# Flipped classroom

On Thursdays from 11.00-13.00, there is a hybrid session where the material presented in video lectures is discussed. These sessions will consist of Q&A followed by a flipped classroom part. Afterwards, the questions we discussed will be added to this document.

Questions from students on Zoom can be added to [this spreadsheet](https://docs.google.com/spreadsheets/d/1kA2EMjc7CJHkiPiZw0ZBBEj00n81hz5aMf4FXHOdZoA/edit?usp=sharing).

## RecSys & Recap (Week 7 / March 24)

### Agenda

* Quick recap of exam organization
* RecSys scenario
  + Q&A
  + Flipped classroom questions and discussion  
    <https://www.wooclap.com/BZSKQI>
* General Q&A
* Solutions to example problems

### Which IR ranking methods can be used for effective content-based recommendation?

* **[Correct]** VSM + TF-IDF
* **[Correct]** BM25
* **[Correct]** QL
* **[Correct]** LDA
* **[Correct]** AWE
* **[Correct]** BERT-based methods

### Which user interactions provide strong signals for effective collaborative filtering?

* **[Incorrect]** Clicks
* **[Correct]** Ratings
* **[Correct]** Likes / dislikes
* **[Incorrect]** Mouse hovering
* **[Correct]** Adding to watch list / basket
* **[Correct]** Purchases

### Which of the following are advantages of content-based recommendation?

* **[Correct]** No cold start problem for items (item interactions are not needed)
* **[Incorrect]** No cold start problem for users (user interactions are not needed)
* **[Incorrect]** Biased towards recommending popular items
* **[Incorrect]** Profiles focus on exploitation and strongly limit exploration

### Which of the following evaluation metrics could be used to evaluate YouTube's next video recommendations? (3-5 videos to choose from)

* **[Correct]** Precision@1
* **[Correct]** Precision@3
* **[Incorrect]** Recall@3
* **[Incorrect]** Total recall / set recall
* **[Incorrect]** MAP (Average Precision)
* **[Correct]** nDCG@3

## Conversational Search (Week 6 / March 17)

### Agenda

* Q&A
* Flipped classroom questions and discussion  
  <https://www.wooclap.com/AOYLGE>
* Exercise solutions (time permitting)

### From the perspective of conversational IR, how would Web search on a laptop be classified?

* **[Correct]** Single initiative, user has control
* **[Incorrect]** Single initiative, system has control
* **[Incorrect]** Mixed initiative

### How can an utterance in a conversation be represented in a machine-readable form?

* **[Incorrect]** Vector of user interactions
* **[Correct]** Content-based vector (e.g., TF-IDF, LSI, AWE)
* **[Correct]** Multinomial distribution over terms
* **[Incorrect]** Average representation of previous utterances

### Which of the following scenarios are examples of a single initiative setting in which the system has control?

* **[Incorrect]** Your internet is down, so you notify the company on twitter and wait for their response
* **[Correct]** Your internet is down, so you call the technical support line and interact with a menu
* **[Incorrect]** You're looking for a movie to watch, so you message CriticBot with a description of your favorite movies, and after some clarifying questions you receive a good recommendation
* **[Incorrect]** You're looking for a movie to watch, so you do a Web search for new, highly-rated movies

### Which of the following scenarios are examples of a mixed initiative setting?

* **[Partially correct]** Your internet is down, so you notify the company on twitter and wait for their response
* **[Incorrect]** Your internet is down, so you call the technical support line and interact with a menu
* **[Correct]** You're looking for a movie to watch, so you message CriticBot with a description of your favorite movies, and after some clarifying questions you receive a good recommendation
* **[Incorrect]** You're looking for a movie to watch, so you do a Web search for new, highly-rated movies

### Which of the following are potential downsides of the "feedback first" approach? (In comparison to not asking for feedback)

* **[Correct]** The system must "understand" the previous turn well enough to identify aspects of the user's information need
* **[Incorrect]** If the system has not interacted with the user before, it is difficult to identify personalized feedback
* **[Correct]** Responding to the feedback request may take the user as long as it takes them to inspect a result snippet
* **[Correct]** Asking for feedback may result in several extra turns if the system's first suggestion is not helpful

### Which of the following evaluation metrics is the MOST appropriate for a conversational search (document retrieval) setting?

