Natural Language Processing

Session 2: **Python Refresher**

Instructor: Behrooz Mansouri
Spring 2023, University of Southern Maine

Previous Session

In previous session we learned about:

- ✓ What is Natural Language Processing
- ✓ What makes Natural Language Processing hard
- ✓ Natural Language Processing Tasks

Google Colab

Google Colaboratory (Colab)

- An executable document allowing you to write, run, and share code on
  - Google Drive (Jupyter notebook stored in Google Drive)
  - GitHub
- A notebook document is composed of cells: Code, text, image or …
- Colab connects your notebook to a cloud-based runtime
  - no required setup on your own machine
- Able to use Code snippets and insert them in your own code

Exercise 1

1. Print odd numbers in range 0 to 100
2. Write a function that returns summation of even numbers in a range (start, end) and both start and end are included in the summation
3. Define an empty list
   a. Add 10 random integer numbers in range [1,100] to your list
   b. Print the maximum number in the list
   c. Print the minimum number in the list
   d. Sort the list
   e. Randomly shuffle the list
   f. Try running the sort part of code with “Run the focused cell”
4. Define a dictionary as: {‘a’: 1, ‘d’:4, ‘b’:2, ‘c’:3}
   a. Sort the dictionary by keys
   b. Sort the dictionary by values in descending order
   c. Add new tuple to the dictionary {‘e’: 5}

Exercise 2

Go to TF-IDF Wikipedia page and copy the first 3-Sentences (above motivation)

Test turning the debug mode on and off with command: !pdb on and !pdb off

a. Define a string with this paragraph as the value
b. Write a program that prints the number of total words and the number of total unique words in this paragraph
c. Which word has the most frequency? (write a program for this)
d. Read the TFIDF.txt file as the string and test if you will get the same results
e. Use NLTK library and remove the stop-words. What is the most frequent word now?

Exercise 3

Target file: quotes.tsv (tab-separated values)

Read the TSV file with pandas library. Install the library with command: ! pip install pandas

Answer the following questions

a. How many quotes are from ‘Alexandre Dumas’?
b. Who has the longest quote? (number of words)
c. In whose quote there is the word ‘one’? Name the author(s)
d. What are the most and least frequent words used in all the quotes?

Exercise 4

Target file: ‘Posts_Coffee.xml’ (Snapshot of https://coffee.stackexchange.com/)

Use the code ‘post_reader.py’ file from here:

https://github.com/ARQMath/ARQMathCode/blob/master/Entity_Parser_Record/post_parser_record.py

Answer the following questions:

a. How many questions/answers have been posted on this website?
b. How many questions, have accepted answers?
c. What is the highest/lowest score given to the questions?
d. What is the average number of words in answers?
e. In which year, the highest number of posts (questions/answers) were posted?

Crawling the Internet

Crawling the internet, also known as web scraping, is the process of automatically collecting information from websites. There are several reasons why one might want to crawl the internet, including:

- Data collection for research or analysis
- Price comparison for online shopping
- Content aggregation for news or media websites
- Creating a search engine

There are several Linux and Windows command line tools that can be used for web scraping, including:

- Linux: wget, curl, and scrapy
- Windows: cURL, wget for Windows, and Scrapy (which is also available on Windows)

Also, there are several web scraping libraries and frameworks available for python such as BeautifulSoup, Scrapy and Selenium which can be useful depending on the complexity of scraping required

The Internet Archive

The Internet Archive is a non-profit organization that was founded in 1996 with the goal of providing universal access to all knowledge. The organization's main website, archive.org, is a digital library that offers free access to millions of books, videos, audio recordings, software, and other cultural artifacts from around the world.

The Internet Archive has several main features:

- **The Wayback Machine**: allows users to access archived versions of websites dating back to the early days of the internet.
- **The Digital Library**: contains millions of books, videos, audio recordings, and other cultural artifacts that can be accessed and downloaded for free.
- **The Audio Archive**: contains over 4 million audio recordings, including music, spoken word, and live recordings.
- **The Video Archive**: contains thousands of videos, including movies, television shows, and educational videos.
- **The Software Collection**: contains thousands of historical software programs, including games, applications, and operating systems.

Stack Exchanges

Stack Exchange is a network of question-and-answer websites on a variety of topics, where users can ask and answer questions and vote on the best answers

The most well-known website in the network is Stack Overflow, which is focused on programming and computer science. Other websites in the network cover a wide range of topics, such as physics, mathematics, and English language learning

The goal of Stack Exchange is to create a library of high-quality, community-driven questions and answers on a wide range of topics

https://stackexchange.com/sites
https://archive.org/download/stackexchange

Next Session

Regular Expressions

In the next session, we will learn Regular expressions and Automata

To do before next session:

- Chapter 2 Jurafsky (Link)