DASC 623: Data Mining and Text Analysis

Assignment #1: Extracting Text from Documents

Fall Quarter 2020

Overview

In this assignment you will create the first step in your NLP data analysis pipeline - ingesting the text data from the file(s) in which the data are contained. Text data can be stored in many different formats, thus you are tasked with defining a set of Python functions that can be used to extract the text from each file type. You should put some time into thinking about how you should treat the data extracted from the different types of files. For example, you should think about in what type(s) of objects should the extracted text be stored?

Part A

This part contains several exercises. For each exercise below you are to define a Python function that can be used to extract text from the file type specified. The first argument of each of these functions should be the path to the file containing the text. Your functions should also include additional arguments that are appropriate for the way that the text is extracted and processed.

Problem #1 Extract text from a .pdf file

Problem #2 Extract text from a .doc file

Problem #3 Extract text from a .docx file

Problem #4 Extract text from a .csv file

Problem #5 Extract text from a .xls file

Problem #6 Extract text from a .xlsx file

Problem #7 Extract text from a .txt file

Problem #8 Extract text from an image file (.pdf, .png,.jpeg)

Part B

This part contains a single exercise. For this exercise you are define a Python function that takes as its first argument an `array` or `list` of filepaths for files that contains text data. The function should identify the file type for each entry in the path argument and based on that determination select the appropriate function that you defined in Part A to extract the text.