Week 3 Practice Quiz

Question 1

Suppose a query has a total of 5 relevant documents in a collection of 100 documents. System A and System B have each retrieved 10 documents, and the relevance status of the ranked lists is shown below:

System A: [+ + - - - - - - - -] System B: [- + - - + - - - - +]

where the leftmost entry corresponds to the highest ranked document, and the rightmost entry corresponds to the lowest ranked document. A "+" indicates a relevant document and a "-" corresponds to a non-relevant one. For example, the top ranked document retrieved by System A is relevant, whereas the top ranked document retrieved by B is non-relevant.

What is the **precision at 10 documents** of both systems?

- P(A) = 2/10 P(B) = 3/10
- P(A) = 2/100 P(B) = 3/100
- P(A) = 2/5 P(B) = 3/5
- P(A) = 8/100 P(B) = 7/100

Question 2

Assume the same scenario as in Question 1. What is the **recall** of both systems?

- R(A) = 2/5 R(B) = 3/5
- R(A) = 2/100 R(B) = 3/100
- R(A) = 2/10 R(B) = 3/10
- R(A) = 8/100 R(B) = 7/100

Question 3

Assume the same scenario as in Question 1. What is the **average precision** of both systems?

- AP(A) = 2/5 AP(B) = 6/25
- AP(A) = 2/10 AP(B) = 3/25
- AP(A) = 2/100 AP(B) = 3/250
- AP(A) = 3/10 AP(B) = 9/20

Question 4

Assume you have two retrieval systems X and Y. If X has a higher MAP (mean average precision), can Y have a higher gMAP (geometric mean average precision)?

- Yes
- No

Question 5

If system A has higher precision at k document than system B for any number of k, does it mean A also has higher recall than B at any position?

- Yes
- No

Question 6

F-measure contains a parameter that weighs between precision and recall. For an automatic system that filters tweets in the search for possible communication between terrorists about attack plans on US soil, when evaluating the system's performance with F-measure, should we use a higher parameter or lower?

- Lower
- Higher

Question 7

What is nDCG capable of but not DCG?

- Compare two systems performed on a set of queries
- Compare two systems performed on one single query
- Work with relevance judgment that is multi-level
- Work with relevance judgment that is binary-level (relevant vs. not relevant)

Question 8

Why is pooling useful?

- So that we don't need to judge every document
- So that we don't need humans to do judgment
- So that all documents can be judged efficiently

Question 9

Which of the following is not correct?

- Precision@10docs is easy for users to interpret.
- MAP and nDCG are good for comparing ranking algorithms.
- DCG is better than nDCG as its value is within [0, 1].

Question 10

If in PR (precision, recall) curves, curve A is above B for all recall, what can you say?

- A is better than B.
- B is better than A.
- There is no clear conclusion about A vs. B.
- A is as good as B.

Week 3 Quiz

Question 1

Suppose a query has a total of 4 relevant documents in the collection. System A and System B have each retrieved 10 documents, and the relevance status of the ranked lists is shown below:

System A: [- + - - - - - - - -] System B: [+ + - - - - - - - -]

where the leftmost entry corresponds to the highest ranked document, and the rightmost entry corresponds to the lowest ranked document. A "+" indicates a relevant document and a "-" corresponds to a non-relevant one. For example, the top ranked document retrieved by System A is non-relevant, whereas the top ranked document retrieved by B is relevant.

What is the **precision at 10 documents** of both systems?

- P(A) = 1/10 P(B)= 2/10
- P(A) = 1/4 P(B)= 2/4
- P(A) = 1/40 P(B)= 2/40
- P(A) = 9/10 P(B)= 8/10

Question 2

Assume the same scenario as in Question 1. What is the **recall** of both systems?

- R(A) = 1/4 R(B)= 2/4
- R(A) = 1/10 R(B)= 2/10
- R(A) = 1/40 R(B)= 2/40
- R(A) = 9/10 R(B)= 8/10

Question 3

Assume the same scenario as in Question 1. What is the **average precision** of both systems?

- AP(A) = 1/8 AP(B) = 1/2
- AP(A) = 1/20 AP(B) = 1/5
- AP(A) = 7/20 AP(B) = 7/10
- AP(A) = 1/10 AP(B) = 1/5

Question 4

Assume you have two retrieval systems X and Y. For a specific query, system X has a higher precision at 10 documents compared to Y. Can system Y have a higher **average precision** on the same query?

