan Jurafsky
  </a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_
view_title">
  Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof
_view_focus">
  Artificial Intelligence
  <br/>
  Natural Language Processing
</td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
    <a href="http://web.stanford.edu/~skatti/" target="_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="http://web.stanford.edu/~skatti/" target="_blank">
      Sachin Katti
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
    Assistant Professor
  </td>
  <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
    Computer Systems
    <br/>
    Networking
  </td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
    <a href="http://cs.stanford.edu/groups/manips/ok.html" target="
_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="http://cs.stanford.edu/groups/manips/ok.html" target="
_blank">
      Oussama Khatib
    </a>

```

```

    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
        Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
        Artificial Intelligence
        <br/>
        Robotics
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
        <a href="http://csl.stanford.edu/~christos/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://csl.stanford.edu/~christos/" target="_blank">
            Christos Kozyrakis
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
        Associate Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
        Computer Systems
        <br/>
        Architecture
        <br/>
        Operating/Distributed Systems
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
        <a href="https://sites.google.com/site/anshulkundaje/" target="
_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="https://sites.google.com/site/anshulkundaje/" target="
_blank">
            Anshul Kundaje
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_
view_title">

```

Assistant Professor	
	views-field views-field-field-research-focus fac_prof_view_focus
Computational Biology	
	views-field views-field-field-faculty-photo views-align-center border-simple
 	
	views-field views-field-title fac_prof_view_name
Monica Lam	
	views-field views-field-field-faculty-title fac_prof_view_title
Professor	
	views-field views-field-field-research-focus fac_prof_view_focus
Computer Systems	
	views-field views-field-field-faculty-photo views-align-center border-simple
 	
	views-field views-field-title fac_prof_view_name
James Landay	
	views-field views-field-field-faculty-title fac_prof_view_title
Professor	
	views-field views-field-field-research-focus fac_prof_view_focus
Human-Computer Interaction (HCI)	
	views-field views-field-field-faculty-photo views-align-center border-simple

```

gn-center border-simple">
    <a href="http://cs.stanford.edu/people/jure/" target="_blank">
        
    </a>
</td>
<td class="views-field views-field-title fac_prof_view_name">
    <a href="http://cs.stanford.edu/people/jure/" target="_blank">
        Jure Leskovec
    </a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_view_title">
    Assistant Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof_view_focus">
    Machine Learning
    <br/>
    Data Science
</td>
</tr>
<tr>
<td class="views-field views-field-field-faculty-photo views-align-center border-simple">
    <a href="http://csl.stanford.edu/~pal/" target="_blank">
        
    </a>
</td>
<td class="views-field views-field-title fac_prof_view_name">
    <a href="http://csl.stanford.edu/~pal/" target="_blank">
        Philip Levis
    </a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_view_title">
    Associate Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof_view_focus">
    Computer Systems
    <br/>
    Networking
</td>
</tr>
<tr>
<td class="views-field views-field-field-faculty-photo views-align-center border-simple">
    <a href="http://vision.stanford.edu/index.html" target="_blank">
        
    </a>

```

```

</td>
<td class="views-field views-field-title fac_prof_view_name">
  <a href="http://vision.stanford.edu/index.html" target="_blank"
>
    Fei-Fei Li
  </a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_
view_title">
  Associate Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof
_view_focus">
  Artificial Intelligence
  <br/>
  Computer Vision
  <br/>
  Machine Learning
</td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
    <a href="http://cs.stanford.edu/~pliang/" target="_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="http://cs.stanford.edu/~pliang/" target="_blank">
      Percy Liang
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
    Assistant Professor
  </td>
  <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
    Artificial Intelligence
    <br/>
    Machine Learning
    <br/>
    Natural Language Processing
  </td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
    <a href="http://nlp.stanford.edu/~manning/" target="_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">

```



```

    <a href="http://nlp.stanford.edu/~manning/" target="_blank">
      Chris Manning
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
    Professor
  </td>
  <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
    Artificial Intelligence
    <br/>
    Natural Language Processing
  </td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
    <a href="http://www.scs.stanford.edu/~dm/" target="_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="http://www.scs.stanford.edu/~dm/" target="_blank">
      David Mazières
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
    Professor
  </td>
  <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
    Computer Systems
    <br/>
    Computer Security
    <br/>
    Operating/Distributed Systems
  </td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
    <a href="http://tiny-tera.stanford.edu/~nickm/" target="_blank"
>
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="http://tiny-tera.stanford.edu/~nickm/" target="_blank"
>
      Nick McKeown
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
    Professor
  </td>
  <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
    Artificial Intelligence
    <br/>
    Natural Language Processing
  </td>
</tr>

```

