**Solution to Session 6**

**Building word embeddings from the course corpus and evaluating the embeddings using wordsim353 data:**

**Combinations of window size and embedding size changed and the changes in the similarity is listed below:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rows** | **Gensim model name** | **Window size** | **Embedding size** | **Similarity calculation** |
| **0** | **gensim\_model** | **3** | **120** | **0.5734946742812335** |
| **1** | **gensim\_model1** | **2** | **90** | **0.5209471279144137** |
| **2** | **gensim\_model2** | **3** | **90** | **0.5655192987627357** |
| **3** | **gensim\_model3** | **2** | **120** | **0.5424663022975982** |
| **4** | **gensim\_model4** | **2** | **100** | **0.5289894021818018** |
| **5** | **gensim\_model5** | **10** | **100** | **0.7080312298073472** |
| **6** | **gensim\_model6** | **15** | **100** | **0.7058135290022604** |
| **7** | **gensim\_model7** | **20** | **200** | **0.7168701903801445** |

**As we see, from rows 0 and 2, having the same window size, but decreasing the embedding size has a negative effect on the similarity calculation on the tested data.**

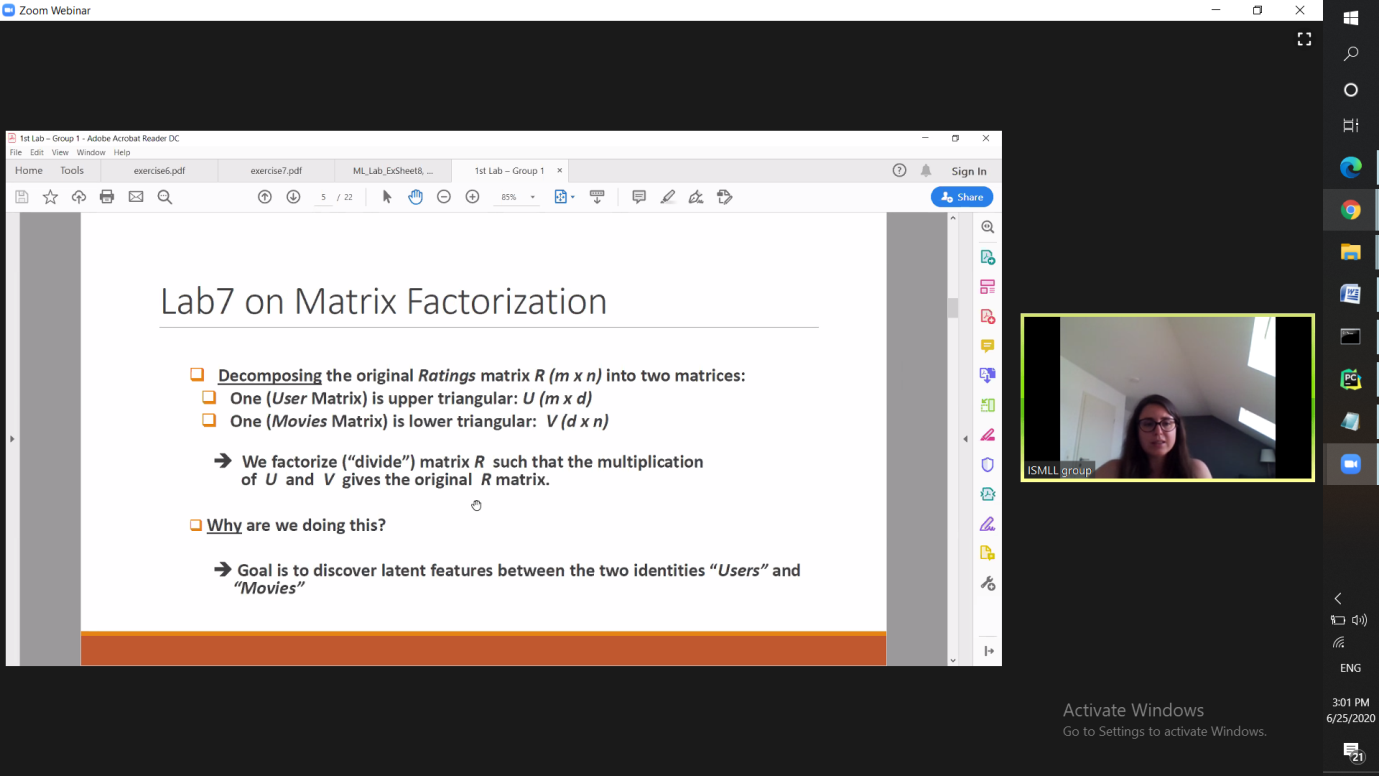
**From rows 1, 3,4 , the same observation is inferred that is as the embedding size increases, the value of the similarity between the wordsim353 data and our gensim model’s predictions are high.**

