CSCE 580: Intr AI
Quiz 1 / September 9, 2025 – Tuesday / Instructions

Prof. Biplav Srivastava, Fall 2025

- 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: JAMES TABAKIAN

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]
Open data is data that is accessible for everyone to use. It is useful for training artificial intelligence models because the computer scientists don't have to manually collect the data themselves.

Filling out the census would be an example of contributing to open data. The results get published by the government which anyone could use.

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]

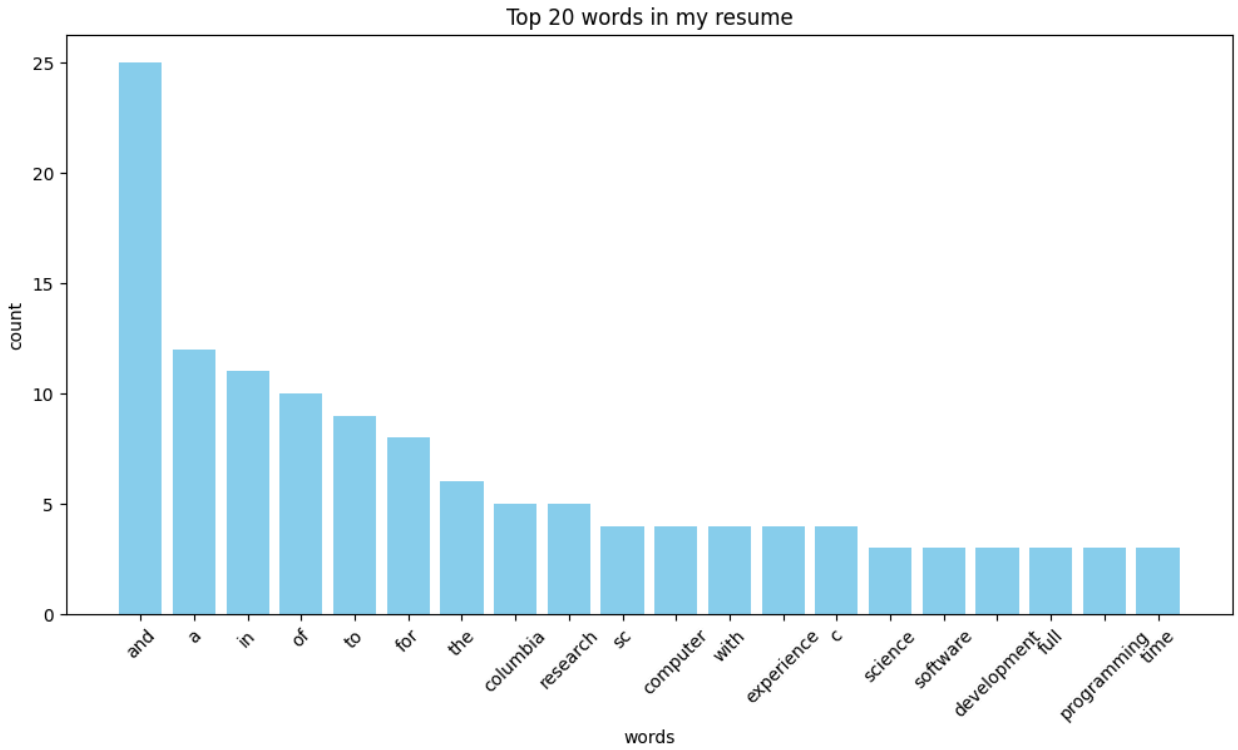
1) The data was never collected.
2) The data was lost or got corrupted

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]

1) Remove row/col with missing data
   Assumption: The missing data occurs randomly
   Risk: Deleting data reduces the sample size therefor reducing statistical power
2) Replace missing data with median, mean, or mode
   Assumption: Missing values are close to the overall median
   Risk: Bias can be created if the data is not missing at random

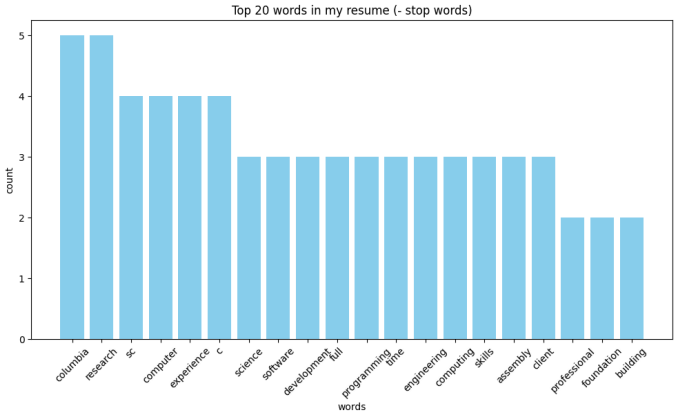
Q2: Programing activity: resume analysis [25 points]
We will work with crowdsourced resume data of students from the class. They are at:
https://drive.google.com/drive/folders/1F6HRaliFWcakVvT605m8Js6a1D40Tx24?usp=sharing
This analysis has to be done as a python notebook or collab. It should be saved 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).