* **[Correct]** Precision@1
* **[Incorrect]** Precision@3
* **[Incorrect]** Precision@10
* **[Correct]** nDCG@3
* **[Incorrect]** nDCG@10
* **[Incorrect]** AP@3 (Average Precision aka MAP)
* **[Incorrect]** AP@10 (Average Precision aka MAP)

### You are evaluating several components in your conversational search system. For which components are you primarily interested in maximizing Recall@10?

* **[Incorrect]** Document retrieval
* **[Correct]** Question retrieval
* **[Incorrect]** Question selection

### You are evaluating several components in your conversational search system. For which components are you primarily interested in maximizing nDCG@3?

* **[Correct]** Document retrieval
* **[Incorrect]** Question retrieval
* **[Correct]** Question selection

### Why is query rewriting often employed for conversational search?

* **[Incorrect]** Conversations consist of many turns, so a query including history may exceed the maximum input length possible with a state-of-the-art BERT/Transformer-based method
* **[Incorrect]** User utterances are often ungrammatical and thus difficult for a search engine to "understand"
* **[Correct]** Later turns in conversations often use pronouns and other anaphora to refer back to earlier topics
* **[Incorrect]** Conversational search requires adding additional structure to an input query

## Counterfactual & Online Evaluation + LTR (Week 5 / March 10)

### Agenda

* Q&A
* Flipped classroom questions and discussion  
  <https://www.wooclap.com/MOXEZF>

### Which of the following are difficulties for counterfactual and online evaluation and LTR?

* **[Incorrect]** User interactions are noisy
* **[Correct]** User interactions are biased
* **[Incorrect]** Collecting user interactions is expensive
* **[Incorrect]** User interactions do not correlate with relevance
* **[Correct]** A model is required to understand user interactions

### Imagine that clicks are NOT biased. Which statements are true in this situation?

* **[Incorrect]** Counterfactual LTR needs to use IPS
* **[Correct]** Clicks can be directly used as relevance labels
* **[Correct]** Clicks are noisy
* **[Incorrect]** Interleaving cannot be used for online evaluation
* **[Incorrect]** It is not possible to estimate the observance probability P(o\_i=1 | R,d\_i)

### Which use cases are better handled by counterfactual evaluation than by a click model?

* **[Incorrect]** Producing a feature for use with LTR
* **[Correct]** Computing a relevance metric
* **[Incorrect]** Modeling biases other than position bias, such as trust bias
* **[Incorrect]** Any modeling of bias

### What are feasible ways to estimate the observance probability P(o\_i=1 | R,d\_i) for IPS (Inverse Propensity Scoring)?

* **[Incorrect]** Randomize top-n results
* **[Correct]** Assume a click model, estimate it from click logs, and use the estimated examination probability P(E\_i = 1)
* **[Incorrect]** P(o\_i=1 | R,d\_i) = 1 / log(i + 1)
* **[Correct]** Swap a random result with a result at rank i

### What are the advantages of A/B testing over Interleaving?

* **[Incorrect]** Requires less data
* **[Incorrect]** Separates users into A/B groups
* **[Correct]** Can evaluate any aspect of a system
* **[Incorrect]** Does not require human annotators

### You are using Team Draft Interleaving to interleave pairs of rankings. Which of the following interleavings COULD be produced with this approach?

