BAIS:6100 Text Analytics

Text Data Collection

Kang-Pyo Lee

Course Schedule (Subject to Change)

Week	Date	Topics	Due
1	Jan 28	Introduction to Text Analytics Introduction to Python, Jupyter Notebook, and UI Interactive Data Analytics Service (IDAS)	
2	Feb 4	Module 1. Python Basics for Text Processing, Part 1 : Strings, Collections, Built-in Functions, Flow Control, and User-Defined Functions	
3	Feb 11	Module 2. Python Basics for Text Processing, Part 2 : Files, Dataframes, and Pattern Matching Using Regular Expressions	HW 1
4	Feb 18	Module 3. Basic Natural Language Processing (NLP) Techniques : Tokenization, Part-of-Speech Tagging, Stemming, Lemmatization, N-grams, Noun Phrase Extraction, Language Detection and Translation, and Gender Prediction Module 4. Keyword Analysis and Visualization	HW 2
5	Feb 25	Test 1	HW 3 (Feb 24)
6	Mar 4	Modules 5 & 6. Text Data Collection Using Twitter APIs and Web Scraping Group Project Announcement	
7	Mar 11	Module 7. Document-Term Representation Module 8. Text Classification	Hw 4
8	Mar 18	Module 9. Text Clustering and Topic Modeling	Project Proposal
9	Mar 25	Module 10. Text Similarity Module 11. Keyword Network Analysis	
10	Apr 1	Test 2	HW 5 (Mar 30)
11	Apr 8	Group Project Presentations and Course Wrap-Up	Project Deliverables

Today's Topic

Text Data Collection

- Module 5. Using Twitter APIs
  - Step 1. Get Twitter API credentials (instructions)
  - Step 2. Communicate with Twitter API using Python (Jupyter notebook available on IDAS)
- **Module 6. Web Scraping**

Homework

Homework 4, which corresponds to Module 6, is due at 6:00 PM on Thu, Mar 11 via ICON Assignments

5 questions, 7 points in total

You will receive a 20% deduction for each day that it is late, including the first/same day

What Is Web Scraping?

Web scraping is data scraping used for extracting data from websites

from Wikipedia

https://en.wikipedia.org/wiki/Web_scraping

Why Web Scraping?

Web scraping is useful when you need to extract information directly from web pages

E.g., you may want to collect data from

- **Google search results**
- **New York Times news articles**
- **Amazon customer reviews**

Why Web Scraping?

Web scraping should be
the last resort you consider

- File download
- Database access
- Application Programming Interfaces (APIs)

} Always preferred over web scraping!

They provide easier access to their data in a more structured manner

What Is a Web Crawler?

A web crawler (aka web spider) is computer software that systematically browses websites and downloads content

Important Things to Consider

- **Being able to collect data from a website does NOT mean that you're allowed to use the data at your disposal**
- **The responsibility of the data falls on the user when collected**
- **You must adhere to the terms of service**
  - E.g., https://www.nytimes.com/content/help/rights/terms/terms-of-service.html#b
  - In general, if you want to use the collected data for
    - personal uses → usually fine
    - **research uses → generally fine, but still be careful**
    - commercial uses → usually not allowed without prior approval
- **You must not violate the politeness policy that states how to avoid overloading web sites**
  - Avoid making massive calls to a website
  - Be careful when you need to collect data from a website on a large scale or in a continuous manner

Limitations of Web Scraping

- **Not easy to automate the web scraping process**
  - Developing and maintaining web scrapers is expensive
    - Each web site has its own structure and there is no "one size fits all" approach in extracting data from websites → You would have to implement a web scraper for each webpage/website
    - Scrapers stop working if the website structure has changed
    - Scrapers could be blocked by the website admin
  - You may need a dedicated server running 24/7 for continuous data scraping
- **Some websites have dynamic content or can identify machine access, which makes web scraping tricky or impossible**
  - We need to simulate human browsing to enable gathering webpage content
  - Any content that can be viewed on a webpage can be scraped