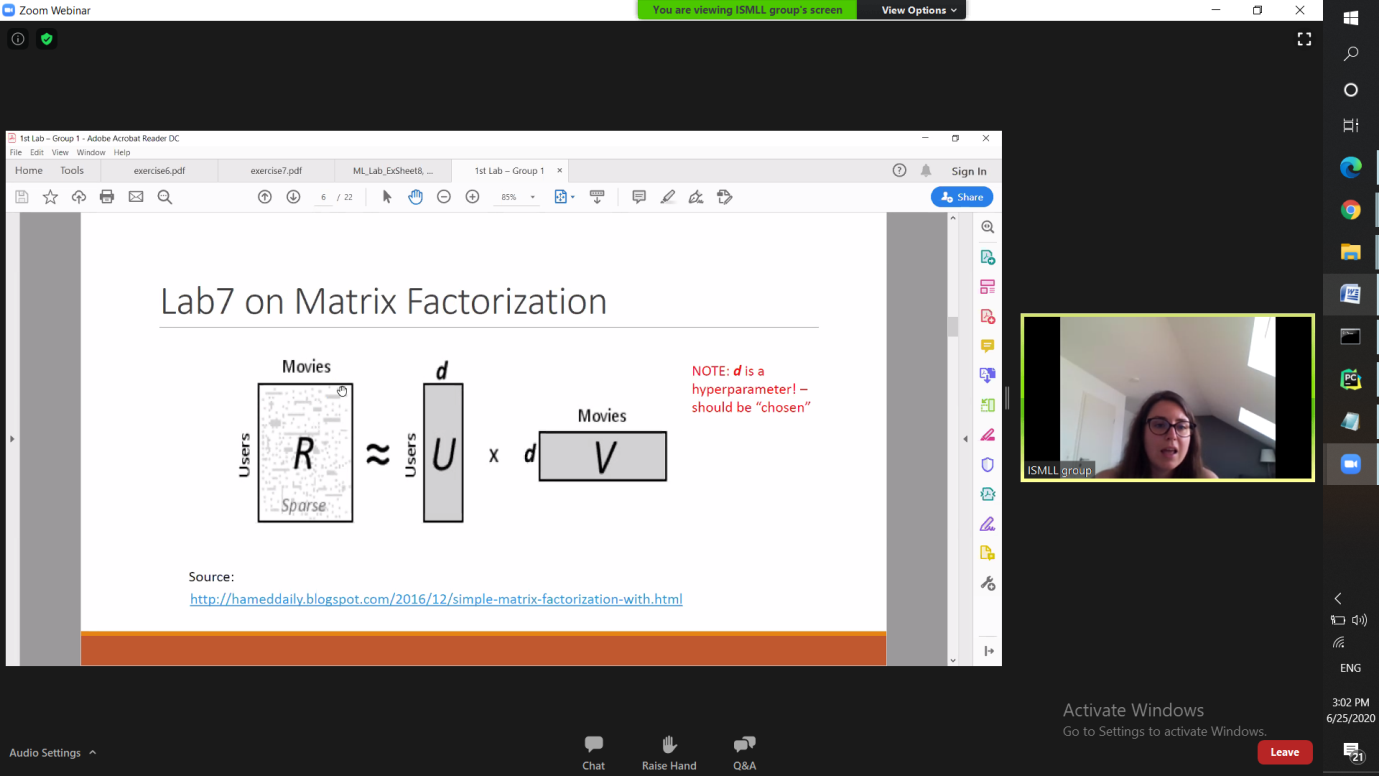
**From rows 5,6, it is observed that when the window size is increased, keeping the embedding size the same, the similarity has decreased.**

