Student name: Nguyễn Minh Tâm
Student ID: 1952968

Question set 8

Part I:

1. Big data can be applied to the education field. We can develop a system that can track the students' activities on e-learning such as BKeL. The system can know how many times students access a page on BKeL and can have an evaluation of the overall performance of students.
   Reason: Because there are still lots of students fail to pass the course every semester
2. Choice 1: 66
   Reason: F_(t+1) = alpha * A_t + (1 - alpha) * F_t = 0.4 * 60 + (1 - 0.4) * 70 = 66
3. Choice 2: content-based uses user profile and item profile, collaborative filtering uses user ratings
   Content-based recommendation uses the user profile and item description to recommend items for the customers.
   Example: If the customer wants to watch a movie, the system will base on the user profile to find what type of films the customer likes and base on the film description to find all films that match the customer.
   Collaborative filtering recommendation uses the user ratings or users' past behaviour to recommend items for the customers.
   Example: If the customer wants to watch a movie, the system will base on the customer's past ratings on the watched film to find the movie that has a similar type to the film having the highest rating.

Part II:

1. In the course assignment, we develop an application that can classify the comments into 3 categories: Positive, Neutral and Negative. During the assignment, our team has conducted 4 different algorithms and compared them together to find the most suitable algorithm that can be applied to the application.
   a) Naïve Bayes
   Naive Bayes is a classification algorithm based on the Bayes Theorem with independent assumption among predictors.
   Advantage of Naïve Bayes in this assignment:
   - Simple and easy to implement
   - Does not require much training data
   - Fast and can make real-time prediction

   Disadvantage of Naïve Bayes in this assignment:

   - Low accuracy

   In this project, we use Multinomial Naive Bayes to classify. It is a variant of Naive Bayes families and is often used in Text Classification.

   b) Support Vector Machine

Support Vector Machine is a supervised machine learning model that uses classification algorithms for two-group classification problems.
Advantage of SVM in this assignment:

- High speed execution
- Conduct good performance with a limited number of samples

Disadvantage of SVM in this assignment:

- Not suitable for large dataset – Long training time for a large dataset
- Choosing a good kernel function is not easy

In this project, we use Linear SVM to classify.

c) Multilayer perceptron
Multilayer perceptron is a feed forward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers connected as a directed graph between the input and output layers. It uses back-propagation for training the network and is also a deep learning method.
Advantage of MLP in this assignment:

- Can be applied to complex non-linear dataset
- Work well with large input data
- Provide quick prediction after training
- The same accuracy ratio can be achieved with smaller data
- Achieve second highest accuracy in this assignment

Disadvantage of MLP in this assignment:

- Computations are difficult and time consuming
- The proper functioning of the model depends on the quality of the training

d) Long short-term memory
The LSTM model architect introduces three main gates and the memory cell in order to capture the long term information as well as the short term use. The hidden idea behind the LSTM model is that for most results of the gate, the output will be processed through a sigmoid activation function yielding out the value between 0 and 1. This indicates the mechanism of the LSTM model to allow information from the long past as well as the current input to contribute to the current predicted result.
Advantage of LSTM in this assignment:

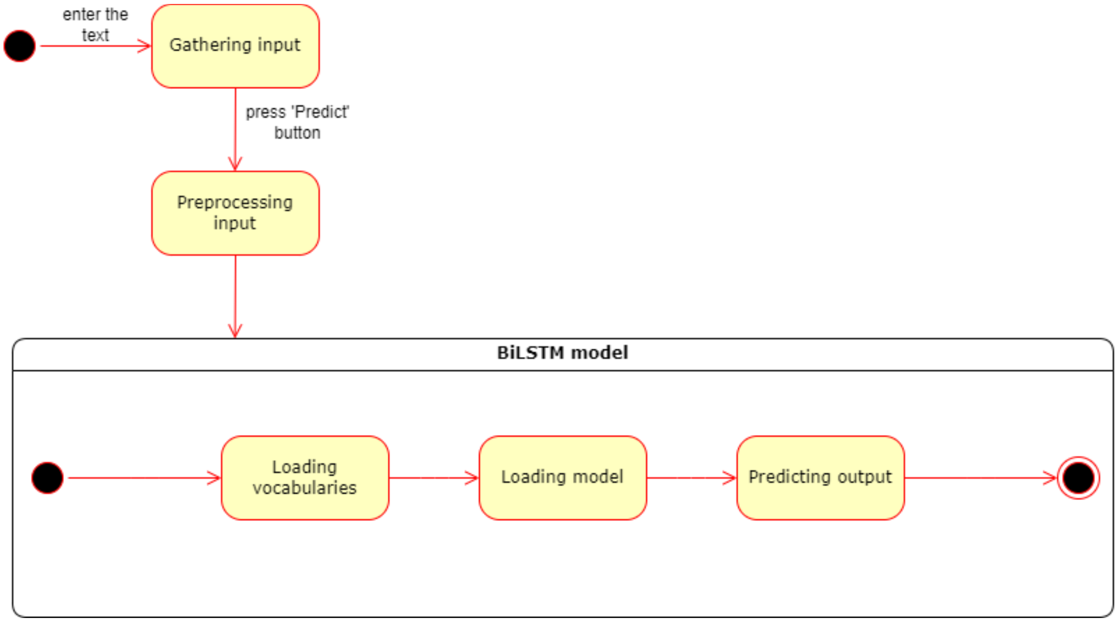
- Provide a large range of parameters such as learning rates, input and output biases
- The complexity to update each weight is O(1)
- We use the Bidirectional LSTM – extension of traditional LSTM, so that we can improve model performance on sequence classification problems -> This algorithm achieves the highest accuracy compared to other algorithms.

Disadvantage of LSTM in this assignment:

- Long training time with large dataset

- Require more memory to train
- Hard to implement dropout in LSTM
- Sensitive to different random weight initializations

2. The statechart diagram of the system in the assignment:


enter the text
Gathering input
press 'Predict' button
Preprocessing input
BiLSTM model
Loading vocabularies
Loading model
Predicting output