Process of Web Scraping

1. Identify the target webpages and the information to extract

2. Understand the HTML structure of the webpages

Use *Developer Tools* in Chrome browser

3. Fetch the whole content from each webpage

Use *requests* package in Python

4. Save the content as an HTML file in your computer

5. Read the content from each file

6. Parse the HTML code and get the necessary information

Use *beautifulsoup* package in Python

Basics of HTML

HyperText Markup Language (HTML) is the standard markup language for creating webpages and web applications

from Wikipedia

HTML5 is the fifth and current major version of the HTML standard

Basics of HTML

```
<!DOCTYPE html>
<html>
 <head>
  <title>This is a title</title>
 </head>
 <body>
  <h1>This is a heading</h1>
  <p>Hello world!</p>
 </body>
</html>
```

- HTML describes the structure of webpages using markup
- HTML elements are the building blocks of HTML pages and can be nested
- HTML elements are represented by tags surrounded by matching angle brackets
- Web browsers do not display the HTML tags, but use them to render the content of the page
- HTML tags most commonly come in pairs
- The first tag in such a pair is the start tag, or opening tag, and the second is the end tag, or closing tag

https://www.w3schools.com/html/html_intro.asp

Basics of HTML

```
<!DOCTYPE html>
<html>
 <head>
  <title>This is a title</title>
 </head>
 <body>
  <h1>This is a heading</h1>
  <p>Hello world!</p>
 </body>
</html>
```

- <!DOCTYPE html> defines this document to be HTML5
- The <html> element is the root element of an HTML page
- The <head> element contains meta information about the document
- The <title> element specifies a title for the document
- The <body> element contains the visible page content
- The <h1> element defines the largest heading
- The <p> element defines a paragraph

https://www.w3schools.com/html/html_intro.asp

Basics of HTML

```
<a href="https://www.google.com">A link to Google</a>
```

- Links are defined with the <a> tag, or anchor tag
- The href attribute holds the URL address of the link
- The text between <a> and describes the text that leads to the specified URL address

https://www.w3schools.com/tags/tag_a.asp

Basics of HTML

```
<table style="width:100%">
  <tr>
    <th>Column 1</th>
    <th>Column 2</th>
  </tr>
  <tr>
    <td>Value 1</td>
    <td>Value 2</td>
  </tr>
</table>
```

- Tables are defined with the <table> tag
- The <tr> element defines a table row
- The <th> element defines a table header
- The <td> element defines a table cell
- You should be consistent in the numbers of <th> elements and <td> elements

https://www.w3schools.com/tags/tag_table.asp

Basics of HTML

```
<div style="color:blue">
  <h3>This is a heading</h3>
  <p>This is a paragraph.</p>
</div>
```

- The tag defines a division or a section in an HTML document
- It is used to group block-elements to format them with CSS

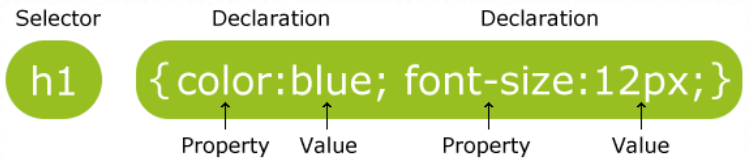
https://www.w3schools.com/tags/tag_div.asp

CSS

```
<h1 style="color:blue;">This is a Blue Heading</h1>
```
Inline CSS

```
<!DOCTYPE html>
<html>
 <head>
  <style>
  body {background-color: black;}
  h1   {color: blue;}
  p    {color: red;}
  </style>
 </head>
 <body>
  <h1>This is a heading</h1>
  <p>This is a paragraph.</p>
 </body>
</html>
```
Internal CSS

- CSS (Cascading Style Sheets) describes how HTML elements are to be displayed on screen
- CSS saves a lot of work, it can control the layout of multiple webpages all at once
- Three types of CSS
  - Inline - using the style attribute in HTML elements
  - Internal - using a <style> element in the <head> section
  - External - using an external CSS file



https://www.w3schools.com/html/html_css.asp

JavaScript

```
<head>
 <script>
 function myFunction() {
    document.getElementById("demo").innerHTML = "Paragraph changed.";
 }
 </script>
</head>
```

- JavaScript, or JS, is the programming language of HTML and the Web
- The three core technologies of World Wide Web content production
  - HTML to define the content of web pages
  - CSS to specify the layout of web pages
  - JavaScript to program the behavior of web pages
- Often used to make webpages interactive and dynamic
- Not so useful to know for web scraping

https://www.w3schools.com/js/default.asp

Useful Features of Google Chrome Browser

Right Click > View Page Source

Ctrl + Shift + I on Windows
Option + Command + I on Mac
for Developer Tools

Right Click > Inspect