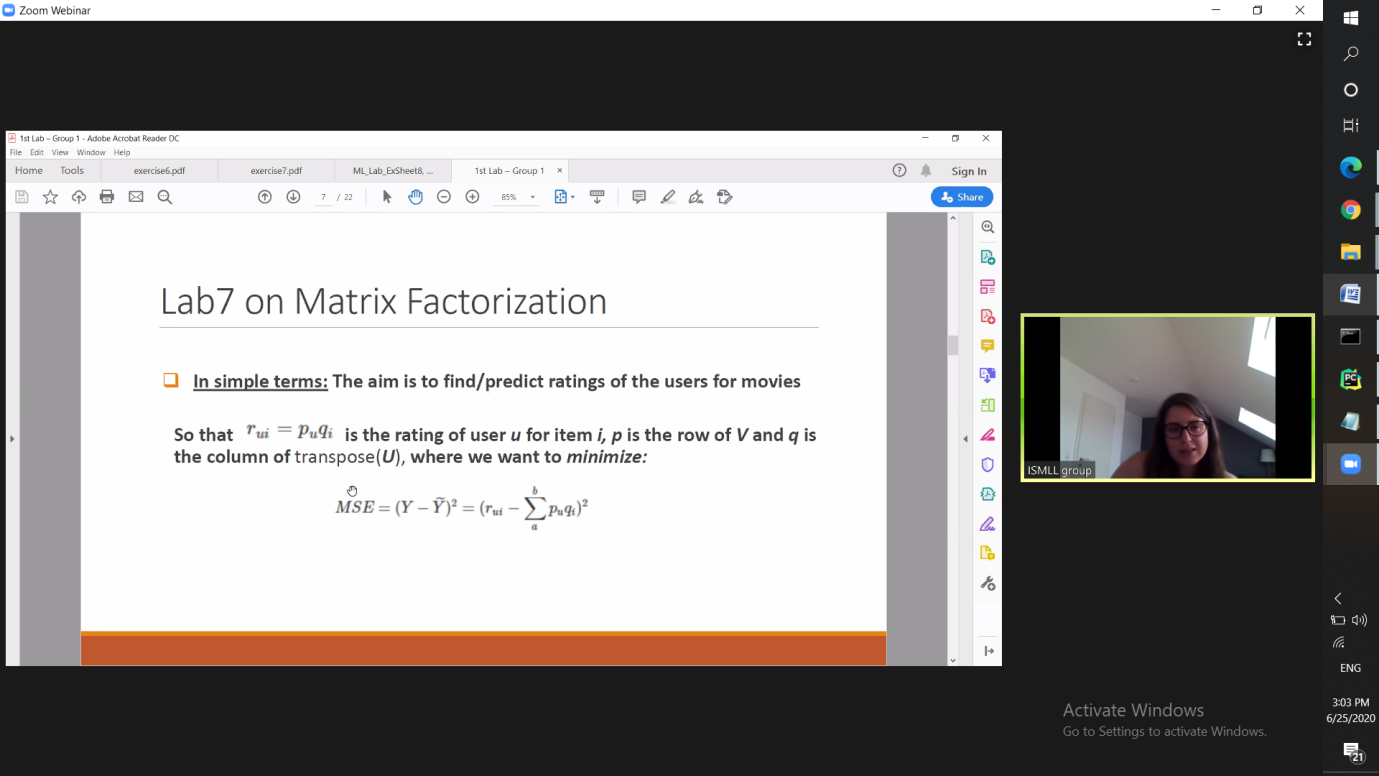
**In row 7, increasing the window size and the embedding size has increased the similarity between the wordsim353 benchmark test and our gensim model’s prediction and the similarity is 0.7168701903801445.**