```

        </td>
        <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
            Professor
        </td>
        <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
            Computer Systems
            <br/>
            Networking
        </td>
    </tr>
    <tr>
        <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
            <a href="http://theory.stanford.edu/people/jcm/home.html" targe
t="_blank">
                
            </a>
        </td>
        <td class="views-field views-field-title fac_prof_view_name">
            <a href="http://theory.stanford.edu/people/jcm/home.html" targe
t="_blank">
                John Mitchell
            </a>
        </td>
        <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
            Professor
        </td>
        <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
            Computer Security
        </td>
    </tr>
    <tr>
        <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
            <a href="http://web.stanford.edu/~subh/" target="_blank">
                
            </a>
        </td>
        <td class="views-field views-field-title fac_prof_view_name">
            <a href="http://web.stanford.edu/~subh/" target="_blank">
                Subhasish Mitra
            </a>
        </td>
        <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
            Associate Professor
        </td>
        <td class="views-field views-field-field-research-focus fac_prof
_view_focus">

```

```

        Computer Systems
    <br/>
    Architecture
    <br/>
    Programming Systems and Verification
</td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://www.andrewng.org/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://www.andrewng.org/" target="_blank">
            Andrew Ng
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Associate Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Artificial Intelligence
        <br/>
        Machine Learning
        <br/>
        Robotics
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://arsenalfc.stanford.edu/kunle/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://arsenalfc.stanford.edu/kunle/" target="_blank">
            Kunle Olukotun
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Computer Systems
        <br/>
        Architecture

```

```

        <br/>
        Programming Systems and Verification
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://web.stanford.edu/~ouster/cgi-bin/home.php" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://web.stanford.edu/~ouster/cgi-bin/home.php" target="_blank">
            John Ousterhout
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Computer Systems
        <br/>
        Operating/Distributed Systems
    </td>
</tr>
<tr>
    <td class="views-field views-field-field-faculty-photo views-align-center border-simple">
        <a href="http://web.stanford.edu/~balaji/" target="_blank">
            
        </a>
    </td>
    <td class="views-field views-field-title fac_prof_view_name">
        <a href="http://web.stanford.edu/~balaji/" target="_blank">
            Balaji Prabhakar
        </a>
    </td>
    <td class="views-field views-field-field-faculty-title fac_prof_view_title">
        Professor
    </td>
    <td class="views-field views-field-field-research-focus fac_prof_view_focus">
        Computer Systems
        <br/>
        Networking
    </td>
</tr>
<tr>

```

```
  |
```

```

</td>
<td class="views-field views-field-title fac_prof_view_name">
  <a href="http://web.stanford.edu/~mendel/" target="_blank">
    Mendel Rosenblum
  </a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_
view_title">
  Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof
_view_focus">
  Computer Systems
  <br/>
  Operating/Distributed Systems
</td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
    <a href="https://cs.stanford.edu/~aviad/" target="_blank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="https://cs.stanford.edu/~aviad/" target="_blank">
      Aviad Rubinstein
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_
view_title">
    Assistant Professor
  </td>
  <td class="views-field views-field-field-research-focus fac_prof
_view_focus">
    Theory
  </td>
</tr>
<tr>
  <td class="views-field views-field-field-faculty-photo views-ali
gn-center border-simple">
    <a href="https://profiles.stanford.edu/dorsa-sadigh" target="_b
lank">
      
    </a>
  </td>
  <td class="views-field views-field-title fac_prof_view_name">
    <a href="https://profiles.stanford.edu/dorsa-sadigh" target="_b
lank">
      Dorsa Sadigh
    </a>
  </td>
  <td class="views-field views-field-field-faculty-title fac_prof_

```

```

view_title">
    Assistant Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof
_view_focus">
    Artificial Intelligence
<br/>
    Robotics
</td>
</tr>
<tr>
<td class="views-field views-field-field-faculty-photo views-align-center border-simple">
<a href="http://robotics.stanford.edu/users/sahami/bio.html" target="_blank">

