

DSCI 560 LAB 02

TEAM DETAILS

Members:

Abhinav Parameswaran - 6936475080

Pavithra Kollipara - 7183732161

Rutuja Bhandigani - 5853375781

Shortlisted Domains:

- **Machine Learning** : Machine Learning is one of the hot-topics trending in tech in today's world, and a lot of students are eager to learn about Machine Learning from top-notch professors at USC. Consequently, there is a substantial surge in student enrollment for the course. Managing such a large class becomes a challenging task due to the high demand. That's why we have specifically focused on this domain to ensure that a maximum number of students can have their questions addressed promptly without having to wait for a long time for the Professor or Teaching Assistant's response.
- **Statistical Programming** : A specialized chatbot in Statistical Programming provides instant expert assistance, on-demand learning, and efficient problem-solving for users working with languages like R or Python. It serves as a quick reference for syntax, fosters community engagement, and keeps users updated on the latest tools and libraries. The chatbot streamlines data exploration, automated report generation, and integrates seamlessly with development environments, enhancing the overall efficiency and learning experience in statistical programming.

List of data sources, links, and descriptions with a sample excerpt of data for each source.

1. UCI Student Performance (CSV)

<https://archive.ics.uci.edu/dataset/320/student+performance>

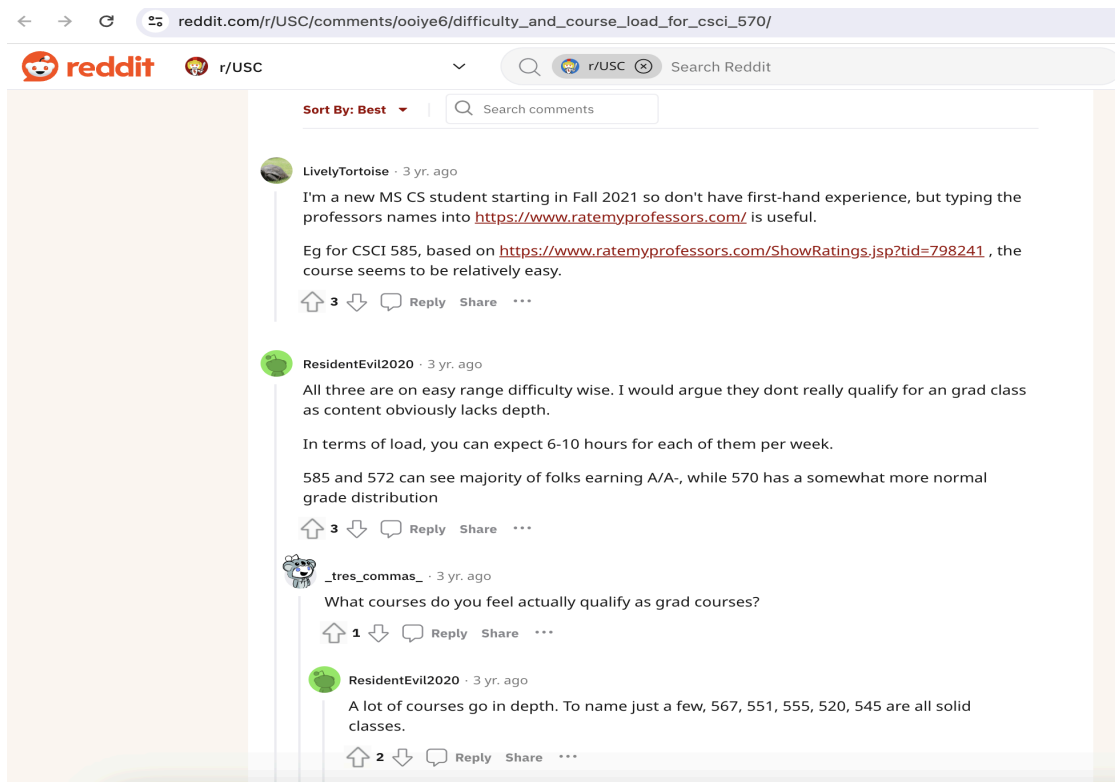
This dataset contains information about students, including student details, study habits, and exam performance. It can be useful for a Teaching Assistant chatbot to understand student performance trends and provide assistance to students with poor performance

3. Reddit Threads (Forum)

https://www.reddit.com/r/USC/comments/ooiye6/difficulty_and_course_load_for_csci_570/

https://www.reddit.com/r/USC/comments/12pznyo/csci_570_shamsian_summer/

If the chatbot serves as a Teaching Assistant (TA), training it on Reddit data can enhance its ability to assist students by providing diverse perspectives, real-world problem-solving scenarios, and up-to-date information within the relevant academic domain. It helps the chatbot understand student queries in varied language and offer contextually appropriate responses based on the wealth of knowledge available on Reddit forums.



