Learning from user generated data - exam 20210629

Q1

(5p) Gap text: Select terms
An information retrieval (IR) system assumes that a user has a/an **[information need]**, which he or she expresses as a **[query]** to a/an **[search engine]**.
The system then tries to find the most **[relevant]** documents/items with respect to the **[query]**.
A recommender system (RS) is a/an **[information filtering]** system, which provides a **[personalized]** perspective on the underlying item catalog. Creating recommendations requires a/an **[user profile]**, which can hold **[explicit]** preference information (e.g., ratings) or **[implicit]** (e.g., song skips) preference information.

Q2

(6p) Dropdown options
Herlocker et al. have identified 10 reasons why people use recommender systems. For each of the following reasons, select the most appropriate implication or example.
Improve Profile → A **user rates lots of items** to receive better recommendations.,
Express Self → Users **share all their ratings**, recommendations, and reviews on social media.,
Find All Good Items → The recommender system achieves a **high recall**.,
Recommend Sequence → An algorithm for **automatic playlist continuation** recommends a coherent list of recommendations.,
Influence Others → Users tag Paris Hilton as "brutal death metal" on Last.fm., **(false tags)**
Find Credible Recommender → A user tries to **fool the system**.

Q3

Dropdown options
Please connect the three monotonicity assumptions to their definitions.
IDF assumption → Rare terms in the collection are **no less important than frequent** ones.,
Normalization assumption → **Long documents are no more important** than short documents.,
TF assumption → **Multiple appearances** of a term in a document are no less important than single appearances.

Q4

Dropdown options
Please choose the correct or most appropriate statement for each type of recommender system.
Demographic filtering → Recommendations for a target user are made by considering his or her **age**.,
Model-based collaborative filtering → Typically adopts **matrix factorization techniques** to create a **joint latent space** of users and items.,
Content-based filtering → **Does not require any behavioral information** from users other than the target user.,
Memory-based collaborative filtering → **Similarity computations** are directly performed on the **user-item-rating matrix**.,
Context-aware recommender system → Uses an **intermediate representation** to connect items to locations.

Q5

Binary choice / true false
Q5 Herlocker et al. identified 8 reasons why people use recommender systems. They include:
influencing others, self-expression, fooling the system, and satisfying their information needs.

Correct - 10 Reasons:
Annotation in Context,
Find Good Items,
Find All Good Items,
Recommend Sequence,
Just Browsing,
Find Credible Recommender,
Improve Profile,
Express Self,
Help Others,
Influence Others

Q6

A recommender system creates a list of recommendations based on a user-provided query, expressing the user's information need.
-> Search Engine not RS

Q7

Common strategies to mitigate unwanted biases of a recommender system include regularization, data rebalancing, and adversarial training.

+ Reranking
+ Filter Items

Q8

Studies have revealed that many common recommendation algorithms favor popular items over lesser popular ones in recommendation lists.

Q9

A parallelized design of hybrid recommenders is an example of an early fusion aggregation approach.

Q10

Filtering and re-ranking are common strategies in a mixed hybridization design.

Q11

The self weighting term in content-boosted collaborative filtering (CBCF) quantifies the confidence the algorithm has in the correlation between users' rating vectors.

Q12

Content-boosted collaborative filtering (CBCF) is based on the idea that real rating data is enriched with the results of a classifier that predicts missing ratings from content-based item features.

Q13

Active user awareness requires the user of a context-aware recommender system to trigger a context-aware update of the recommendation list.

Q14

The user background describes general, rather static information about the user, e.g., knowledge or demographics.

Q15

Context acquisition can be explicit, implicit, or inferring.

Q16

The item context in a multimedia recommender system encompasses additional data that is not encoded in the audiovisual signal of the item, e.g., tags or album cover artwork.

Q17

The evaluation metrics MRR and NDCG consider the position of relevant items in the recommendation list.

Q18

Beyond-accuracy metrics include serendipity, coverage, and diversity.

Q19

Mean absolute error (MAE) disproportionally penalizes larger discrepancies between predicted and true ratings, in contrast to RMSE which does not.

Q20

The average precision (AP) metric accounts for relevant items not in the recommendation list, therefore, implicitly factoring in recall.

Q21

The reciprocal rank metric is the inverse of the highest rank at which the first relevant item is found in the recommendation list.

Q22

User-based CF (in the memory-based variant) tends to scale better with the number of users and items because only items the active user rated are considered when computing user similarity.

Q23

The graph-based transitivity approach can uncover transitive user-item relationships and adopts graph traversal algorithms on a tripartite graph (users, items, ratings).
Biparitite

Q24

In case of binary ratings (i.e., user interacted with item, or did not), no user bias factor is required when computing memory-based CF.

Q25

Model-based CF projects the user-item-rating matrix into a low-dimensional space, using techniques such as SVD.

Q26

The regularization term in the optimization function of an SGD-trained MF model is used to account for user biases during cold start.

Q27

Missing ratings are a problem in SVD-based matrix factorization. This problem can be mitigated by performing multidimensional scaling, e.g., Sammon's mapping with a magic factor of 0.45, on the sparse rating matrix.

Q28

Casefolding as a text preprocessing technique to create a TF-IDF item representation may introduce semantic ambiguities of terms, i.e., terms with different meanings are mapped to the same term.

Q29

The use of a logarithmic term when computing TF and IDF values is motivated by Zipf's law about the frequency of word occurrences in English texts.

Q30

Lemmatization or stemming is used to identify different parts-of-speech in texts when creating a VSM for item representation.

Q31

The use of an exponential correction term when computing TF and IDF values is motivated by Zipf's law about the frequency of word occurrences in English texts.

Q32

The third monotonicity assumption (i.e., normalization assumption) is accounted for by using cosine similarity between the TF-IDF representations of items.

- [unsure]

Q33

The IDF monotonicity assumption states that terms that appear in only a few documents of the corpus are more important than terms that appear in many documents.

Q34

The topic modeling approaches pLSA and LDA assume conditional independence of terms and documents, given a topic z.

- probably true unsure about LDA

Q35

LDA uses principal components analysis (PCA) to discover latent topics from TF-IDF representations of documents.