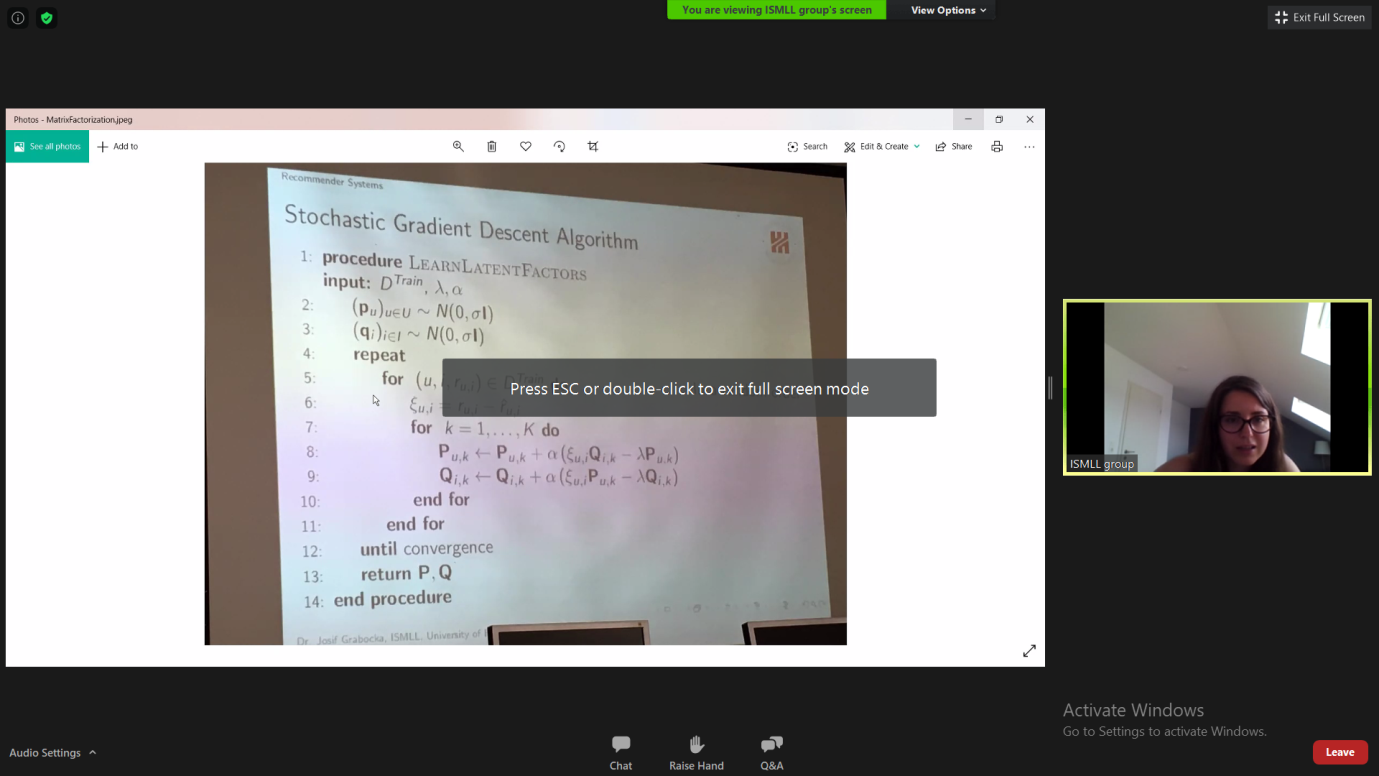
**The optimal one through the above experiment is achieved by having the window size as 20 and embedding size as 200.**

**The difference in similarity between row 5 and 7 is so low, but the time taken to execute row7 is way higher than time taken for row 5. SO we could also use the settings from row 5 as the time taken is less.**