4. General Conversational Chatbot Dataset (JSON)

<https://www.kaggle.com/datasets/niraliivaghani/chatbot-dataset?resource=download>



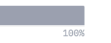


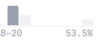



```
intents 3.json
JSON

323 "responses": [
324   "My College has total 2 floors "
325 ],
326 "context_set": ""
327 },
328 {
329   "tag": "syllabus",
330   "patterns": [
331     "Syllabus for IT",
332     "what is the Information Technology syllabus",
333     "syllabus",
334     "timetable",
335     "what is IT syllabus",
336     "syllabus",
337     "What is next lecture"
338   ],
339   "responses": [
340     "Timetable provide direct to the students OR To know about syllabus visit <a target=\"_blank\" href=\"TIMETABLE LINK\"> here</a>"
341   ],
342   "context_set": ""
343 },
344 {
345   "tag": "library",
346   "patterns": [
347     "is there any library",
348     "library facility",
349     "library facilities",
350     "do you have library",
351     "does the college have library facility",
352     "college library",
353     "where can i get books",
354     "book facility",
355     "Where is library",
356     "Library",
357     "Library information",
358     "Library books information",
359     "Tell me about library",
360     "how many libraries"
361   ],
362   "responses": [
```

5. Transcript PDF (PDF)

<https://huggingface.co/datasets/jamescalam/youtube-transcriptions?row=0>

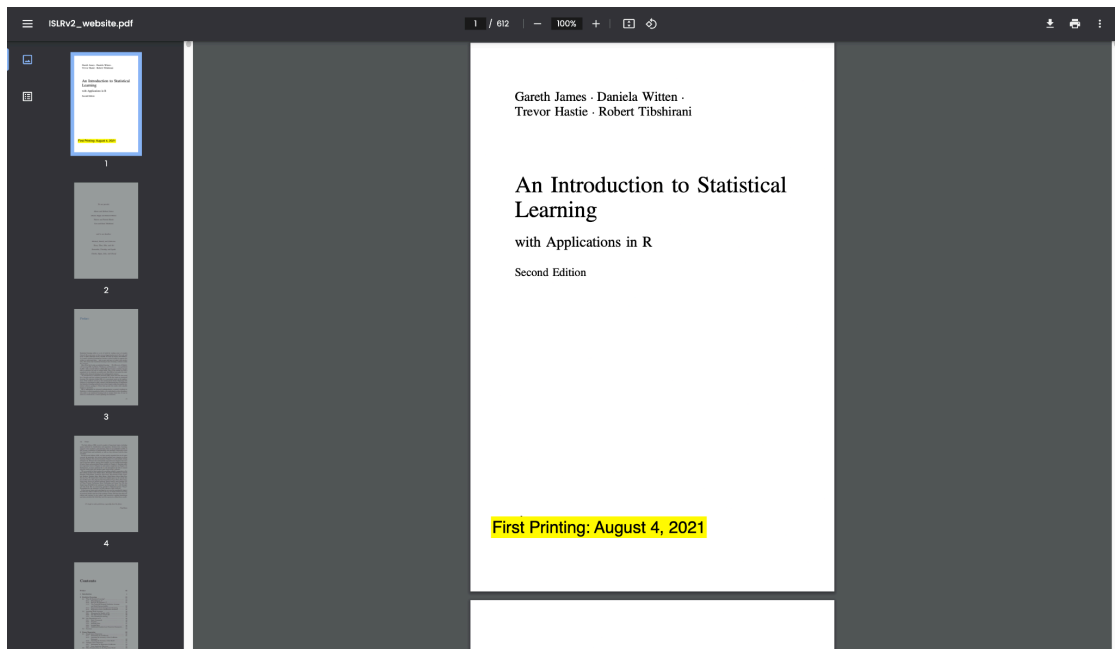
Video transcripts as a dataset is beneficial for training a chatbot specific to that domain. It allows the chatbot to learn from the spoken content, improving its understanding and ability to generate contextually relevant responses within the video domain. In the case of the above linked dataset, the video describes the steps involved in training and testing an Italian BERT (transformer).

