Quiz #7: Recommendation Systems

Name: GRADER ID: ____________

1) (6 pts) Given 300,000 news articles, the first task is to a) divide the articles into several categories, b) randomly select one category and then divide the selected category into several categories. Articles within the same category should have similar semantically (e.g., sports vs. politics). Briefly describe how TF-IDF can be used to achieve this task. Your description should start from reading the files. (3 pts) Now that you have the TF-IDF results, briefly describe how you can build a user profile (i.e., content-based recommendation) and use MinHash and LSH to recommend articles for the user. You should discuss how the choices of b and r would affect your recommendation results (3 pts)

ⓐ [1 POINT] ① Read the news article and tokenize them

[1POINT] ② Construct item profiles of the articles using TF-IDF

ⓐ Remove most frequent words and stop-words

ⓑ Remove rare words

ⓒ Concentrate on useful words with high TF-IDF scores.

ⓓ Use these words to best characterise the topic of the document

ⓔ Use cosine distance/Jaccard sim to measure similarity

[1 POINT] ③ Using ②ⓔ divide into several categories like sports, news etc. Then use sports to differentiate between golf tennis football etc.

ⓑ ① Construct user profiles by creating vectors with same components that describe users preferences.

[1POINT] ② Find similar users/items by creating minhash signatures from TF-IDF scores to reduce complexity

③ Use LSH to find the similarity by creating bands and placing item profiles in buckets.

④ Identify in which bucket we look for items that have small distance from user

2) (1pt) Briefly explain one advantage and one disadvantage of using Decision Trees for finding recommendations compared to using the Cosine Distance.

[0.5] Advantage - more accurate and works on small ~~dataset~~ problem size

[0.5] disadvantage - Consider different predicates/complex combination

3) (3 pts) What are the two common evaluation metrics of recommendation systems discussed in the article "Recommender Systems, Prem Melville and Vikas Sindhwani, Encyclopedia of Machine Learning, 2010" in addition to Precision, Recall, and AUC? The main difference between the two evaluation metrics is that one of them emphasizes on a type of error. What is that type of error?

[2 POINTS] The two most commonly used metric is Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

[1 POINT] RMSE puts more emphasis on larger absolute ~~[illegible]~~ errors.

[1 POINT]

⑤ Use a prediction heuristic - Estimate degree to which a user would prefer an item by computing cosine distance b/w user profile and item profile.

⑥ Idea of LSH is to reduce complexity of comparing large number of pairs. So bands and rows help in this trade off as hashing samples is less costly.

[1 POINT]

→ If you have r>>b,
· makes dissimilar pairs even more dissimilar
reduces False positive and increase False Negative

→ If you have b>>r,
reduces False Negative and increase False Positive

b	FP	FN
↑	↑	↓
r		
↑	↓	↑