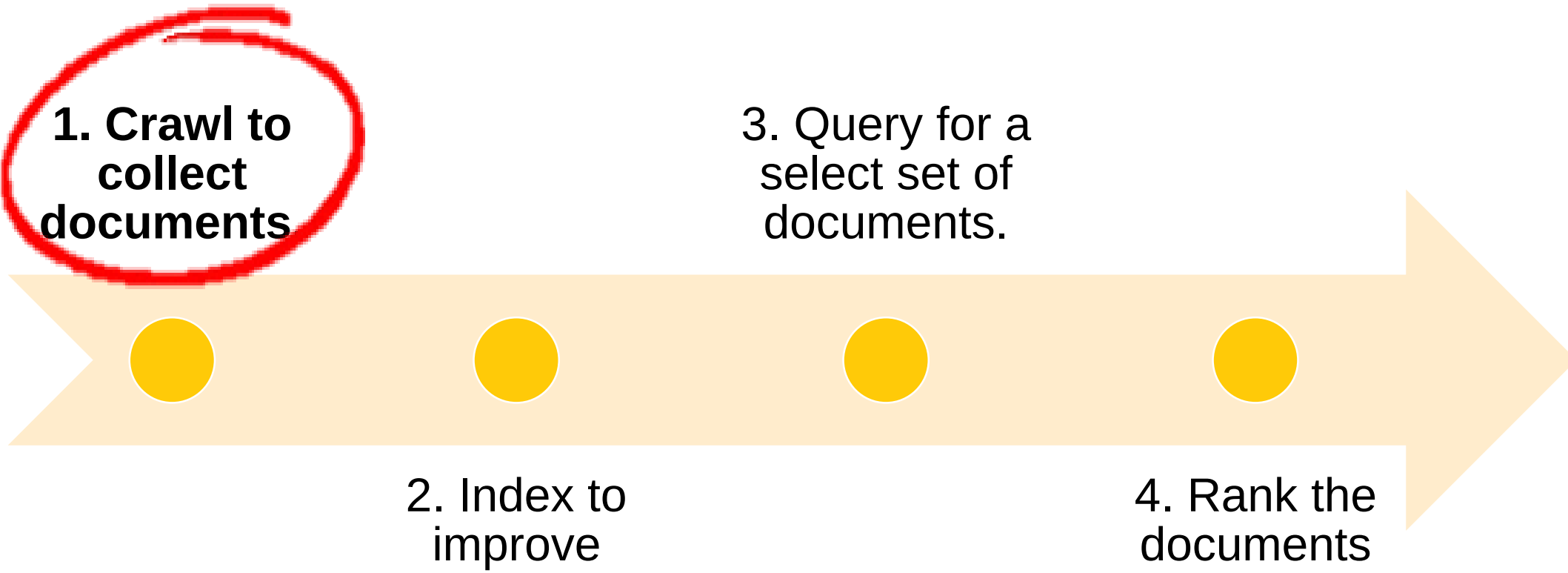
# ENGR 102 PROGRAMMING PRACTICE

## WEEK 10

# Searching & Ranking

# Search Engine



# Search Engine

- Create a Python module (`mysearchengine.py`).

- The module will have two classes:
  - one for crawling and creating the database, and
  - the other for doing full-text searches by querying the database, as well as ranking.

# Crawler Code

- **`requests`**: download web pages
- **`BeautifulSoup`**: build a structured representation of web pages.
- Using **`requests`** and **`BeautifulSoup`**, you can build a crawler that will take a list of URLs to index and crawl their links to find other pages to index.