title string · lengths	published string · lengths	url string · lengths	video_id string · lengths	channel_id string · classes	id string · lengths	text string · lengths	start float64	end float64
 67-78 14.5%	 23 13%	 28 100%	 11 100%	 UCv83t05ce... 13%	 18-20 53.5%	 0-60 29.8%	 0-3.78k 93.1%	 2.18-3.78k 93.1%
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t0.0	Hi, welcome to the video.	0	9.36
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t3.0	So this is the fourth video in a...	3	11.56
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t9.36	from Scratch mini series.	9.36	15.84
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t11.56	So if you haven't been following along.	11.56	18.48
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t15.84	we've essentially covered what you can...	15.84	20.6
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t18.48	So we got some data.	18.48	23.72
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t20.6	We built a tokenizer with it.	20.6	25.76
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t23.72	And then we've set up our input pipeline	23.72	28.48
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t25.76	ready to begin actually training ou...	25.76	32.36
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t28.48	is what we're going to cover in this...	28.48	35.96
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t32.36	So let's move over to the code.	32.36	39.56
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t35.96	And we see here that we have essentially...	35.96	40.48
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t39.56	we've done so far.	39.56	48.8
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t40.4800000000000004	So we've built our input data, our input...	40.48	51.52
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t48.8	And we're now at a point where we have ...	48.8	54.04
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t51.51999999999999996	PyTorch data loader, ready.	51.52	56.4
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t54.04000000000000006	And we can begin training a model wit...	54.04	61.84
Training and Testing an Italian BERT -...	2021-07-06 13:00:03 UTC	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-t56.4	So there are a few things to be aware...	56.4	64.88
Training and Testing	2021-07-06	https://youtu.be/3SPdoyi6ZoQ	3SPdoyi6ZoQ	UCv83t05cePwHMT1952IVVhw	3SPdoyi6ZoQ-	So I mean, first,	64.88	67.98

6. Textbook (PDF)

https://hastie.su.domains/ISLRv2_website.pdf

Utilizing a textbook PDF as a dataset for training the chatbot can enhance its knowledge in the specific domain covered by the textbook. By incorporating the textual content, the chatbot can learn from the structured information within the textbook, broadening its understanding and enabling it to provide more accurate and informed responses related to the domain of interest.



Functionalities of the *data_exploration.py* file

subreddit_scrapper(): In the *data_exploration* python file, the function *subreddit_scrapper* does not use the reddit api as such for simplicity of code for now. However, in our final project, we propose to use the API for more accurate and efficient data retrieval.

csv_data_operations(): This function reads a csv file into a dataframe using pandas, displays the first few records of the dataset, calculates the size and dimension of the dataset, identifies missing data in the dataset and provides some basic statistics about the dataset.

scrape_pdf(): This function is used for extracting text from a given pdf file. It uses the *pdfplumber* library to open the pdf file, iterates through each page of the pdf, extracts text content and prints the 1st page on terminal.

Output (screenshots)

```

beckendrof@beckendrof-virtual-machine:~/Desktop/abhinav_6936475080/scripts$ python3 data_exploration.py
/home/beckendrof/Desktop/abhinav_6936475080/scripts/data_exploration.py:5: DeprecationWarning:
Pyarrow will become a required dependency of pandas in the next major release of pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and better interoperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54466

import pandas as pd
1. CSV data
1. ASCII data from QnA forum
3. PDF - text data

Enter your choice
1

Retrieving data from csv file...

  school sex  age address famsize Pstatus  Medu  Fedu  Mjob  Fjob  reason guardian  ...  internet  romantic  famrel  freetime  goout  Dalc  Walc  health  absences  G1  G2  G3
0    GP   F   18     U    GT3      A      4      4  at_home  teacher  course  mother  ...    no      no      4      3      4      1      1      3      6      5      6      6
1    GP   F   17     U    GT3      T      1      1  at_home  other  course  father  ...    yes     no      5      3      3      1      1      3      4      5      5      6
2    GP   F   15     U   LE3      T      1      1  at_home  other  other  mother  ...    yes     no      4      3      2      2      3      3     10      7      8     10
3    GP   F   15     U    GT3      T      4      2  health  services  home  mother  ...    yes     yes     3      2      2      1      1      5      2     15     14     15
4    GP   F   16     U    GT3      T      3      3   other  other  home  father  ...    no      no      4      3      2      1      2      5      4      6     10     10

[5 rows x 33 columns]
Dimensions of the dataset: (395, 33)
school      0
sex          0
age         0
address     0
famsize     0
Pstatus     0
Medu        0
Fedu        0
Mjob        0
Fjob        0
reason      0
guardian    0
traveltime  0
studytime   0
failures    0
schoolsup   0
famsup      0
paid        0
activities  0
nursery     0
higher      0
internet    0
romantic    0
famrel      0
freetime    0
goout       0
Dalc        0
Walc        0
health      0
absences    0
G1          0
G2          0
G3          0
dtype: int64

   age      Medu      Fedu  traveltime  studytime  failures  famrel  ...  Dalc      Walc      health  absences      G1      G2      G3
count 395.000000 395.000000 395.000000 395.000000 395.000000 395.000000 395.000000 ... 395.000000 395.000000 395.000000 395.000000 395.000000 395.000000 395.000000
mean  16.696203  2.749367  2.521519  1.448101  2.035443  0.334177  3.944304 ...  1.481013  2.291139  3.554430  5.708861  10.908861  10.713924  10.415190
std    1.276043  1.094735  1.088201  0.697505  0.839240  0.743651  0.896659 ...  0.890741  1.287897  1.390303  8.003096  3.319195  3.761505  4.581443
min    15.000000  0.000000  0.000000  1.000000  1.000000  0.000000  1.000000 ...  1.000000  1.000000  1.000000  0.000000  3.000000  0.000000  0.000000
25%    16.000000  2.000000  2.000000  1.000000  1.000000  0.000000  4.000000 ...  1.000000  1.000000  3.000000  0.000000  8.000000  9.000000  8.000000
50%    17.000000  3.000000  2.000000  1.000000  2.000000  0.000000  4.000000 ...  1.000000  2.000000  4.000000  4.000000  11.000000  11.000000  11.000000
75%    18.000000  4.000000  3.000000  2.000000  2.000000  0.000000  5.000000 ...  2.000000  3.000000  5.000000  8.000000  13.000000  13.000000  14.000000
max    22.000000  4.000000  4.000000  4.000000  4.000000  3.000000  5.000000 ...  5.000000  5.000000  5.000000  75.000000  19.000000  19.000000  20.000000

[8 rows x 16 columns]