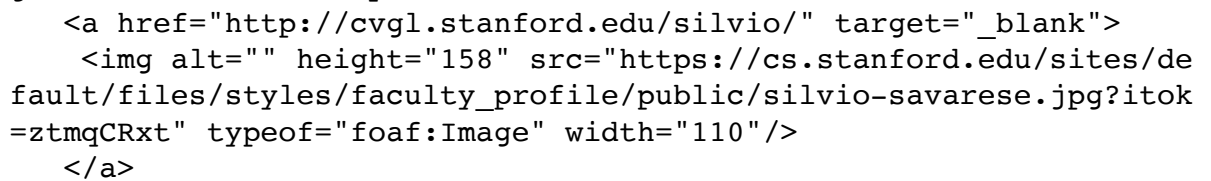
</a>
</td>
<td class="views-field views-field-title fac_prof_view_name">
<a href="http://robotics.stanford.edu/users/sahami/bio.html" target="_blank">
    Mehran Sahami
</a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_view_title">
    Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof_view_focus">
    Education
</td>
</tr>
<tr>
<td class="views-field views-field-field-faculty-photo views-align-center border-simple">
<a href="http://robotics.stanford.edu/~jks/" target="_blank">

</a>
</td>
<td class="views-field views-field-title fac_prof_view_name">
<a href="http://robotics.stanford.edu/~jks/" target="_blank">
    Kenneth Salisbury
</a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_view_title">
    Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof_view_focus">
    Artificial Intelligence
<br/>

```

Robotics	
----------	--

--	--

	http://cvgl.stanford.edu/silvio/ Silvio Savarese	Assistant Professor	Artificial Intelligence Computer Vision Robotics
--	--	---------------------	--

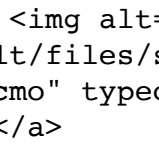
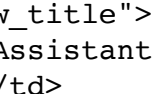
	http://robotics.stanford.edu/~shoham/ Yoav Shoham	Professor	Artificial Intelligence
--	--	-----------	-------------------------

--	--


```

gn-center border-simple">
    <a href="http://theory.stanford.edu/~liyang/" target="_blank">
        
    </a>
</td>
<td class="views-field views-field-title fac_prof_view_name">
    <a href="http://theory.stanford.edu/~liyang/" target="_blank">
        Li-Yang Tan
    </a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_view_title">
    Assistant Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof_view_focus">
    Theory
</td>
</tr>
<tr>
<td class="views-field views-field-field-faculty-photo views-align-center border-simple">
    <a href="http://robots.stanford.edu/" target="_blank">
        
    </a>
</td>
<td class="views-field views-field-title fac_prof_view_name">
    <a href="http://robots.stanford.edu/" target="_blank">
        Sebastian Thrun
    </a>
</td>
<td class="views-field views-field-field-faculty-title fac_prof_view_title">
    Professor
</td>
<td class="views-field views-field-field-research-focus fac_prof_view_focus">
    Artificial Intelligence
    <br/>
    Machine Learning
    <br/>
    Robotics
</td>
</tr>
<tr>
<td class="views-field views-field-field-faculty-photo views-align-center border-simple">
    <a href="http://theory.stanford.edu/~valiant/" target="_blank">
        