- Yes
- No

Question 5

Can a retrieval system have an F1 score of 0.75 and a precision of 0.5?

- No
- Yes

Question 6

For any ranked list of search results, precision at 10 documents is **always** higher than precision at 20 documents.

- False
- True

Question 7

What can you say about the precision-recall (PR) curve?

- It is always monotonically decreasing.
- It is always monotonically increasing.
- The ideal system should have the PR curve as a horizontal line.

Question 8

Which is correct about average precision?

- It combines precision and recall.
- It does not show the difference between ranks of relevant documents.

Question 9

Which of the following is NOT true about Cranfield evaluation methodology?

- It simulates real document collections.
- It simulates user queries.
- It does not involve humans to make relevance judgments.

Question 10

Which of following is wrong about nDCG@k?

- It has a range between 0 and 1.
- It discounts only top ranked documents.
- It can be used to compare across queries.

Week 4 Practice Quiz

Question 1

You are given a vocabulary composed of only three words: “text,” “mining," and “research.” Below are the probabilities of two of these three words given by a unigram language model:

Word	Probability
text	0.4
mining	0.2

What is the probability of generating the phrase “text mining research” using this unigram language model?

- 0.032
- 0.08
- 0
- 0.4

Question 2

You are given the query Q= “food safety” and two documents:

D1 = “food quality regulations”

D2 = “food safety measures”

Assume you are using the maximum likelihood estimator **without** smoothing to calculate the probabilities of words in documents (i.e., the estimated $p(w|D)$ is the relative frequency of word w in the document D). Based on the unigram query likelihood model, which of the following choices is correct?

- P(Q|D1) = 0 P(Q|D2) = 1/9
- P(Q|D1) = 1/3 P(Q|D2) = 1/9
- P(Q|D1) = 1/3 P(Q|D2) = 0
- P(Q|D1) = 1/2 P(Q|D2) = 1/2

Question 3

Probability smoothing avoids assigning zero probabilities to unseen words in documents.

- True
- False

Question 4

Assume you are given two scoring functions:

$S1(Q,D)=P(Q|D)$

$S2(Q,D)=\log P(Q|D)$

For the same query and corpus, $S1$ and $S2$ will give the same ranked list of documents.

- True
- False

Question 5

Assume you are using linear interpolation (Jelinek-Mercer) smoothing to estimate the probabilities of words in a certain document. What happens to the smoothed probability of the word when the parameter λ is **decreased**?

- It becomes closer to the maximum likelihood estimate of the probability derived from the document.
- It becomes closer to the probability of the word in the collection language model.
- It does not change.

Question 6

In the query likelihood model, why is smoothing necessary?

- Without smoothing, if a term in the query is not in the document, then the likelihood becomes $-\infty$.
- Without smoothing, if a term in the document is not in the query, then the likelihood becomes $-\infty$.
- Without smoothing, if a term is neither in the query nor the document, then the likelihood becomes $-\infty$.

Question 7

Which of the following is NOT correct about the unigram model?

- The probability of generating the word A OR B is the sum of the probability of generating A and the probability of generating B.
- The probability of generating the words A AND B is the product of the probability of generating A and the probability of generating B.
- The probability of generating the word sequence "A" "B" "C" is the same as generating "C" "B" "A."

Week 4 Quiz

Question 1

Assume you are using a unigram language model to calculate the probabilities of phrases. Then, the probabilities of generating the phrases “study text mining” and “text mining study” are **not** equal, i.e., P(“study text mining”) ≠ P(“text mining study”).

- False
- True

Question 2

You are given a vocabulary composed of only four words: “the,” “computer,” “science,” and “technology.” Below are the probabilities of three of these four words given by a unigram language model.

Word	**Probability**
the	0.4
computer	0.2
science	0.3

What is the probability of generating the phrase “the technology” using this unigram language model?

- 0.04
- 0.5
- 0.0024
- 0.1

Question 3

You are given the query Q= “online courses” and two documents:

D1 = “online courses search engine”

D2 = “online education is affordable”

Assume you are using the maximum likelihood estimator **without** smoothing to calculate the probabilities of words in documents (i.e., the estimated $p(w|D)$ is the relative frequency of the word w in the document D). Based on the unigram query likelihood model, which of the following choices is correct?