```

```

* beckendrof@beckendrof-virtual-machine:~/Desktop/abhinav_6936475080/scripts$ python3 data_exploration.py
/home/beckendrof/Desktop/abhinav_6936475080/scripts/data_exploration.py:5: DeprecationWarning:
Pyarrow will become a required dependency of pandas in the next major release of pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and better interoperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54466

import pandas as pd
1. CSV data
1. ASCII data from QnA forum
3. PDF - text data

Enter your choice
2

Scrapping CSCI 570 related posts on r/USC...
Saving relevent post data to csv...
CSV file created

```

```

* beckendrof@beckendrof-virtual-machine:~/Desktop/abhinav_6936475080/scripts$ python3 data_exploration.py
/home/beckendrof/Desktop/abhinav_6936475080/scripts/data_exploration.py:5: DeprecationWarning:
Pyarrow will become a required dependency of pandas in the next major release of pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and better interoperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54466

import pandas as pd
1. CSV data
1. ASCII data from QnA forum
3. PDF - text data

Enter your choice
3

Retrieving data from pdf file....
<Page:1>

DSCI 552: Machine Learning for Data
Science (Spring 2024)
Units: 4
Instructor: Mohammad Reza Rajati, PhD
PHE 412
rajati@usc.edu – Include DSCI 552 in subject.
Office Hours: Right after the lecture, by appointment
Webpage: Personal Homepage at Intelligent Decision Analysis
TA(s): Will be introduced on Piazza.
Lecture 1: Tuesday, Thursday, 10:00 pm –11:50 am GFS 116
Lecture 2: Tuesday, Thursday, 4:00 pm –5:50 pm OHE 122 & Online
Webpages: Piazza Class Page for discussions, announcements, and course materials
and USC DEN Class Page for exams and grades
and GitHub for code submission
– All HWs, handouts, solutions will be posted in PDF format
– Student has the responsibility to stay current with webpage material
Prerequisite: Prior courses in multivariate calculus, linear algebra, probability, and statistics.
– This course is a prerequisite to DSCI 558.
Other Requirements: Computer programming skills.
Using Python is mandatory.
Students must know Python or must be willing to learn it.
Tentative Grading: Assignments 45%
Midterm 1 20%
Midterm 2 25%
Final Project 10%
Participation on Piazza* 5%
Letter Grade Distribution:
≥ 93.00 A 73.00 - 76.99 C
90.00 - 92.99 A- 70.00 - 72.99 C-
87.00 - 89.99 B+ 67.00 - 69.99 D+
83.00 - 86.99 B 63.00 - 66.99 D
80.00 - 82.99 B- 60.00 - 62.99 D-
77.00 - 79.99 C+ ≤ 59.99 F