* **[Incorrect]** Rankings: [1, 2, 3, 4] and [2, 4, 3, 1]. Interleaving: [1, 4, 2, 3]
* **[Correct]** Rankings: [1, 2, 3, 4] and [2, 4, 3, 1]. Interleaving: [2, 1, 4, 3]
* **[Correct]** Rankings: [1, 2, 3, 4] and [2, 4, 3, 1]. Interleaving: [2, 1, 3, 4]
* **[Correct]** Rankings: [1, 2, 3, 4] and [2, 4, 3, 1]. Interleaving: [1, 2, 3, 4]

### In which ways does evaluation using interleaving differ from counterfactual evaluation (CE)?

* **[Incorrect]** Interleaving is performed offline, whereas CE is online
* **[Correct]** Interleaving provides a relative preference, whereas CE can be used to calculate a metric
* **[Incorrect]** Interleaving is more efficient than A/B testing, whereas CE is not
* **[Incorrect]** CE requires an intervention, whereas interleaving is purely observational

### After creating and displaying an interleaved result set, how can we choose the ranker?

* **[Incorrect]** By computing our relevance metric
* **[Correct]** By counting the number of clicks for each ranker
* **[Incorrect]** By (position) debiasing the clicks for each ranker and then counting

### Which of the following are potential disadvantages of counterfactual LTR?

* **[Correct]** Requires logged data from a deployed system
* **[Incorrect]** Requires human relevance judgments for training
* **[Correct]** Optimizes a lower bound on some relevance metric
* **[Incorrect]** Requires an additive linearly decomposable relevance metric
* **[Incorrect]** Is not compatible with the nDCG metric
* **[Incorrect]** Requires knowing exact examination probabilities

## Learning to Rank and Interactions (Week 4 / March 3)

### Agenda

* BM25 IDF
* Q&A
* Flipped classroom questions and discussion  
  <https://app.wooclap.com/IFLLLO>

### Why would you use LTR (learning to rank)?

* **[Correct]** To optimize a ranking metric of interest directly
* **[Incorrect/Partial]** To improve semantic matching of queries and documents
* **[Incorrect]** To reduce the time required for ranking (query latency)
* **[Correct]** To synthesize various ranking signals
* **[Correct]** To consider signals that only depend on a query or document

### Which of the following methods are typically viewed as LTR?

* **[Incorrect ]QLM**
* **[Incorrect]** LDA
* **[Incorrect]** LSI
* **[Incorrect]** BERT
* **[Correct]** RankNet
* **[Incorrect]** BM25
* **[Correct]** Regression

### Which of the following quantities should be used as features for LTR?

* **[Correct]** Document length
* **[Correct]** QLM score
* **[Incorrect]** Relevance judgement for query-document
* **[Correct]** Number of incoming links (for web pages)
* **[Incorrect]** Index size
* **[Correct]** Number of clicks on query-document
* **[Correct]** PBM attractiveness parameter for query-document
* **[Correct]** TF-IDF for (query term)-document
* **[Correct]** Number of characters in a query
* **[Incorrect]** Number of documents in a collection

### Which of the following quantities should be used as labels for LTR?

* **[Incorrect]** Document length
* **[Incorrect]** QLM score
* **[Correct]** Relevance judgement for query-document
* **[Incorrect]** Number of incoming links (for web pages)
* **[Incorrect]** Index size
* **[Correct]** Number of clicks on query-document
* **[Correct]** PBM attractiveness parameter for query-document
* **[Incorrect]** TF-IDF for (query term)-document
* **[Incorrect]** Number of characters in a query
* **[Incorrect]** Number of documents in a collection

### Which of the following metrics are the best fit for optimizing with LamdaRank?

* **[Incorrect]** Precision@1
* **[Incorrect]**Recall@1
* **[Incorrect]**Precision@10
* **[Incorrect]**Recall@10
* **[Correct]**AP
* **[Correct]** NDCG
* **[Incorrect]** Accuracy

### What can a click model be used for?

* **[Correct]** To estimate position bias
* **[Incorrect]** To estimate attractiveness rather than relevance
* **[Incorrect]** To better rank all documents by producing features for LTR
* **[Correct]** To improve document ranking for head and torso (non-tail) queries

### You are using a click model to simulate clicks. So far, you simulated no clicks for the first three results. Which probability you need to calculate to simulate a click/no click on the 4th result?