- P(Q|D1) = 1/16 P(Q|D2) = 0
- P(Q|D1) = 0 P(Q|D2) = 1/4
- P(Q|D1) = 1/16 P(Q|D2) = 1/4
- P(Q|D1) = 1/2 P(Q|D2) = 1/2

Question 4

Assume the same scenario as in Question 3, but using linear interpolation (Jelinek-Mercer) smoothing with $\lambda=0.5$. Furthermore, you are given the following probabilities of **some** of the words in the collection language model:

Word	**P(w|C)**
online	1/4
courses	1/4
education	1/8

Based on the unigram query likelihood model, which of the following choices is correct?

- P(Q|D1) = 1/16 P(Q|D2) = 1/32
- P(Q|D1) = 1/16 P(Q|D2) = 0
- P(Q|D1) = 1/16 P(Q|D2) = 1/16
- P(Q|D1) = 1/32 P(Q|D2) = 1/32

Question 5

If word count for every term doubles in one document:

- $p(w|d)$ remains the same if using Jelinek-Mercer smoothing.
- $p(w|d)$ remains the same if using Dirichlet-prior smoothing.
- If not using any smoothing, query likelihood would change for some queries.

Question 6

Assume you are using Dirichlet Prior smoothing to estimate the probabilities of words in a certain document. What happens to the smoothed probability of the word when the parameter μ is **increased**?

- It becomes closer to the probability of the word in the collection language model.
- It becomes closer to the maximum likelihood estimate of the probability derived from the document.
- It does not change.
- It tends to 1.

Week 5 Practice Quiz

Question 1

In PageRank, what is NOT the benefit of introducing random jumping?

- Otherwise PageRank will favor nodes with fewer incoming links
- Otherwise disconnected page always has zero probability
- Otherwise for zero-outlink nodes will receive all the probability

Question 2

Modern web search engines often combine many features (e.g., content-based scores, link-based scores) to rank documents.

- True
- False

Question 3

PageRank only uses the inter-document links when calculating a document's score, without considering the content of the document.

- True
- False

Question 4

Can a crawler that only follows hyperlinks identify hidden pages that do not have any incoming links?

- No
- Yes

Question 5

Which one is NOT a feature of incremental crawling?

- Can learn from past experience (updated daily vs. monthly)
- Target at frequently updated pages
- Target at frequently accessed pages
- Target at a subset of pages (e.g., all pages about "automobiles")

Question 6

Which of the following sites will get a larger value in PageRank?

- A site that outlinks to many others but none others link to it
- A site that has many others linking to it but it does not link to others

Question 7

Which of the following is NOT a reason that PageRank is efficient?

- Matrix M is usually sparse
- Dimension of matrix M is usually large
- Each iteration of PageRank is simple (matrix multiplication)

Question 8

In PageRank, one needs to estimate the transition probability between pages. Which one of the following statement is true:

- The probability of jumping to page A from page B is the same as that from page B to page A
- The sum of probabilities of jumping from one page to all other pages is one
- The number of parameters in the transition matrix is smaller than the degree of freedom of the transition matrix

Question 9

In the original PageRank algorithm, what kind of pages in the following situation will get most probability

- The page that only has incoming links
- The page that has equal number of incoming and outgoing links

Question 10

Which of following feedback does not involve humans?

- Relevance
- Pseudo
- Implicit

Week 5 Quiz

Question 1

Which of the following is **not** true about GFS?

- The file data transfer happens directly between the GFS client and the GFS master.
- The file data transfer happens directly between the GFS client and the GFS chunkservers.
- The GFS keeps multiple replicas of the same file chunk.

Question 2

MapReduce allows parallelizing the creation of the inverted index.

- True
- False

Question 3

In MapReduce, the Reduce function is called for each unique key of the output key-value pairs from the Map function.

- True
- False

Question 4

Which of the following would cause a web page P to have a higher PageRank score?

- Add to another page Q a link that points to page P
- Add to page P a link that points to another page Q

Question 5

Imagine if the web is fully connected with N pages such that for any pair of pages, P and Q, there exists a link from P to Q, then which of the following is true?