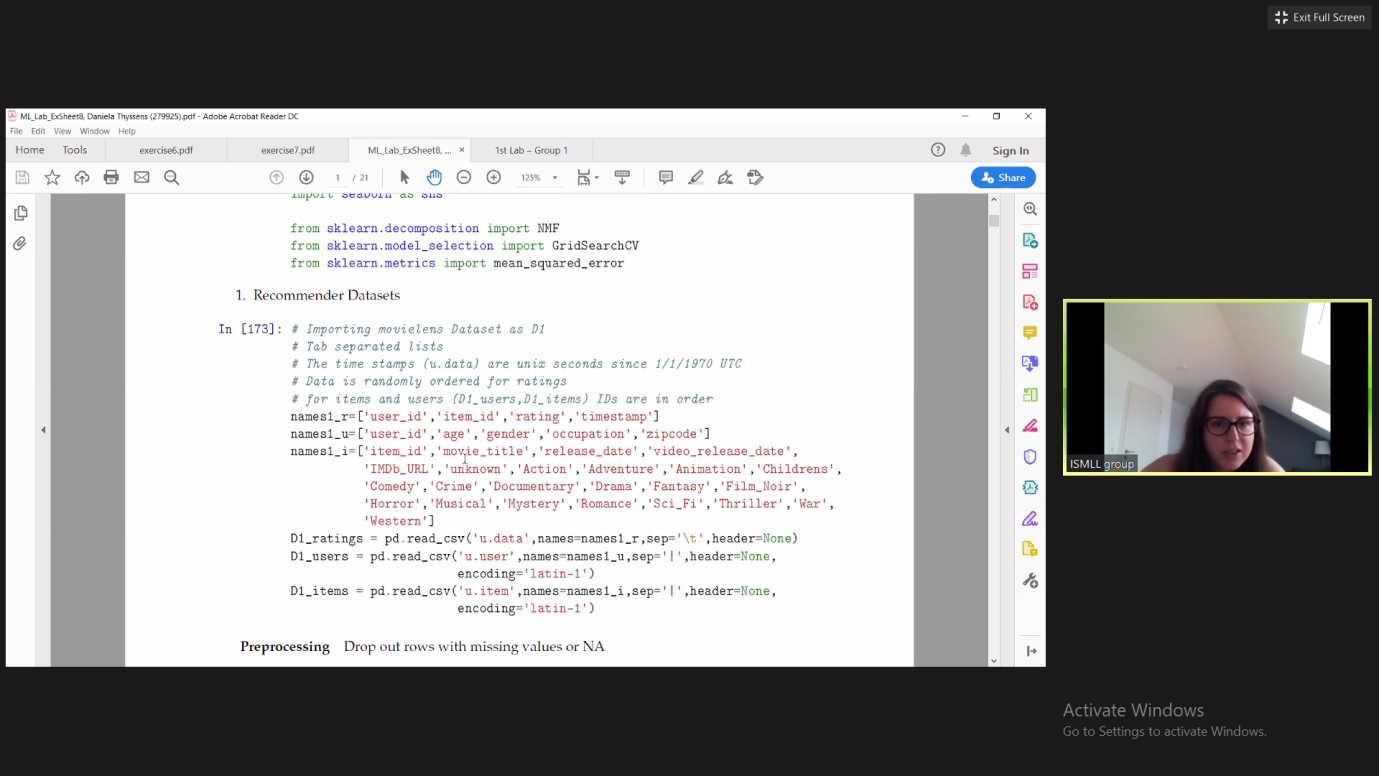
****

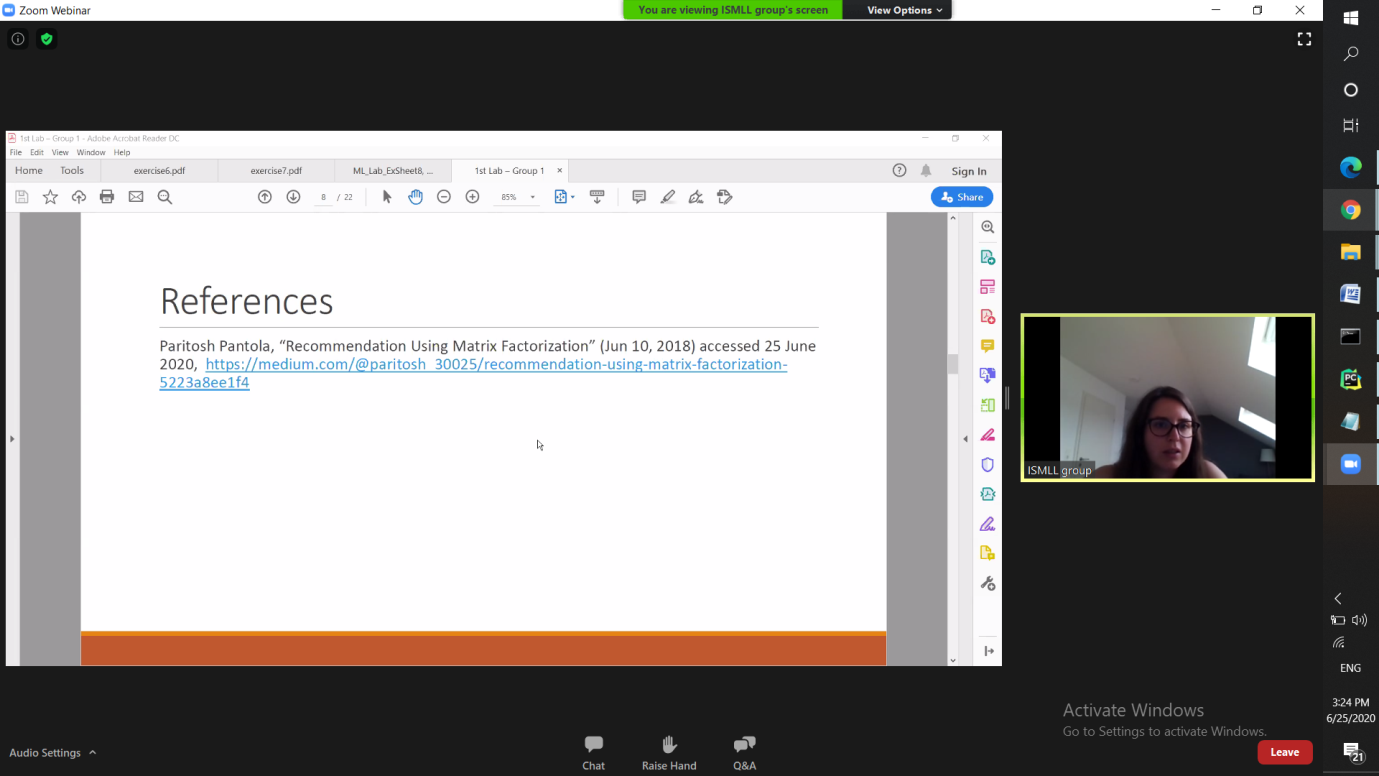
****

