

DSCI-560 Lab 2

NOTE: You must use GitHub to store all source codes and reports associated with the lab assignments. You must commit/push all of the modifications for the assignments with detailed descriptions at least once daily when working on the lab assignment. Please submit the GitHub link and record of your commits and detailed commit description to Brightspace before the due date.

1. Team Formation

All the assignments should be performed in teams of 3. Create a team of 3, and decide a name for your team. **Document the team name and the name and USC ID of all three members.** Each member must have the same details in their assignment submission for this week.

Having the 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 that your team gets maximum time to work on the assignments.

2. Data Selection, Search, Find, and Collect

For this assignment, you will start to think about how you might approach the final project around building a Teaching Assistant assisting chatbot. 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 data sets hold.

3. Examples of Tools for Data Collection

You will concentrate on simple text data from websites and a few common data document types 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 looking into the extraction tools for those sources.

There are many options to search, find, and collect 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 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 some basic operations on the dataset including displaying the first few records, calculating the size and dimensions of the dataset, identifying missing data, etc.

For websites, extract the text using web scraping libraries.

For PDF documents, extract text using the 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 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 a lot of attempts to build realistic chatbots, most people would rather speak to a real person because their 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 Blackboard 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 to show how your scripts can be used to convert various types of data sources into common data in the same 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.