Solutions:

I have 3 features_matrices with 3 diffrent group of features as explained bel

1. Feature Engineering Approach 1:

    a. Combined the user information of user_assesment_scores and user_intere
    taken an assesment.
    b. On analysing the data, although all the given 10000 users have express
    have taken the assesment. Hence there was loss of information in combinin
    c. Applied log transformation to the user_assesment_scores to achieve nor
     'ratings' using unsupervised discretization with k-means algorithm. Used
    d. Though the information was not present for all the users, the similari
    e. For example, consider the users (user_handle: 12 and 4343), both the u
    and scored close scores of 170 and 164 respectively.

2. Feature Engineering Approach 2:

    a. Combined the user information of user_course_views and course_tags for
    b. Joining user_course_views with course_tags on course_id to get the cou
    c. Creating a new Feature combining Course_tag and Level of the course. D
    d. Aggreating(sum) the number of seconds an user has viewed a course desc
    e. Removing the entries where a user a viewed a course for '0' seconds¶
    f.Applying log transformation to the aggregated number of seconds to conv
    g. Creating a new feature 'Rating':
        0 - if the user has not viewed a course
        1 - if the log_transformed_seconds in the range (0 - 1st quantile),
        2 - if the log_transformed_seconds in the range (1st quantile - 2nd
        3 - if the log_transformed_seconds in the range (2nd quantile - 3rd q
        4 - if the log_transformed_seconds in the range (3rd quantile - max)

    h. Although there was little loss in information the similarity was consi
    for example, consider the top 3 similar users of user_handle : 10000 - us
    the course 'apex-absolute-beginner-guide-coding-salesforce' with 'beginne
    has literally the viewed same courses.

3. Feature Engineering Approach 3:

    a.The user_interest was the only with the information of all the given 10
    a course atleast once.
    b. Used the interest_tag as the features for this model.

Solution for the questions given:

1. The metric chosen for similarity calculation is Cosine Similarity with Mea

   a. Adjusted Cosine is invariant to scaling of the vector. Hence it is rob
      while increasing the number of users or for more expressive features t
   b. On calculating the coefficient based on all the elements in the vector
      the similarity between users that may not have many features in common.
   c.Hence it takes care of 2 main issues, dealing with uncommon features be
   d. Adjsuted Cosine is statistically simliar to Pearson Correlation robust
   e. Projection wise cosine similarity is dependent on only the direction o
   f. The cosine similarities of a subset of the original data are the same

2. I would consider the following steps to scale the product.

   a. Starting with storing the data, I have used SQLite for solving the pro
   database or file storage like HDFS, Azure Data lake could be efficient.
   b. Considering the data is stored on a distributed environment (Databrick
   to implement parallel processing in the pipelines. For example consider d
   be accomplished in parallel which is highly efficient.
   c. Rendering the product to end-user should be a more containerized versi
   or deploying the model on the cloud will provide substantial increase in
   d. Using DAG engines like Apache Spark or Tensor Flow for numerical compu
   process and computations internally through lazy evaluation. DAG engine w
   graph assisting performance improvement. This could prove incredibly effi
   which increases the computation at scale.
   e. Perform computation on GPU
   f. Perform random or stratified sampling (to ensure users from all groups

3. Considering the scale and diversity of the users the company serves, for a

   a. Build a more robust API by changing the DB storage to distributed stor
   b. As mentioned, definitely gather more data on the users (Job/student, p
      Completed the Course/Not Completed/ Assessment Scores/ User Reviews/ e
      expansive and expressive features. Also in the case of a Supervised Le
   c. Track user journey through clickstream data to identify friction point
   d. Develop end-to-end pipelines by covering more features like mentioned
   e. Careful consideration of the points mentioned above for scaling the so