How to evaluate performance of recommender systems?

Qualitative evaluation through metrics

- There is a common practice to keep a **hold out set** and evalaute the quality using **human evaluators**
- Serependity, Diversity, Novelty are other attributes that are also important for a recommender system
- Serependity: Ability of the model to **help users discover new interests**.
    - If the ML system treats the **user in isolation,** it may **not know the user is interested** in a given item, but the model might **still recommend** it because **similar users** are interested in that item when it looks at **aggreagted preferences**.
- Recommender system that should suggest novel, relevant and unexpected items also help in **diversification and popularity bias**.

Quantitative evaluation through metrics

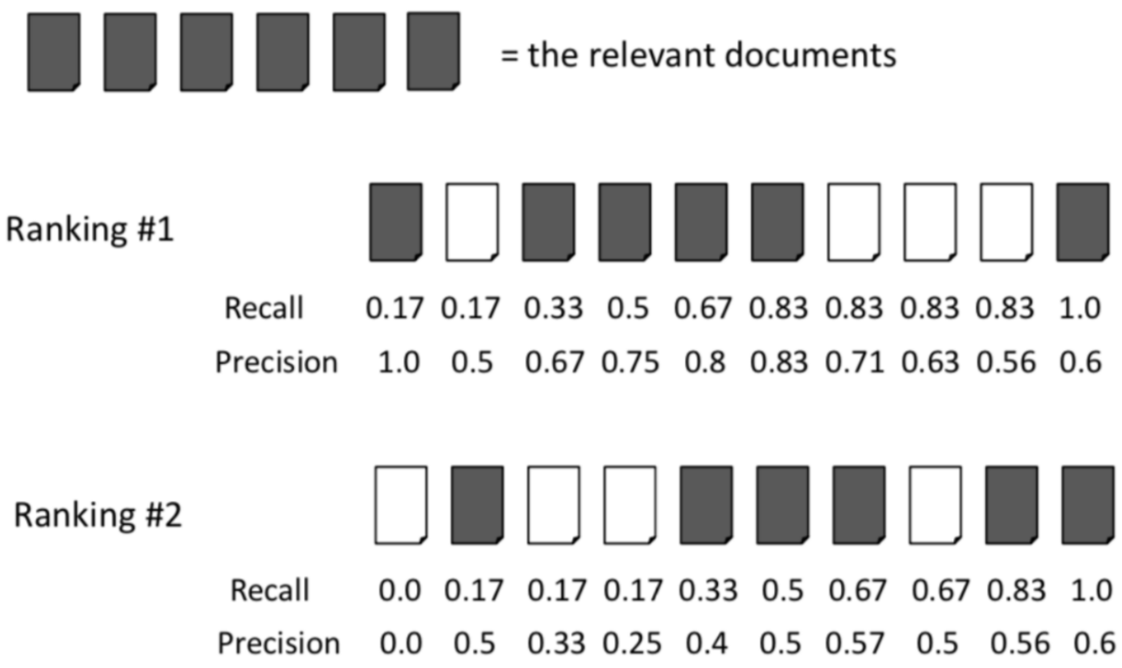
Precision at k: Proportion of recommended items in the top-k set that are relevant

- Precision@k = (# of recommended items @k that are relevant) / (# of recommended items @k)
- Ex: If precision at 10 in a top-10 recommendation problem is 80% - 80% of the recommendation I make are relevant to the user.

Recall at k: Proportion of relevant items found in the top-k recommendations

- Recall@k = (# of recommended items @k that are relevant) / (total # of relevant items)
- Ex: If recall at 10 is 40% in our top-10 recommendation system. This means that 40% of the total number of the relevant items appear in the top-k results.

Illustration of precision and recall at 10 for two ranking models



Precision vs. Recall

- Precision as metric in recommenders optimise the **ability of how good we are in retrieving relevant items** (already good rated) but **too bigger** precision **lower the ability to give unique and new recommendations** (false positives).
- Optimising for recall makes sense when **number of relevant items are less than recommended items**
- Here let us look at 100 recommend items & let us look at both precision and recall

Metrics for ranking:

- Coverage, Hit Rate, F1 are some other accuracy related metrics for recommender systems
- MAP, DCG, NDCG, MRR are some other metrics that also considers order (rank of items) while evaluating recommender systems: Please refer to this article: https://towardsdatascience.com/ranking-evaluation-metrics-for-recommender-systems-263d0a66ef54

Colab paid products - Cancel contracts here