    </a>
</td>

```

Gregory Valiant	Assistant Professor	Theory	
Jennifer Widom	Professor	Computer Systems Data Science	
Keith Winstein	Assistant Professor		

<div> <div>views-field views-field-field-research-focus fac_prof_view_focus</div> <div>Computer Systems</div> <div></div> <div>Networking</div> </div>
<div> <div>views-field views-field-field-faculty-photo views-align-center border-simple</div> <div> </div> </div> <div> <div>views-field views-field-title fac_prof_view_name</div> <div>Mary Wootters</div> </div> <div> <div>views-field views-field-field-faculty-title fac_prof_view_title</div> <div>Assistant Professor</div> </div> <div> <div>views-field views-field-field-research-focus fac_prof_view_focus</div> <div>Theory</div> </div>
<div> <div>views-field views-field-field-faculty-photo views-align-center border-simple</div> <div> </div> </div> <div> <div>views-field views-field-title fac_prof_view_name</div> <div> Matei Zaharia </div> </div> <div> <div>views-field views-field-field-faculty-title fac_prof_view_title</div> <div>Assistant Professor</div> </div> <div> <div>views-field views-field-field-research-focus fac_prof_view_focus</div> <div>Computer Systems</div> <div></div> <div>Operating/Distributed Systems</div> <div></div> <div>Data Science</div> </div>

Draw bar charts on research area contributions of Stanford CS faculty

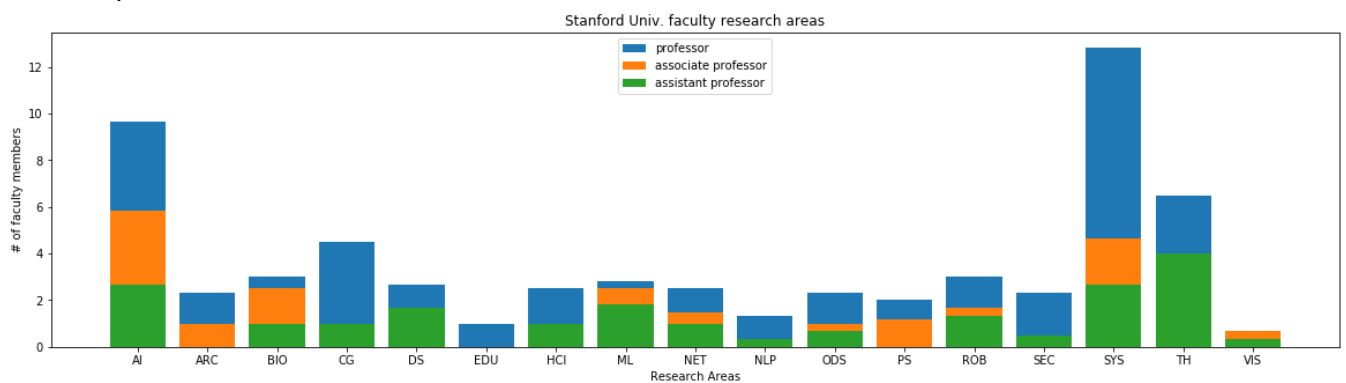
- The followings are research fields

```
['Architecture', 'Artificial Intelligence', 'Computational Biology',  
'Computer Graphics', 'Computer Security', 'Computer Systems', 'Computer Vision', 'Data Science', 'Education', 'Human-Computer Interaction (HCI)', 'Machine Learning', 'Natural Language Processing', 'Networking', 'Operating/Distributed Systems', 'Programming Systems and Verification', 'Robotics', 'Theory']
```

- In plotting, use the following abbreviations:

```
['ARC', 'AI', 'BIO', 'CG', 'SEC', 'SYS', 'VIS', 'DS', 'EDU', 'HCI', 'ML',  
'NLP', 'NET', 'ODS', 'PS', 'ROB', 'TH']
```

- 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 [7]: # YOUR CODE MUST BE HERE
#.replace('<br/>', '\\\')
def cleaning(L):
    faculty = '<td class="views-field views-field-field-faculty-title fac_prof_view_title">'
    focus = '<td class="views-field views-field-field-research-focus fac_prof_view_focus">'
    L=str(L)
    L=L.replace('<td>','').replace('</td>','')

    if faculty in L:
        L=L.replace(faculty,'')
    elif focus in L:
        L=L.replace(focus,'').replace('<br/>','&')

    L=L.replace('\n','').replace('[','').replace(']','')
    return L.split(',')

field=['Architecture', 'Artificial Intelligence', 'Computational Biology', 'Computer Graphics', 'Computer Security', 'Computer Systems', 'Computer Vision', 'Data Science',
        'Education', 'Human-Computer Interaction (HCI)', 'Machine Learning', 'Natural Language Processing', 'Networking', 'Operating/Distributed Systems',
        'Programming Systems and Verification', 'Robotics', 'Theory
']

Prof, Area = [], []
for i in cleaning(soup.body.findAll('td', {"class":"views-field views-field-field-faculty-title fac_prof_view_title"})):
    Prof.append(i.strip())
for i in cleaning(soup.body.findAll('td', {"class":"views-field views-field-field-research-focus fac_prof_view_focus"})):
    Area.append(i.strip())
Dic=list(zip(Prof,Area))
print(Dic)

