Project Phase I: Read out loud

Instructions:

- The aim of this project is to give you hands-on experience with a real-life machine learning application.
- There are two phases of this project. In phase I, you are required to do data collection and implement a basic logistic regression model, while in phase II you will use the collected data to build language recognition models (more details on phase II will be shared later).
- This is a group project and the maximum number of members per group is 6. **However, please clearly note that each member of the group is required to complete and submit phase I individually.**
- Phase I will serve as a backbone for phase II. Any irregularities here will not only affect your project but also your group's project. In case of any ambiguity ask the TAs first.
- **Carefully read the submission instructions at the end of the project. Late days policy is not applicable to project phase I.**
- The deadline to submit phase I is **Friday, 26th November 2021.**

Problem:

You are required to record the quick answers to the following questions in the language specified. Please carefully read the recording guidelines below before starting recording.

Urdu:

1. ایک ایسی جگہ کے بارے میں بتائیں جہاں آپ چھٹیاں گزارنا چاہیں گے
2. لمز یونیورسٹی میں آپ کا سب سے دلچسپ تجربہ کیا رہا
3. کورونا وائرس کی وجہ سے ہونے والے لوک ڈاؤن میں آپ کو سب سے زیادہ مشکل کیا درپیش رہی
4. اپنے مشین لرنگ کے کورس کے اب تک کے تجربہ کے بارے میں کچھ بتائیں
5. آپ مشین لرنگ کے کورس میں کیا بہتری دیکھنا چاہیں گے
6. اپنے پسندیدہ سائنس فکشن کریکٹر کے بارے میں کچھ بتائیں
7. وہ کیا وجوہات ہیں جن کی بنا پر آپ بیچلرز کے بعد مزید تعلیم حاصل کرنا چاہیں یا نہ چاہیں گے
8. اپنی پسندیدہ کتاب کے بارے میں کچھ بتائیں
9. ڈرامے یا فلم کے بارے میں کچھ بتائیں پسندیدہ اپنے
10. کیا آپکو کائینات میں ایلینز کی موجودگی کا خیال اچھا لگتا ہے

11. ورلڈ کپ میں پاکستان کرکٹ ٹیم کی کارکردگی کے بارے میں آپ کے کیا خیالات ہیں
12. اگر آپ تاریخ کے ایک شخص کے ساتھ رات کا کھانا کھا سکتے ہیں تو یہ کون ہوگا اور کیوں
13. لمزمیں اپنی پسندیدہ جگہ کے بارے میں ہمیں بتائیں
14. ۔اور دلچسپ پایا ہمیں کسی بھی کورس کے بارے میں بتائیں جو آپ نے لمز میں پڑھا
15. آپ کس شہر میں رہنا پسند کریں گے اور کیوں
16. اپنے بچپن کی پسندیدہ یاد کے بارے میں ہمیں کچھ بتائیں
17. اپنا ویکنڈ گزارنے کے پسندیدہ طریقے کے بارے میں کچھ بتائیں
18. آپ کس قسم کی فلمیں یا ٹی وی سیریز دیکھنا پسند کرتے ہیں؟
19. کیا آپ گھر پر فلمیں دیکھنا پسند کرتے ہیں یا تھیٹر میں اور کیوں؟
20. کیا آپ فلموں کی بجائے ٹی وی سیریز دیکھنا پسند کرتے ہیں اور کیوں؟
21. اگر آپ ساری زندگی صرف ایک کھانا کھا سکتے تو وہ کیا ہوگا؟
22. اگر آپ وقت پر واپس جا سکتے، تو آپ کس سال کا سفر کریں گے اور کیوں؟
23. اسکول میں آپ کا پسندیدہ مضمون کیا تھا
24. آپ خود کو پانچ سال میں کہاں دیکھتے ہیں
25. آپ کی پسندیدہ چھٹی کی خاندانی روایت کیا ہے
26. ایک مشغلہ جسے آپ اپنانا پسند کریں گے
27. اگر آپ پانچ سال پیچھے جا سکتے اور اپنے آپ کو کچھ بتا سکتے، تو آپ کیا کہیں گے
28. آپ کا سب سے قابل فخر کارنامہ کیا ہے
29. آپ کا پسندیدہ مصنف کون ہے اور کیوں
30. آپ کونسی سپر پاور حاصل کرنا چاہیں گے اور کیوں

English:

1. Tell us about a place where you would want to spend your holidays.
2. What has been one of the most interesting experiences you have had at LUMS?
3. What was the most difficult thing you faced during the COVID-19 lockdown?
4. Tell us about your experience in the Machine Learning Course so far
5. What improvement would you want to see in the Machine Learning Course?
6. Tell us something about your favorite science fiction character.
7. Give us some reasons for your decision to pursue, or to not pursue, higher education after your undergrad.
8. Tell us something about your favorite book.