* **[Incorrect]** P(C\_4 = 1)
* **[Incorrect]** P(C\_4 = 1 | C\_3 = 0)
* **[Incorrect]** P(C\_4 = 1 | C\_3 = 0, C\_2 = 0)
* **[Correct]** P(C\_4 = 1 | C\_3 = 0, C\_2 = 0, C\_1 = 0)

### How can P(C\_4 = 1 | C\_3 = 0, C\_2 = 0, C\_1 = 0) be calculated using PBM?

* **[Incorrect]** alpha\_qd4
* **[Correct]** alpha\_qd4 \* gamma\_4
* **[Incorrect]** a\_qd4 \* (1 - a\_qd3) \* (1 - a\_qd2) \* (1 - a\_qd1)

### How can P(C\_4 = 1 | C\_3 = 0, C\_2 = 0, C\_1 = 0) be calculated using CM?

* **[Incorrect]** alpha\_qd4
* **[Incorrect]** alpha\_qd4 \* gamma\_4
* **[Correct]** a\_qd4 \* (1 - a\_qd3) \* (1 - a\_qd2) \* (1 - a\_qd1)

### You used a click model to estimate attractiveness for each query-document pair (alpha\_qd). Which of the following is an LTR approach that can use alpha\_qd as a feature?

* **[Correct]** Point-wise
* **[Correct]** Pair-wise
* **[Correct]** List-wise
* **[Incorrect]** Transformer with cross-entropy loss

### In the following scenarios, which click model is the best fit?

### SCENARIO 1: A businessman looking for the homepage of the New York Times for his morning reading.

### SCENARIO 2: An American basketball fan searching for information and history for the NBA.

* **[Incorrect]** CM in both #1 and #2
* **[Incorrect]** PBM in both #1 and #2
* **[Correct]** CM in #1 and PBM in #2
* **[Incorrect]** CM in #2 and PBM in #1

## Document representation and matching (Week 3 / Feb 24)

### Agenda

* Announcements: 75 person limit lifted
* Q&A
* Flipped classroom questions and discussion  
  <https://www.wooclap.com/UCDULN>

### In IR, we usually rank with unigram language models (LMs). Is it possible to use bi-gram LMs for ranking?

* **[Correct]** Yes, if we use p(t\_i | t\_{i-1}) instead of p(t\_i) in the Query Likelihood Model (QLM)
* **[Incorrect]** Yes, if we use a mixture of two unigram language models: one for t\_i and one for t\_{i-1}
* **[Incorrect]** No, because we assume that all terms in documents/queries are independent of each other
* **[Incorrect]** No, because there is no clear way to incorporate bi-gram LMs into matching functions

### What happens to the BM25 score of a document if all TFs in that document are very large?

* **[Incorrect]** The document gets the highest BM25 score
* **[Incorrect]** The document gets the lowest BM25 score
* **[Incorrect]** The BM25 score simplifies to |q| \* k\_1, where |q| is the query length and k\_1 is a parameter
* [**[Correct]** The BM25 score simplifies to the sum of IDFs](https://www.desmos.com/calculator/1t3c49oi4v)
* **[Incorrect]** In this case, BM25 cannot be computed

LSI and Evaluation

You use two low-rank approximations for LSI:

* LSI 1: k = 1000
* LSI 2: k = 100

Which of the following evaluation metrics are expected to be LOWER for LSI 2 than for LSI 1?

* **[Correct]** Precision@1
* **[Correct]** Recall@1
* **[Correct]** Precision@10
* **[Correct]** Recall@10
* **[Correct]** Total precision
* **[Correct]** Average precision (AP)
* **[Correct]** DCG
* **[Correct]** RBP
* **[Correct]** ERR

### Stemming and Evaluation

Given a collection of text documents, you build two inverted indices:

\* Index A - no text preprocessing.

\* Index B - stemming.

For a precisely crafted query, which of the following evaluation metrics are expected to be LOWER for B than for A?

