Week 10 - Lab Session Results

March 5, 2025

Content-based recommendation

Exercise 1

In this exercise, you will build user profiles using TF-IDF vectors and use them to get the recommended items.

Based on the TF-IDF vectors obtained in the Exercise 2 from Week 09, represent each user in the same vector space. Amongst other feasible solutions, you can represent a user (creating a user profile) by computing the weighted mean of the items vectors, from the items that have been rated by the users in the training set. Reflect on this way of creating the user profile; is there a better way to make use of low ratings?

For all users, compute the cosine similarity with each product that they have not rated in the training set (**unobserved ratings**). Take the top-5 items with the highest cosine similarity as the top-5 recommended items.

What are the top-5 recommended items for user `A39WWMBA0299ZF`? Print out the top-5 items for said user and their similarity score, rounded to three decimal places.

```
981 users loaded in the train set
949 users loaded in the test set

Top-5 recommended items for user 'A39WWMBA0299ZF':
[('B019FWRG3C', 0.419),
 ('B00W259T7G', 0.19),
 ('B00IJHY54S', 0.088),
 ('B0006O10P4', 0.081),
 ('B00006L9LC', 0.079)]
```

Exercise 2

In this exercise, you will evaluate the content-based recommender system in Exercise 1.

Compute the hit rate for the content-based recommender system from Exercise 1. Evaluate the hit rate based on the top-5, top-10 and top-20 recommendations, averaged over the total number of users. Round your final answer to 3 decimal places. Remember that, as we are evaluating the system, you should compute the hit rate over the **test set**. How well/bad does this content-based approach perform compared to the collaborative filtering approaches?

```
Hit Rate (top-5): 0.419
Hit Rate (top-10): 0.448
Hit Rate (top-20): 0.510
```

Exercise 3

In this exercise, you will create a content-based recommender system based on word2vec embeddings and evaluate its performance with hit rate.

Repeat Exercise 1 and 2, this time representing the items and users in a word2vec vector space. You may use the gensim library and download the 300-dimension embeddings from Google. Source: https://radimrehurek.com/gensim/models/word2vec.html#pretrained-models

Remember to follow the same preprocessing pipeline as instructed in Lab W9, skipping the stemming step. Think on why we should not perform stemming when working with word2vec embeddings.

```
[==================================================] 100.0% 1662.8/1662.8MB
downloaded

Total number of items: 84
Vocabulary size before preprocessing: 545
Vocabulary size after preprocessing: 484

Top-5 recommended items for user 'A39WWMBA0299ZF':
[('B000LIBUBY', 0.782),
 ('B019FWRG3C', 0.78),
 ('B000W0C07Y', 0.767),
 ('B0012XPRO8', 0.758),
 ('B00HLXEXDO', 0.747)]

Hit Rate (top-5): 0.412
Hit Rate (top-10): 0.427
Hit Rate (top-20): 0.490
```