```



```

scripts > reddit.csv > data
1 Title,Comment
2 Three weeks absence from class in the first month of college,"I'm in the MSCS program and started this fall, just double
3 What are the best resources for CSCI 570 Algorithms,This.. and around 300 leetcode problems during the semester. So
4 SGM 124 is unstable,Even the ceiling can't handle csci-570
5 Shamsian for CSCI 170? Is he good?,"Shamsian is the best, I took his CSCI 570 Algo course for grads, and by far the best
6 Grade requirement for Masters,"Which course is this, if CSCI 570, let me know"
7 Masters Courses Graded on a Curve?,Csci 570
8 CSCI : 570,There is only one class for which it is a pre req CSCI 670. Which is a PhD level course. So if you are not a
9 F grade in a course in grad school - will they be kicked out?,sounds like CSCI 570 to me
10 How hard is the M.S. Computer Science?,"One class per semester should be very manageable, even if you work full time. Yo
11 Dream grad school...not so much anymore.,Don't let one person ruin your experience. When I took CSCI 570 people were sup
12 "Difficulty and Course load for CSCI 570: Algorithms, CSCI 572: Information retrieval, CSCI: 585: Database systems.",I d
13 Taking 570 over the Summer,"If you're a student with disabilities, I strongly recommend that you do not take CSCI 570. T
14 CSCI 570 (Shamsian) Summer,"I took the undergrad version of 570 (CSCI 270) at USC, and 570 was a literal mirror of the c
15 MS Computer Science Course Recommendations,Sorry! CSCI 402 EE 457 EE 450 CSCI 570 576 577a 577b 580 561 571 585 590
16 "Difficulty and Course load for CSCI 570: Algorithms, CSCI 572: Information retrieval, CSCI: 585: Database systems.",On
17 Dream grad school...not so much anymore.,USC's grad program is definitely a bit differently focused from undergrad but
18 Dream grad school...not so much anymore.,CSCI 570. A required course for weeding out some CS students
19 Professors at USC,"CSCI-570 was a death trap for me because I wasn't a CS major, and yeah the class was rough but he was
20 Computer Science for Scientists and Engineers Master Program at USC: Is the program difficult for someone who has no kno
21 Famous courses to take at USC,CSCI 570 if you're passionate about crying in your free time
22 Help Deciding Between USC and UCLA for CS Masters,Have you seen the size of our CS master's courses? The program is stra
23 Life-changing classes?,CSCI 402 by Bill Cheng and CSCI 570 by Shamsian. Absolutely breathtaking!

```

Describe your vision of the final system that people would care about.

The envisioned domain-specific chatbot for Teaching Assistants will serve as an intelligent assistant capable of addressing a wide range of questions on platforms like Piazza or discussion forums related to a specific subject. While resembling ChatGPT, it specifically targets the educational domain, aiming to provide accurate and helpful responses to student queries, thereby enhancing the learning experience and facilitating efficient communication on academic platforms.

Describe what might be missing in these existing chatbots.

Existing chatbots may have limitations or areas where improvements could be made.

Limited Context Awareness: Chatbots might struggle to understand the context of a conversation over multiple turns, leading to responses that lack coherence or relevance.

Natural Language Understanding: Chatbots may face difficulties in handling ambiguous queries or understanding nuanced language, leading to misinterpretations.

Lack of Semantic Understanding: Understanding the meaning behind words and sentences might be challenging, resulting in incorrect or irrelevant responses.

Handling Complex Queries: Chatbots might struggle to handle complex queries that require a deep understanding of a specific domain or intricate information.

Limited Personalization: Lack of effective personalization may lead to generic responses, overlooking the user's preferences and history.

Forgetting User History: Some chatbots may not effectively retain or utilize user history, impacting the ability to provide contextually relevant responses.

Limited Learning Capabilities: Chatbots may lack effective mechanisms for continuous learning from user interactions, hindering their ability to improve over time.

Discuss how your dataset might improve the overall performance and correctness.

Utilizing datasets such as UCI Student Performance (CSV) enhances the Teaching Assistant chatbot's ability to offer personalized assistance by understanding individual student profiles. The inclusion of Syllabus PDF details improves the chatbot's accuracy in responding to student inquiries about course structures and policies. Training on Reddit Threads broadens the chatbot's knowledge, incorporating diverse perspectives and real-world scenarios. Integrating a General Conversational Chatbot Dataset enhances the chatbot's natural language understanding, while Syllabus PDF and Textbook PDF datasets enrich its context-awareness in the machine learning domain, and also provides data about how the course is structured. This comprehensive approach enhances the overall performance and correctness of the Teaching Assistant chatbot.

Video Demo Link: https://www.youtube.com/watch?v=8AjY61_6QR0