* **[Correct]** Precision@1
* **[Correct]** Recall@1
* **[Correct]** Precision@10
* **[Correct]** Recall@10
* **[Correct]** Total precision
* **[Incorrect]** Total recall
* **[Correct]** Average precision (AP)
* **[Correct]** NDCG
* **[Correct]** RBP
* **[Correct]** ERR

### Which document ranking functions are expected to perform best for patent search?

* **[Incorrect]** VSM
* **[Incorrect]** QLM
* **[Incorrect]** BM25
* **[Correct]** LSI
* **[Correct]** LDA

### What are the disadvantages of using Average Word Embeddings (AWEs) for ranking?

* **[Incorrect]** AWE only allows exact term matching
* **[Correct]** The AWE of a relevant document may be far from the AWE of a query in the latent space
* **[Incorrect]** It is not feasible to compute the AWE of a query during runtime
* **[Correct]** The AWE of a document may have a different meaning compared to the document itself
* **[Correct]** To use AWE effectively, one needs a lot of training data

### Why would you use re-ranking on top of another ranking?

* **[Incorrect]** The more ranking methods we use on top of each other the better
* **[Incorrect]** One ranking method is usually not enough, we use re-ranking to fix its potential issues
* **[Correct]** Re-ranking uses slower but more effective methods
* **[Incorrect]** Re-ranking introduces personalization
* **[Incorrect]** Through re-ranking, we can control what search results we show to a user

### Which pre-processing strategy would work best when ranking with AWEs (average word embeddings)?

* **[Incorrect]** Only minimal text analysis (tokenization, lowercasing) -- no stopwording or stemming
* **[Correct]** Remove stopwords, do not perform stemming
* **[Incorrect]** Keep stopwords, perform stemming
* **[Incorrect]** Remove stopwords, perform stemming

### Which pre-processing strategy would work best when ranking with a neural transformer-based method like BERT?

* **[Correct]** Only minimal text analysis (tokenization, lowercasing) -- no stopwording or stemming
* **[Incorrect]** Remove stopwords, do not perform stemming
* **[Incorrect]** Keep stopwords, perform stemming
* **[Incorrect]** Remove stopwords, perform stemming

### Additional evaluation scenarios (lawyer, RecSys): see previous week

### OPEN question for later discussion:

We’ve seen AWE as an example of a straightforward way to use word embeddings and, at the other end of the spectrum, we’ve seen the KNRM approach for characterizing the distribution of similarities between query and document terms. Given a similarity matrix as input, what are some other feasible ways to compute a relevance score?

## Evaluation (Week 2 / Feb 17)

### Agenda

* Brief guest presentation on IvI's Inclusive AI program from Maartje ter Hoeve
  + Website (with sign up form): <https://uva-iai.github.io/>
  + E-mail IAI: info-iai@uva.nl
  + Maartje’s email: m.a.terhoeve@uva.nl
* Announcements: assignment released; lectures posted; mic for class
* Q&A: evaluation / text analysis / indexing
* Flipped classroom questions and discussion  
  <https://wooclap.com/XFOJBH>

### Which of the following strategies are feasible to select documents for relevance assessment?

* **[Incorrect]** Random sampling
* **[Incorrect]** Manual selection
* **[Correct]** Depth-k pooling
* **[Partially correct]** Using a classifier

### In depth-k pooling, any document not in the depth​-k pool is assumed to be non-relevant. What are the reasons behind this assumption?

* **[Incorrect]** Search engines do not return relevant documents below position k.
* **[Correct]** We assume that search engines are likely to assign low ranks to non-relevant documents.
* **[Incorrect]** We assume that users are not likely to examine documents below position k.
* **[Incorrect]** Documents below position k are not used during ranking.
* **[Correct]** Given a query, most documents in any collection are non-relevant.