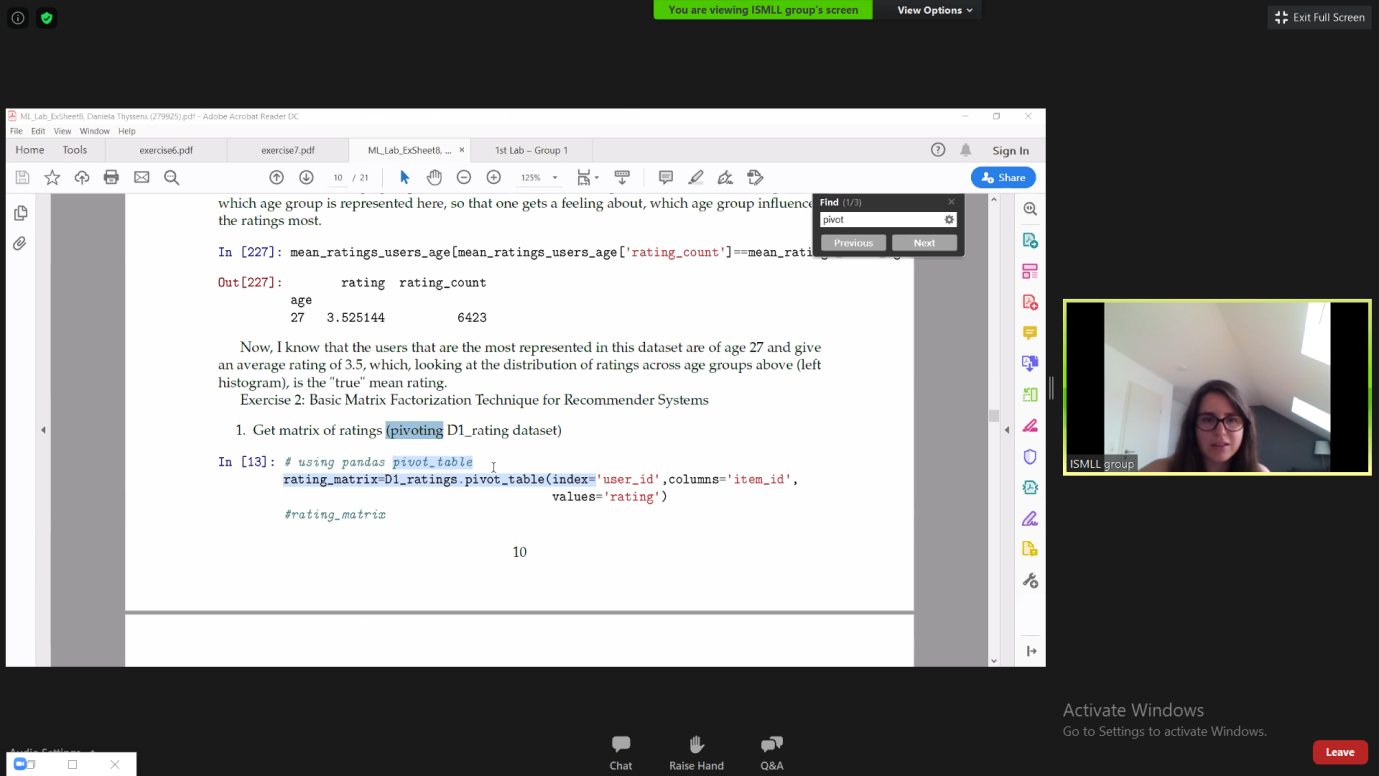
****

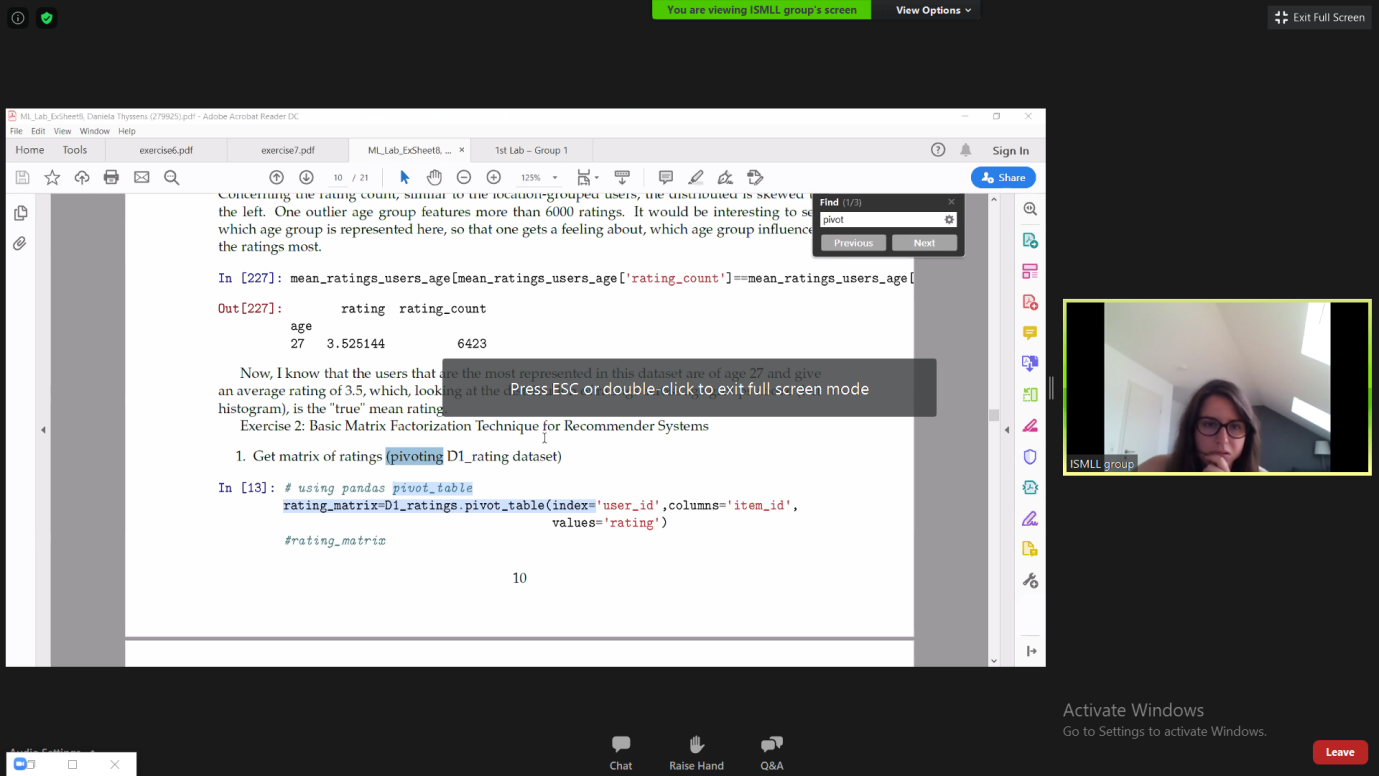
****

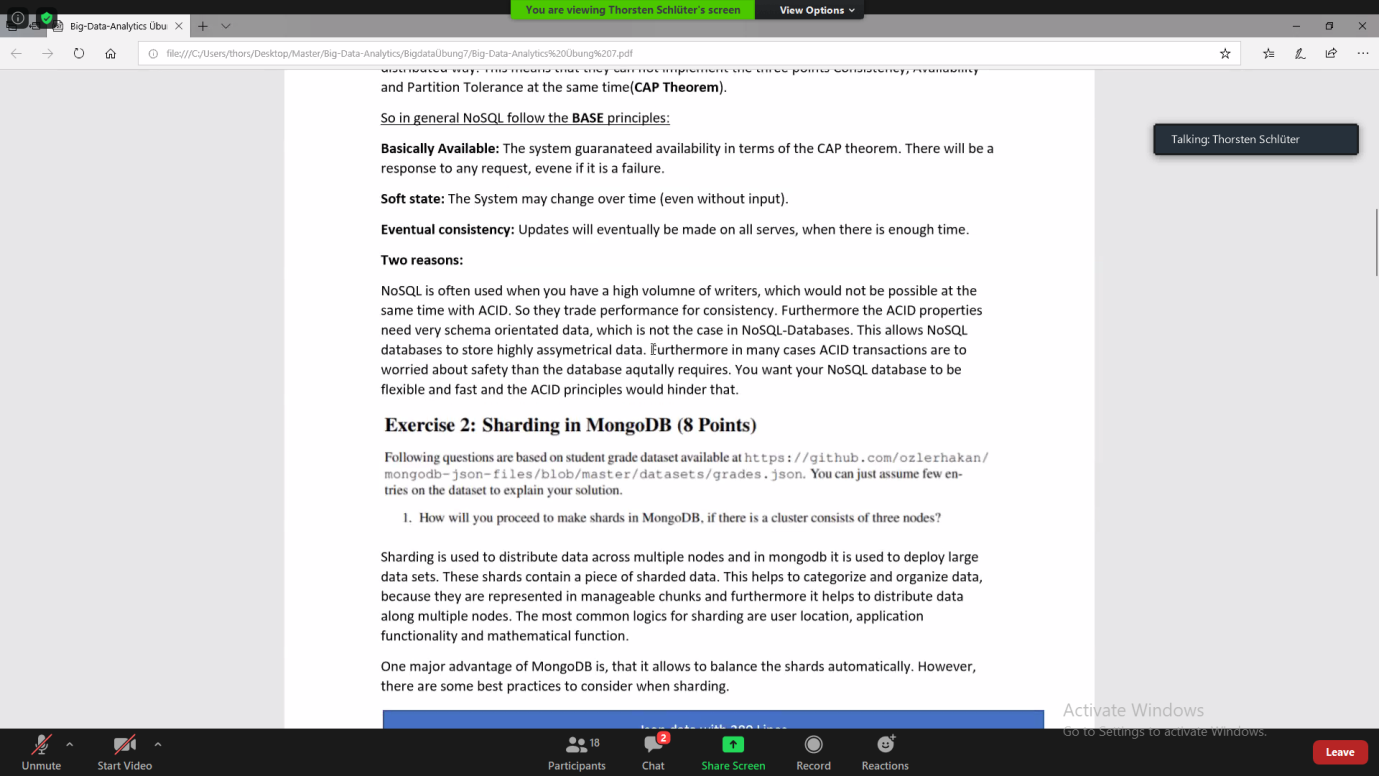
[**https://medium.com/@paritosh\_30025/recommendation-using-matrix-factorization-5223a8ee1f4**](https://medium.com/@paritosh_30025/recommendation-using-matrix-factorization-5223a8ee1f4)

****

****

****

****

****