DSCI-560 Lab 2

NOTE: You must use GitHub to store all source code and reports associated with the lab assignments. You must commit/push all modifications to the assignments with detailed descriptions at least once daily while working on the lab assignment. Please submit the GitHub link and a record of your commits, along with detailed commit descriptions, to Coursistant before the due date. The tools and libraries listed in this assignment are simply suggestions. Please use whatever tools and libraries are necessary to complete the project.

1. **Team Formation**

   All the assignments should be performed in teams of 3. Form a team of 3 students (no more, no less) and decide on a name for your team. ***Pick a team name and include the names and USC ID of all three members***. Each member must have the same details in their assignment submission for this week.

   Having teams created by the assignment deadline would help your team plan and start working on the next assignment as soon as it is released, so your team has maximum time to work on the assignments.

2. **Real-World Applications and Domain-Specific Data**

   As I have discussed during the lecture, your final project will be a group chat/messaging system with a live-in AI agent that will assist you in the way you communicate with others in the group and provide you with all of the services that you would like it to perform for you in a specific domain associated with the group. For example, the Parents of the High School Athletes might want to communicate via a group chat system. But in addition to chatting with others, there are common services that they would use as a group. The additional services include efficient ways to organize fundraisers, create and sign up for volunteer activities at the sports event, etc. Another example might be a classroom, where students might ask other students, team members, or instructors about specific class topics. While there may be specific systems that cater to these domains, they often require manual configurations and customizations by someone who knows how to use the tools and services.

   For your final project, I would like to reduce the burdens of having to learn and use the tools from the users. You will attempt to achieve this by building a smart AI agent that will assist the users to communicate and provide domain specific services. In essence, you will build an assistant that will recognize and do everything for you through the simplest but one of the most popular ways of communicating today, texting/messaging. I would like you to come up with the most creative ways to communicate with one another. So simple that you won't even have to select the name out of your contact list. One example task of your AI might be figuring out who to send your message. This might be done through AI asking your questions. It might be done by looking at the historical behavior of the user. It might also be done by looking at the message content. In addition to messaging, the agent

ought to be able to answer most users' questions using a domain-specific knowledge base. Lastly, the system should automate a common set of services used by the group.

For this assignment, you will discuss with your teammates your preference and select a specific domain. Brainstorm with your team what data might be relevant for the final project and how you might find and collect data. To ensure individual contribution, document your individual shortlisted options for a domain and the reasoning behind the kinds of data that you will incorporate.

You will learn to use the necessary tools and APIs to build, evaluate, and improve the quality of data used in your system. Data sources should be publicly available datasets, including ASCII text in forums/applications, office documents, websites, PDFs, scanned PDFs, images, audio, and video recordings.

Document a few publicly available dataset links that you could use for training the chatbot, and a brief description of the data these datasets hold.

3. Examples of Tools for Data Collection

You will focus on simple text data from websites and a few common data file formats, such as CSV and PDF, for now. However, you must recognize that real-time data from chat clients and extracted data from instructional videos will become very important in the final project. Therefore, it is recommended that you start researching extraction tools for those sources.

There are many options for searching, finding, and collecting data. You are expected to individually collect a few data samples related to your team's dataset domain. *(This is a part of data exploration and does not have to be the final datasets you would be working on*

Open a terminal or command prompt and install the necessary libraries:

```
pip install requests pandas beautifulsoup4 pdfplumber
pip install pytesseract
```

The following are some examples of the libraries and tools for data collection:

- requests: https://pypi.org/project/requests/
- pandas: https://pandas.pydata.org/
- beautifulsoup4: https://pypi.org/project/beautifulsoup4/
- pdfplumber: https://pypi.org/project/pdfplumber/
- pytesseract: https://pypi.org/project/pytesseract/

4. Data Collection

For this assignment, you will retrieve three different types of data:

i. CSV or Excel
ii. ASCII Texts like Forum Postings and HTML
iii. PDF and Word Documents that require conversion and OCR

Choose any publicly available dataset from the data sources and types that you have chosen.

Create a Python file named "data_exploration.py" to retrieve the dataset using its API and store it in CSV / Excel format.

Run basic operations on the dataset, including displaying the first few records, calculating the dataset's size and dimensions, identifying missing data, etc.

For websites, extract the text using web scraping libraries.

For PDF documents, extract text using any PDF-to-text libraries.

*[**Note:** These tasks are just to help you understand how the data exploration needs to be done, so do not worry about which source and libraries/tools you pick. Your focus should be on understanding how the libraries work and how data can be extracted from these sources.]*

5. Submission

Individually submit a document that lists the team details, your shortlisted set of domains, publicly available datasets for each of them, and the reasoning behind your choices. Provide a good reason behind your topic choice. Make a list of data sources, the links, and brief descriptions with a sample excerpt of your data for each source.

Submit the data exploration Python file along with the document. In the report, describe what the script does (conversion tasks and tools to keep only the relevant data) to create a clean single dataset.

While there are many attempts to build realistic chatbots, most people would rather speak to a real person because chatbots' capabilities are very limited. Describe what might be missing in these existing chatbots. Discuss how your dataset might improve the overall performance and correctness.

Please submit all documents, answers to the questions, source codes, and reports on Coursistant (https://usc.xlearnedu.com) by the due date and time. Provide a document in ***PDF format (No other format would be considered)***. Please mention your ***Name and USC ID*** at the end of the document.

Please create a demo video showing how your scripts can convert various data sources into a common format. Upload the demo video to YouTube and submit the link. The main purpose of the video is to convince me that you did the tasks.

There will be a 50% penalty for all late submissions.