### Pick the best evaluation metrics for the following search tasks:

#### SCENARIO: A businessman searching for the New York Times homepage for his breakfast reading.

* **[Correct]** Precision@1
* **[Correct]** Recall@1
* **[Incorrect]** Precision@10
* **[Incorrect]** Recall@10
* **[Incorrect]** Total precision
* **[Incorrect]** Total recall
* **[Correct]** Average Precision (AP)
* **[Incorrect]** NDCG
* **[Partially correct]** RBP
* **[Incorrect]** ERR

#### SCENARIO: An American basketball fan searching for information and history for the NBA. Some of the returned pages provide a lot of relevant details, for example, team rankings, match scores, the latest news, etc. Some pages are just marginally relevant. Others are less interesting or irrelevant.

* **[Incorrect]** Precision@1
* **[Incorrect]** Recall@1
* **[Incorrect]** Precision@10
* **[Incorrect]** Recall@10
* **[Incorrect]** Total precision
* **[Incorrect]** Total recall
* **[Partially correct]** Average precision (AP)
* **[Correct]** NDCG
* **[Correct]** RBP
* **[Correct]** ERR

#### SCENARIO: A lawyer searching for all evidence relevant to one of her cases. The lawyer is evaluated on whether she wins the case. She bills clients hourly, so she doesn't mind reading through all the documents that are returned.

* **[Incorrect]** Precision@1
* **[Incorrect]** Recall@1
* **[Incorrect]** Precision@10
* **[Incorrect]** Recall@10
* **[Incorrect]** Total precision
* **[Correct]** Total recall
* **[Incorrect]** Average precision (AP)
* **[Incorrect]** NDCG
* **[Incorrect]** RBP

#### **[Incorrect]** ERR

#### SCENARIO: You are developing a recommendation system that will be used to display three side-by-side music playlist recommendations to the user.

* **[Incorrect]** Precision@1
* **[Incorrect]** Recall@1
* **[Correct]** Precision@3
* **[Incorrect**] Recall@3
* **[Incorrect]** Precision@10
* **[Incorrect]** Recall@10
* **[Incorrect]** Total precision
* **[Incorrect]** Total recall
* **[Incorrect]** Average precision (AP)
* **[Incorrect]** NDCG
* **[Incorrect]** RBP

### In ML classification, a widely used evaluation measure is accuracy, i.e., the percentage of correct predictions. Why is accuracy not a good measure for IR?

* **[Incorrect]** In IR, we evaluate rankings of objects, not labels for individual objects.
* **[Incorrect]** As opposed to clearly interpretable classification labels in ML, the notion of relevance in IR is vague.
* **[Incorrect]** We do not know relevance labels for all documents.
* **[Correct]** The number of non-relevant documents is much larger than the number of relevant documents.
* **[Incorrect]** Accuracy can be used for IR evaluation.

### For which of the following tasks is it important to identify all relevant documents (given a user’s query)?

* Medical search
* Web search
* Sponsored search / ad search
* Patent search
* News search

## IR0 recap (Week 1 / Feb 10)

### What are potential problems with stemming?

* [Incorrect] An inverted index becomes larger
* [Incorrect (not a problem)] More relevant documents can be retrieved
* [Partially correct] Too many documents may be retrieved

### What information can be stored in an inverted index without introducing redundancy?

* [Correct] Term frequency
* [Incorrect] document length
* [Correct] BM25 score for term and document (a type of weight)

### Indexing stop words

Given a collection of documents, we build two indexes:  
*A: minimal pipeline (tokenization incl punct removal, lowercasing)*

*B: A plus stopword removal.*

Which of the following statements are true?

* [False] Index A is smaller than Index B
* [Partially true] Index B is smaller than Index A, if both contain document ids
* [True] Index B is no bigger than Index A, if both contain document ids

### Indexing and stemming

Given a collection of documents, we build two indexes:  
*A: minimal pipeline (tokenization incl punct removal, lowercasing)*

*B: A plus stemming*

Which of the following statements are true?

* [False] Index A is smaller than Index B
* [depends on specifics] Index B is smaller than Index A, if only index B contains term positions

### Open-ended: What is Information Retrieval?

### Open-ended: With recommender systems, what is the query?

### From Q&A: How do we decide which words should be stopwords?

### From Q&A: What is the relationship between the various data structures? (inverted index, forward index, web graph, page attribute file)