- All the pages will have a PageRank score of 1/N.
- At least one page will have a PageRank score larger than 1/N.
- At least one page will have a PageRank score smaller than 1/N.

Question 6

Which of the following is/are the difference(s) between pseudo and implicit feedback?

- Pseudo feedback assumes top ranked documents are relevant.
- Implicit feedback assumes user-clicked documents are relevant.
- Both methods do not involve user activity.

Question 7

What is true about Rocchio feedback?

- Negative examples are as important as positive examples.
- All words are used (no truncation) so that performance and efficiency can be guaranteed.
- It works the best if you discard original query weights during feedback.
- It can be used for relevance feedback and also pseudo feedback.

Question 8

The advantages of incremental crawling are:

- It targets frequently updated pages.
- It targets frequently accessed pages.
- It learns from past experience.
- It adds overhead.

Question 9

What is NOT the reason that PageRank works better than "citation counting"?

- PageRank considers "indirect citations."
- PageRank has smoothing.
- It utilizes text information other than link information.

Question 10

Which of the following is NOT true about the PageRank algorithm?

- The transition matrix is symmetric.
- The values of the elements in the transition matrix range from [0, 1]
- It's an iterative algorithm.

Week 6 Practice Quiz

Question 1

In collaborative filtering, to measure user similarity using cosine, which one of the following is correct?

- It will not be biased by the user activity (less/more activity of users).
- It solves problem of missing values.

Question 2

Which of the following tasks can be solved as a classification problem?

- Ranking
- Recommendation
- Spamming filtering

Question 3

In recommendation/filtering, which heuristic in the following is NOT correct?

- If a user likes one item, then he/she will dislike other items not similar to the one liked.
- Similar users will like the same items.
- The same user will like similar items.

Question 4

Comparing the logistic regression learning-to-rank method vs. the query likelihood model, which of the following is NOT true?

- The logistic regression method can take the query likelihood output as features.
- The query likelihood method does not require time for training.
- If trained until converged, the logistic regression will output the same solution (probability) as the query likelihood model as they are using the same relevance judgement.

Question 5

What is the advantage of Learning to Rank over BM25?

- Learning to Rank can combine many more features than BM25.
- BM25 is much slower to train than Learning to Rank.

Question 6

Can the regression based approach Learning to Rank utilize multi-grade relevance judgement (for example, not very relevant, not relevant, mediocre, relevant, very relevant)?

- Yes
- No

Question 7

Which is/are the reason(s) learning-based algorithms became popular in text retrieval?

- More features are available now for determining the results.
- More methods are available now for combining.
- There are more information needs such as removing spamming or diversified ranking.

Question 8

When a new user comes, which of the following will NOT help for recommendation?

- Ask user to first select a few items that he likes
- Ask user to provide a short description of himself
- Recommend user with random selected items

Week 6 Quiz

Question 1

Information filtering systems are more suitable to help users satisfy long-term information needs than short-term ad hoc information needs.

- True
- False

Question 2

In content-based filtering, an item is recommended to a user based on whether other “similar” users like the item or not.

- False
- True

Question 3

In recommendation systems, one uses Beta-Gamma threshold learning for trade-off between exploration and exploitation: $\theta = \alpha * \theta zero + (1 - \alpha) * \theta optimal$. Which of the following is true?

- α should be smaller for new users
- α should be larger for new users
- α should the same for all users no matter if they are new

Question 4

Content-based filtering and collaborative filtering can be combined in a recommender system.

- True
- False

Question 5

Recommendation is one type of Pull mode of information access.

- True
- False

Question 6

In Netflix, if a user has watched a lot of thriller movies, then it recommends "Inception" and "The Silence of the Lambs" to the user, what is this an example of?

- This is content-based filtering.
- This is collaborative filtering.

Question 7

In Spotify, if a user has indicated himself/herself as youth, then Spotify recommends songs that are most listened by users under 20 years old. What is this an example of?

- This is content-based filtering.
- This is collaborative filtering.

Question 8

When adding social network information into recommendation systems, such as friends' info and friends' liked items, this can be used to help:

- Content-based filtering
- Collaborative filtering

Question 9

Which of the following scenario is not suitable for collaborative filtering?

- Where there is only few existing users
- Where there is only few items (music, movies, etc) offering