9. Tell us something about your favorite TV show or movie.
10. Do you like the possibility of Aliens existing in this universe?
11. What are your thoughts about the Pakistan Cricket Team's performance in the World cup?
12. If you could have dinner with one person from history, living or dead, who would it be and why?
13. Tell us about your favorite place at LUMS.
14. Tell us a little about any course that you took and found interesting at LUMS.
15. Which city would you most like to live in and why?
16. What is your favorite memory while growing up?
17. What would be your ideal way to spend the weekend?
18. What type of movies or tv series do you like to watch?
19. Do you prefer to watch movies at home or at the theater and Why?
20. Do you prefer watching TV series instead of movies and why?
21. If you could only eat one meal for the rest of your life, what would it be?
22. If you could go back in time, what year would you travel to and why?
23. What was your favorite subject in school?
24. Where do you see yourself in five years?
25. What is your favorite family holiday tradition?
26. What's one hobby you'd love to get into and why?
27. If you could go back in time five years and tell yourself something, what would you say?
28. What is your proudest accomplishment?
29. Who is your favorite author?
30. What superpower would you like to have and why?

For this group, you are supposed to record these sentences word by word.

Mixed:

1. ویسے تو میں ہر کھیل کو اچھا سمجھتا ہوں but hockey is one of my favorite sports.

2. آج بہت زیادہ گرمی ہے so it will rain probably.
3. گذشتہ سالوں میں سائنس نے حیرت انگیز ترقّی کی ہے and inventions of new wonders has astonished the mankind.
4. مواصلات سے مراد ایک ایسا طریقہ کار ہے جس کے ذریعے information is transferred easily from one place to another.
5. Plantation of trees and flowers is quite useful and necessary جس کے ذریعے ایک صحت مند ماحول کی نشونما ہوسکتی ہے
6. Human life is getting in serious danger day by day جیسے جیسے ماحول کی آلودگی میں اضافہ ہورھا ہے
7. Islamabad is not only one of the most beautiful cities of Pakistan بلکہ اسے ہمارے ملک کا دولحاکومت بھی شمار کیا جاتا ہے
8. The increasing population of Pakistan is approximately two twenty million or اسکو دنیا کا پانچواں کثیر آبادی والا ملک کہا جاتا ہے.
9. میں اپنے دوستوں کے ساتھ مل کے باغ کی سیر کو گیا and we spent a quality time there together after so long.
10. صحت کی تندارستی کے لیے صرف ورزش ہی ضروری نہیں but it is also necessary to take good and healthy diet.
11. English language is the most common language جو دنیا میں سب سے زیادہ بولی جاتی ہے .
12. ٹیلیوژن ہماری تفریح کا ہی سب سے بڑا ذریعہ نہیں but it is also an important source of knowledge as well.
13. Northern areas of Pakistan are full of tourists کیوں کے خوبصورت پہاڑ اور دریا سیر و تفریح کی اہم ترین وجہ ہیں .
14. My best friend is fond of reading books اِس لیے وہ زیادہ وقت مطالعہ کرتے ہوئے پایا جاتا ہے
15. There are billions of stars in the space مگر یہ آسمان کا بہت معمولی حصہ ہیں .
16. Allama Iqbal is our national hero and a great poet اور انہوں نے اپنی شاعری سے مسلمانوں کو بیدار کیا
17. محنت کرنے والا کبھی ناکام نہیں ہوتا as hard work is the key to success.
18. We got extremely late for the event yesterday کیوں کے جلسے جلوس کی وجہ سے سڑکیں بند تھیں .
19. ایک کامیاب ملک کے لیے ضروری ہے that there should be fair and just leaders.
20. Her car met an accident on the main road yesterday کیوں کے وہ بہت تیزی اور لاپرواہی سے گاڑی چلا رہی تھی .
21. Ali and Usama both are best friends since childhood اِس لیے دونوں ایک دوسرے کی ہر مشکل میں مدد کرتے ہیں .
22. Travelling is one of his favorite hobby اسی لیے وہ دُنیا کے مشہور ممالک گھوم چکا ہے
23. Smoking is injurious to health or اس کا زائد اِستعمال کینسر کا باعث بنتا ہے .
24. مسلسل سفر کرنے کے بعد they all went to sleep immediately.