```

[('Professor', 'Computer Graphics&Human-Computer Interaction (HCI)'), ('Professor', 'Computer Systems&Programming Systems and Verification'), ('Assistant Professor', 'Computer Systems&Operating/Distributed Systems&Data Science'), ('Associate Professor', 'Computer Systems&Architecture&Programming Systems and Verification'), ('Associate Professor', 'Artificial Intelligence&Computational Biology'), ('Associate Professor', 'Artificial Intelligence&Computational Biology'), ('Assistant Professor', 'Human-Computer Interaction (HCI)'), ('Assistant Professor', 'Artificial Intelligence&Robotics'), ('Professor', 'Computer Security&Theory'), ('Assistant Professor', 'Artificial Intelligence&Machine Learning'), ('Professor', 'Theory'), ('Professor', 'Computer Systems&Architecture'), ('Professor', 'Computational Biology&Computer Systems'), ('Associate Professor', 'Artificial Intelligence&Computational Biology'), ('Assistant Professor', 'Computer Systems&Computer Security'), ('Associate Professor', 'Computer Systems&Programming Systems and Verification'), ('Assistant Professor', 'Artificial Intelligence&Machine Learning'), ('Assistant Professor', 'Computer Graphics'), ('Professor', 'Computer Graphics'), ('Professor', 'Computer Systems&Data Science'), ('Associate Professor', 'Artificial Intelligence'), ('Professor', 'Artificial Intelligence&Computer Graphics'), ('Professor', 'Computer Graphics&Computer Systems'), ('Professor', 'Computer Systems'), ('Professor', 'Computer Systems&Architecture'), ('Professor', 'Computer Graphics'), ('Professor', 'Artificial Intelligence&Natural Language Processing'), ('Assistant Professor', 'Computer Systems&Networking'), ('Professor', 'Artificial Intelligence&Robotics'), ('Associate Professor', 'Computer Systems&Architecture&Operating/Distributed Systems'), ('Assistant Professor', 'Computational Biology'), ('Professor', 'Computer Systems'), ('Professor', 'Human-Computer Interaction (HCI)'), ('Assistant Professor', 'Machine Learning&Data Science'), ('Associate Professor', 'Computer Systems&Networking'), ('Associate Professor', 'Artificial Intelligence&Computer Vision&Machine Learning'), ('Assistant Professor', 'Artificial Intelligence&Machine Learning&Natural Language Processing'), ('Professor', 'Artificial Intelligence&Natural Language Processing'), ('Professor', 'Computer Systems&Computer Security&Operating/Distributed Systems'), ('Professor', 'Computer Systems&Networking'), ('Professor', 'Computer Security'), ('Associate Professor', 'Computer Systems&Architecture&Programming Systems and Verification'), ('Associate Professor', 'Artificial Intelligence&Machine Learning&Robotics'), ('Professor', 'Computer Systems&Architecture&Programming Systems and Verification'), ('Professor', 'Computer Systems&Operating/Distributed Systems'), ('Professor', 'Computer Systems&Networking'), ('Assistant Professor', 'Computer Systems&Data Science'), ('Professor', 'Theory'), ('Professor', 'Computer Systems&Operating/Distributed Systems'), ('Assistant Professor', 'Theory'), ('Assistant Professor', 'Artificial Intelligence&Robotics'), ('Professor', 'Education'), ('Professor', 'Artificial Intelligence&Robotics'), ('Assistant Professor', 'Artificial Intelligence&Computer Vision&Robotics'), ('Professor', 'Artificial Intelligence'), ('Assistant Professor', 'Theory'), ('Professor', 'Artificial Intelligence&Machine Learning&Robotics'), ('Assistant Professor', 'Theory'), ('Professor', 'Computer Systems&Data Science'), ('Assistant Professor', 'Computer Systems&Networking'), ('Assistant Professor', 'Theory'), ('Assistant Professor', 'Computer Systems&Operating/Distributed Systems&Data Science')]

Ethics:

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

What to submit

- Run all cells
- Goto "File -> Print Preview"
- Print the page
- Submit in class
- No late homeworks accepted
- Your homework will be graded on the basis of correctness and programming skills

Deadline: 4/18