Top 20 words in my resume
count
words
0
5
10
15
20
25
and
a
in
of
to
for
the
columbia
research
sc
computer
with
experience
c
science
software
development
full
programming
time

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)

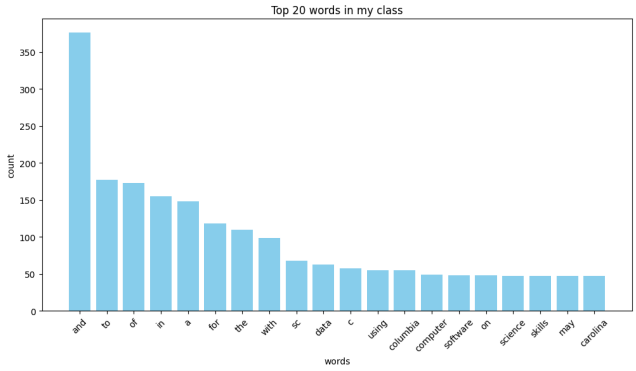
Do: Remove stop_words from resume_words. Let us call them **specific_words**.
Plot the histogram for **specific_words.**



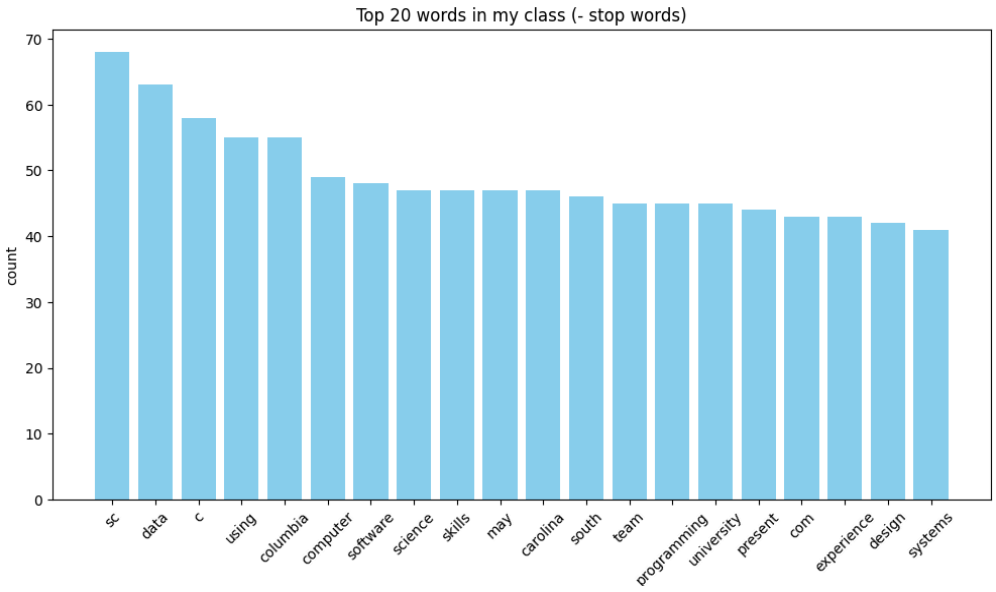
c. Analyze: Note which words emerge now. Was removing stop_words helpful in revealing more about you (from the resume).

Removing stop words was helpful because the new list of 20 words shows off more of my skills and background instead of including common words that are used in the majority of sentences.

- 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).

Removing stop words was helpful because now it is easy to see on the list what skills and background the students come from in my class instead of half of the words being everyday common words.

- Task 3: [5 points]
    1. Analyze: specific_words from your resume and that of class. Which words are unique to you?

I own a computer-building business so words like “client”, “building”, and “assembly” are common on my resume while they do not appear in the top 20 words for the class.

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]

The data contains ID numbers, (Dates and time), shifts, and dispatch units

2. What % of data is missing, by each column? [3 points]

```
XREF ID                    0.000000%
DISPATCH UNIT              0.000000%
DISPATCH CREATED DATE      0.000000%
INCIDENT NUMBER            0.000000%
1ST UNIT ON SCENE         19.454545%
ALARM DATE TIME            1.409091%
CALL COMPLETE              1.409091%
SHIFT                      3.136364%
```

3. What data issues are there (e.g., different formats) and how we can resolve them [5 points]

The first unit on scene is not recorded a lot of the time and the alarm date time, call complete time, and shift are occasionally not reported. We can resolve this by either dropping entries with missing data, or by replacing missing data with its mean, median, or mode.

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]

87 Days, 6 Hours, 43 Minutes

2. How many fire units, on an average, are usually sent for a fire alarm? [5 points]

1.48 Fire Units

3. Which shift is the busiest among A, B, C ? [5 points]

Shift A is the busiest with 590 Alarms

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]

DayOfWeek	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Total
Hour								
0	7	6	4	1	2	7	4	31
1	8	7	7	4	3	5	7	41
2	4	3	2	2	3	7	7	28
3	6	8	9	1	10	4	4	42
4	2	4	5	2	6	4	4	27
5	8	3	6	5	3	2	4	31
6	5	5	6	6	8	6	9	45
7	13	13	6	7	11	4	7	61
8	11	9	13	7	6	4	11	61
9	8	14	11	12	12	7	11	75
10	16	12	9	10	11	13	11	82
11	15	13	12	17	16	11	10	94
12	15	12	11	17	19	9	17	100
13	16	11	10	9	12	14	18	90
14	15	19	11	15	15	14	13	102
15	13	19	23	22	15	13	14	119
16	17	18	10	15	16	11	9	96
17	19	11	17	8	23	14	13	105
18	16	16	22	20	8	13	8	103
19	16	8	15	12	12	15	7	85
20	15	12	21	14	12	18	9	101
21	11	13	8	10	12	11	7	72
22	4	11	15	8	12	16	12	78
23	7	6	5	6	7	15	5	51
Total	267	253	258	230	254	237	221	1720

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]

I used a k-means silhouette to get a score of .4118 and dbscan to get a score of .2445. They both performed the strongest at two clusters. The k-means method performed better because its score was closest to 1.

2. Using the best result, try to interpret (label) the clusters. What do they represent? [5 points]

The first cluster occurs between the hours 9:00 and 22:00 when the station is the busiest. The second cluster occurs between 23:00 and 8:00 when the station is not busy. It was busy during the first cluster because that's when the majority of the population is awake and the opposite is true for the second cluster.