25. More pollution is found in cities جب کے دیہات فضائی آلودگی سے پاک رہتے ہیں .
26. Corruption has been a leading disaster for many years in Pakistan اسی لیے ہمارے ملک میں یہ معاشی استحصال کی بڑی وجہ ہے .
27. Our government should build more schools and universities تاکہ ملک کے تعلیمی نظام کو بہتر بنایا جا سکے .
28. Karachi is considered one of the biggest cities of Pakistan کیوں کے یہ RAQBAY اور آبادی کے لحاظ سے کافی بڑا شہر ہے .
29. I have often seen her mistreating people شاید وہ ایک سخت مزاج اور با تمیز لڑکی ہے
30. She cleans her room on regular basis کیوں کے اسے صاف ماحول میں رہنا پسند ہے .
31. Today موسم خشگوار ہے so let's go باہر
32. اج mom cooked دال dinner میں
33. وہ daily exercise کرتا ہے
34. my day was good لیکن I wasn't too خوش جیتنا ہوم ورک مجھے ملا تھا ویسے تو .
35. children بہیر ground میں hide and seek کھیل رہے ہیں .
36. One should drink at least پندرہ glasses of پانی دن میں
37. وہ day کی five times ki نمازیں regulary پڑھتی ہے .
38. Today weather بہت زیادہ pleasant ہے .
39. Ayesha آج school نہیں اسکتی because رات کو she met a car accident.
40. All five دوست went to سنیما to watch a موودی

Disclaimer: Please do not record your personal information such as your name, roll number, cell number, CNIC, etc. as part of your answers.

Recording Instructions:

- You are required to use Praat for recording.
- Record mono sound with a sampling frequency of 16000 Hz.
- Watch this short introductory video on Praat: https://www.youtube.com/watch?v=ibLra0Yk36A
- Record each answer separately and save them in separate files e.g. 1.wav, 2.wav where 1 & 2 are question numbers
- Each recording must have a .wav extension.
- Utter sentences at your normal pace, don't be too fast, and don't be too slow either.
- Use a good headphone mic for recording.

- Don't place your mic directly in front of your mouth and nose to avoid breathing noise.
- Don't use a lot of uhhs, umms, hmm, breaks, and try to use proper words.
- Try to avoid background noise as much as possible
- Make sure your recorded answer doesn't exceed 12 - 15 seconds. For your own convenience, you should write your answers beforehand on paper/ MS word and then record them. There are significant acoustic differences between read and spontaneous speech.

- Name all the recordings "<language code>-<recording number>-<roll number>". Language codes are "ur" for Urdu, "en" for English, and "ue" for mixed sentences.

- Place all the recordings of your group in a single folder and upload them on google drive. In case any member fails to record data, or does not record all the data, the group will have to face a penalty.

Dataset:

For this part, you will be using the data recorded by your group. You have approximately 400-600 recordings (depending on your group size) which you will split into train and test data by yourself. Please note that it is **mandatory** to use data from all the members of the group. Otherwise, your scores might be lower than expected, and we might penalize you for that.

Note: Each member of the group has to attempt the Logistic regression part individually.

Feature Extraction:

In the feature extraction step, you will represent each WAVE file by 13-dimensional Mel-frequency Cepstral Coefficients (MFCCs). MFCCs are a widely used representation of human speech. A code snippet is provided here that shows how to install the required library and read & represent a sample WAVE file with an MFCC vector.

Note: You yourself have to append $x_0 = 1$ to handle bias.

Part 1:

Implement Multinomial Logistic Regression from scratch keeping in view all the discussions from the class lectures to classify the audios into the three classes specified. Feel free to read Chapter 5 of the Speech and Language Processing book to get an in-depth insight into the Multinomial Logistic Regression classifier. Specifically, you'll need to implement the following:

- Softmax function
- Cross-entropy loss function (for multinomial logistic regression)
- Batch Gradient Descent

- Prediction function that predicts the label of test recordings using learned multinomial logistic regression
- Evaluation function that calculates classification accuracy, macro-average (precision, recall, and F1), and confusion matrix on the test set.
- Report plots with no. of iterations/ epochs on the x-axis and training & validation loss on the y-axis. Try out different combinations of learning rates and epochs.

Part 2:

Use scikit-learn's Logistic Regression implementation to train and test the logistic regression on the provided dataset. Use scikit-learn's accuracy_score function to calculate the accuracy and confusion_matrix function to calculate confusion matrix on the test set.

Submission Instructions:

Each group member is supposed to submit this part individually. Use Google Colab for this assignment since you will be uploading all your group's data on Google drive, and will use that link in your Colab file. All students are supposed to submit:

1) The .ipynb file.
2) The .py file.
3) The link for the google drive folder in the first cell of your .ipynb file.
4) The written transcripts of their answers for the English and Urdu part of data generation. Make a word file, and write down all your recorded answers there, and submit its .pdf version. You can type your Urdu answers in Roman Urdu.

Bonus:

If you can speak a language, or know someone who can speak a language, other than Urdu or English, record around 100 sentences in that language and include that in your data and classification. There will be a 10% bonus for this part.