# Using requests module

- Makes it easy to download web pages
- Input: a URL

```
import requests

r = requests.get('http://cs.sehir.edu.tr')
print(r.content[0:50])
```

# BeautifulSoup

- Parse a web page and build a structured representation.
- Access any element of the page by
  - type, ID, or any of its properties
  - get a string representation of its contents.
- Install module: **beautifulsoup4**

# Beautiful Soup

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')
print(soup.prettify())
```

```
html_doc = """
<html><head><title>The Dormouse'
<body>
<p class="title"><b>The Dormouse

<p class="story">Once upon a tim
<a href="http://example.com/elsi
<a href="http://example.com/laci
<a href="http://example.com/till
and they lived at the bottom of

<p class="story">...</p>
</body>
</html>
"""
```

```
<html>
 <head>
  <title>
   The Dormouse's story
  </title>
 </head>
 <body>
  <p class="title">
   <b>
    The Dormouse's story
   </b>
  </p>
  <p class="story">
   Once upon a time there were three little sisters; and their names were
   <a class="sister" href="http://example.com/elsie" id="link1">
    Elsie
   </a>
   ,
   <a class="sister" href="http://example.com/lacie" id="link2">
    Lacie
   </a>
   and
   <a class="sister" href="http://example.com/tillie" id="link3">
    Tillie
   </a>
   ;
and they lived at the bottom of a well.
  </p>
  <p class="story">
   ...
  </p>
 </body>
</html>
```

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

# Beautiful Soup

```
<html>
 <head>
  <title>
   The Dormouse's story
  </title>
 </head>
 <body>
  <p class="title">
   <b>
    The Dormouse's story
   </b>
  </p>
  <p class="story">
   Once upon a time there were three little
   <a class="sister" href="http://example.co
    Elsie
   </a>
   ,
   <a class="sister" href="http://example.co
    Lacie
   </a>
   and
   <a class="sister" href="http://example.co
    Tillie
   </a>
   ;
and they lived at the bottom of a well.
  </p>
  <p class="story">
   ...
  </p>
 </body>
</html>
```

```
print(soup.title)
# <title>The Dormouse's story</title>

print(soup.title.name)
# 'title'

print(soup.title.string)
# 'The Dormouse's story'

print(soup.title.parent.name)
# 'head'

print(soup.p)
# <p class="title"><b>The Dormouse's story</b></p>

print(soup.p['class'])
# 'title'
```

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

# Beau

```
<html>
 <head>
  <title>
   The Dormouse's st
  </title>
 </head>
 <body>
  <p class="title">
   <b>
    The Dormouse's s
   </b>
  </p>
  <p class="story">
   Once upon a time there were three little sisters; and their names were
   <a class="sister" href="http://example.com/elsie" id="link1">
    Elsie
   </a>
   ,
   <a class="sister" href="http://example.com/lacie" id="link2">
    Lacie
   </a>
   and
   <a class="sister" href="http://example.com/tillie" id="link3">
    Tillie
   </a>
   ;
and they lived at the bottom of a well.
  </p>
  <p class="story">
   ...
  </p>
 </body>
</html>
```

```
print(soup.a)
# <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>

print(soup.find_all('a'))
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]

print(soup.find(id="link3"))
# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>

for link in soup.find_all('a'):
    print(link.get('href'))
# http://example.com/elsie
# http://example.com/lacie
# http://example.com/tillie
```

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

# Beautiful Soup

```
<html>
 <head>
  <title>
   The Dormouse's story
  </title>
 </head>
 <body>
  <p class="title">
   <b>
    The Dormouse's story
   </b>
  </p>
  <p class="story">
   Once upon a time there were three little sisters; and their names were
   <a class="sister" href="http://example.com/elsie" id="link1">
    Elsie
   </a>
   ,
   <a class="sister" href="http://example.com/lacie" id="link2">
    Lacie
   </a>
   and
   <a class="sister" href="http://example.com/tillie" id="link3">
    Tillie
   </a>
   ;
and t
  </p
  <p
   ..
  </p
 </bo
</htm
```

```
print(soup.find_all('b'))
# [<b>The Dormouse's story</b>]
```

```
print(soup.find_all(["a", "b"]))
# [<b>The Dormouse's story</b>,
#  <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]
```

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

# Beautiful Soup

