CSCE 580: Intr AI   Prof. Biplav Srivastava, Fall 2025

Quiz 1 / September 9, 2025 – Tuesday / Instructions

- All the responses should be in your Github before **the end of day on Tuesday (Sep 16, 2025)** – next Tuesday.
- For coding part (Q3 and Q4), implement python notebooks or in Collab. Call them " –Quiz1-Response-Q3 or Q4". If Github, put it in your Github repo under "Quiz1" sub-folder. All files, including doc, data and code, will be under this. Examples: "<>/Quiz1/Responses.pdf, "<>/Quiz1/Q3-code.ipynb".
- For questions/ clarifications, send an email to Instructor biplav.s@sc.edu and TAs vishalp@email.sc.edu, kausik@email.sc.edu.

Total points = (20 + 25 + 55): 100 points, Obtained =

Student Name:

The quiz is to test your understanding of concepts of intelligent agents and practical problem solving.

Q1: About data for AI [4 + 16 = 20 points]
Instructions: Give your answers in bullet points.

a) What is open data? Given an example of open data that you produce which others can use? [2 + 2 = 4]
   * Data that is available to be modified or used by anyone freely. Lab reports are often published for open access to be repeated and verified, and to provide data on a subject.
b) You are analyzing a dataset and some attributes are missing.

b.1) What could be any 2 reasons why they are missing? [2 + 2 = 4]

  * The value could have not been provided
  * The value is being redacted.

b.2) What are any 2 ways you can still proceed with data analysis despite the missing values. For each, mention what assumption you are making and what are its risks. [(2+2+2) * 2 = 12]
* You could omit the value, but this will reduce the available data.
* You could impute a new value (through median or mean), but this assumes no outliers and average cases.

Q2: Programing activity: resume analysis [25 points]

We will work with crowdsourced resume data of students from the class. They are at:
https://drive.google.com/drive/folders/1F6HRaliFWcakVvT605m8Js6a1D40Tx24?usp=sharing

This analysis has to be done as a python notebook or collab. It should be saved as "<>/Quiz1/Q2-code.ipynb".

- Task 1 [10 points]
    - Do: Read your resume in text and get a list of words. Let us call them **resume_words**
    a. Do: Plot a histogram of top 20 resume_words, i.e., bar graph of words (x-axis) and counts (y-axis).

b. Context: The list of common English words are called **stop_words**. They are usually articles, determiners and prepositions, along with their variations. A list of 127 are at: https://gist.github.com/sebleier/554280 (Raw file is at: https://gist.githubusercontent.com/sebleier/554280/raw/7e0e4a1ce04c2bb7bd41089c9821dbcf6d0c786c/NLTK's%2520list%2520of%2520english%2520stopwords)
Do: Remove stop_words from resume_words. Let us call them **specific_words**.
Plot the histogram for **specific_words.**
c. Analyze: Note which words emerge now. Was removing stop_words helpful in revealing more about you (from the resume).

- Task 2 [10 points]
  - Context: Take all resumes in folder
  1. Do: Read all resume in text and get a list of words. Let us call them **resume_words**
  2. Do: Plot a histogram of top 20 resume_words, i.e., bar graph of words (x-axis) and counts (y-axis).
  3. Do: Remove stop_words from resume_words. Now plot the histogram for **specific_words**
  4. Analyze: Note which words emerge now. Was removing stop_words helpful in revealing more about the class (from the resumes).
- Task 3: [5 points]
  1. Analyze: specific_words from your resume and that of class. Which words are unique to you?

Q3: Programming activity: data analysis for social impact [55 points]

We are provide you access to redacted version of real data about firefighting at a firestation's services in the Midlands in 2025. There are omitted fields to maintain confidentiality (addresses, names). The data has 8 columns and 2,200 rows.

See: https://drive.google.com/drive/folders/1nJTZJZ_M9e7whJy4cMzNCXTXfN7zYvPs?usp=sharing

Write python code/ demonstrate its working in notebook, and report on the following questions along with your code.

a) Data issues: [15 points]

1. What is the range of data for the cases (dispatches) ? [2 points]
2. What % of data is missing, by each column? [3 points]
3. What data issues are there (e.g., different formats) and how we can resolve them [5 points]
4. Resolve data issues. Assign IDs. Pick a method for handling missing data and use consistently. Describe your data cleaning strategy, as appropriate. Do remainder of the tasks with data resolved. [5 points]

b) Exploratory data analysis – about file alarms. Answer from your analysis. [20 points]

1. On an average, in how much time is a call (alarm) resolved from the time it is created to closed ? [5 points]
2. How many fire units, on an average, are usually sent for a fire alarm? [5 points]
3. Which shift is the busiest among A, B, C ? [5 points]
4. Create a matrix of number of file alarms organized by the day of week (x_axis) and hour of the day (y-axis). It will also have totals for each row and column. See illustration below. [5 points]

HOUR	SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY	TOTAL
0:00	8	2	2	1	0	6	1	20
1:00	3	1	3	4	2	2	3	18
2:00	2	4	3	3	3	0	2	17
3:00	2	1	2	3	2	0	1	11
4:00	3	2	1	2	1	1	2	12
5:00	4	2	2	1	1	3	3	16
6:00	2	4	4	4	4	0	3	21
7:00	0	7	6	0	3	3	3	22
8:00	2	6	7	5	5	9	3	37
9:00	5	5	4	5	2	6	4	31
10:00	7	6	6	10	5	8	4	46
11:00	13	4	6	5	7	12	7	54
12:00	9	6	8	6	4	8	10	51
13:00	5	7	7	5	5	5	4	38
14:00	7	3	13	8	14	9	8	62
15:00	6	4	6	6	7	7	10	46
16:00	5	8	8	5	6	5	9	46
17:00	5	14	6	4	7	9	6	51
18:00	3	8	9	7	2	7	8	44
19:00	2	7	7	6	1	4	5	32
20:00	5	3	10	3	6	5	6	38
21:00	7	1	5	6	4	6	5	34
22:00	2	4	3	3	3	3	1	19
23:00	2	2	4	0	2	1	0	11
Count of Inic	109	111	132	102	96	119	108	777

c) Unsupervised learning [20 points]

1. Cluster the data based on any two methods in sci-kit and report on their cluster quality. Which method performs better ? [15 points]
2. Using the best result, try to interpret (label) the clusters. What do they represent? [5 points]