```
<html>
 <head>
  <title>
   The Dormouse's story
  </title>
 </head>
 <body>
  <p class="title">
   <b>
    The Dormouse's story
   </b>
  </p>
  <p class="story">
   Once upon a time there were three little sisters; and their names were
   <a class="sister" href="http://example.com/elsie" id="link1">
    Elsie
   </a>
   ,
   <a class="sister" href="http://example.com/lacie" id="link2">
    Lacie
   </a>
   and
   <a class="sister" href="http://example.com/tillie" id="link3">
    Tillie
   </a>
   ;
and they lived at the bottom of a well.
  </p>
  <p class="story">
   ...
  </p>
 </body>
</html>
```

```
import re
for tag in soup.find_all(re.compile("^b")):
    print(tag.name)
# body
# b
```

```
for tag in soup.find_all(re.compile("t")):
    print(tag.name)
# html
# title
```

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

# Beautiful Soup

```
<html>
 <head>
  <title>
   The Dormouse's story
  </title>
 </head>
 <body>
  <p class="title">
   <b>
    The Dormouse's story
   </b>
  </p>
  <p class="story">
   Once upon a time there were three little sisters; and their names were
   <a class="sister" href="http://example.com/elsie" id="link1">
    Elsie
   </a>
   ,
   <a class="sister" href="http://example.com/lacie" id="link2">
    Lacie
   </a>
   and
   <a class="sister" href="http://example.com/tillie" id="link3">
    Tillie
   </a>
   ;
and they lived at the bottom of a well.
  </p>
  <p class="story">
   ...
  </p>
 </body>
</html>
```

```
def has_class_but_no_id(tag):
    return tag.has_attr('class') and not tag.has_attr('id')

print(soup.find_all(has_class_but_no_id))
# [<p class="title"><b>The Dormouse's story</b></p>,
#  <p class="story">Once upon a time there were…bottom of a well.</p>,
#  <p class="story">...</p>]
```

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

# Beautiful Soup

```
<html>
 <head>
  <title>
   The Dormouse's story
  </t
 </hea
 <body
  <p
   <b
    The Dormouse's story
   </b>
  </p>
  <p class="story">
   Once upon a time there were three little sisters; and their names were
   <a class="sister" href="http://example.com/elsie" id="link1">
    Elsie
   </a>
   ,
   <a class="sister" href="http://example.com/lacie" id="link2">
    Lacie
   </a>
   and
   <a class="sister" href="http://example.com/tillie" id="link3">
    Tillie
   </a>
   ;
and they lived at the bottom of a well.
  </p>
  <p class="story">
   ...
  </p>
 </body>
</html>
```

```
import re
print(soup.find(string=re.compile("sisters")))
# 'Once upon a time there were three little sisters; and their names were\n'
```

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

# Beautiful Soup

```
<html>
 <head>
  <title>
   The Dormouse's story
  </title>
 </head>
 <body>
  <p class="title">
   <b>
    The Dormouse's story
   </b>
  </p>
  <p class="story">
   Once upon a time there were thr
   <a class="sister" href="http://example.com/elsie" id="link1">
    Elsie
   </a>
   ,
   <a class="sister" href="http://example.com/lacie" id="link2">
    Lacie
   </a>
   and
   <a class="sister" href="http://example.com/tillie" id="link3">
    Tillie
   </a>
   ;
and they lived at the bottom of a well.
  </p>
  <p class="story">
   ...
  </p>
 </body>
</html>
```

```
print(soup.find_all(class_=re.compile("itl")))
# [<p class="title"><b>The Dormouse's story</b></p>]

def has_six_characters(css_class):
    return css_class is not None and len(css_class) == 6

print(soup.find_all(class_=has_six_characters))
# [<a class="sister" href="http://example.com/elsie" id="l
#  <a class="sister" href="http://example.com/lacie" id="l
#  <a class="sister" href="http://example.com/tillie" id="
```

https://www.crummy.com/software/BeautifulSoup/bs4/doc/