Início	sexta, 20 de novembro de 2020 às 14:11
Estado	Prova submetida
Data de submissão:	sexta, 20 de novembro de 2020 às 15:54
Tempo gasto	1 hora 42 minutos
Nota	**186,67** de um máximo de 200,00 (**93**%)

Informação   ⚑ Destacar pergunta

Information Retrieval Systems and Tasks

Pergunta 1   Incorreta   Pontuou -1,7 de 5,0   ⚑ Destacar pergunta

In information retrieval, it is common to identify tasks to be accomplished by an automatic system. Consider the description of the following information processing problems and select one that can be solved by an IR system.

Selecione uma opção de resposta:

○ *Catalog search*. Use a web form to search the catalog of products from a company, in the company information system.   ✗

○ [No answer.]

○ *Enterprise information*. Perform keyword search on the documents generated internally by an organisation.

○ *Interaction with an automatic answering system*. Find the answer to a user query following the branches of a multiway tree used by an automatic answering system.

○ *Personnel Information*. Export, from an institutional database, personnel records.

Your answer is incorrect.

Pergunta 2

An IR system is composed of several modules contributing to data processing, retrieval, result ranking, and user interface. From the following functions, select the one which has NOT a central role in the modules of an information retrieval system.

Selecione uma opção de resposta:

○ [No answer.]

○ Building inverted indexes, based on terms extracted from documents.

○ Text processing, identifying words and other lexical elements, to build a vocabulary.

⦿ Classifying documents, assigning them categories according to the domain of the documents in the collection.   ✔

○ Calculating scores for documents in an answer, using the query, the index and user context.

Your answer is correct.

Pergunta 3

IR tasks capture the different collection types, interaction modes and expected results from a retrieval system. Which of the following is NOT an IR task?

Selecione uma opção de resposta:

○ Medical Records: system that explores methods for searching unstructured information found in patient medical records.

○ [No answer.]

○ Complex Answers: system that is capable of answering complex information needs by collating relevant information from an entire corpus.

○ Real-Time Summarization: system that constructs real-time update summaries from social media streams in response to users' information needs.

⦿ Time-bound Query Efficiency: system that queries a database and gets the list of documents created during a range of dates in less than a prescribed time.   ✔

Your answer is correct.

Pergunta 4

Relevance is a central concept in IR. Which of the following definitions captures its meaning?

Selecione uma opção de resposta:

- ⦿ A document is relevant if it satisfies the information need that started the search process.   ✔

- ○ A document is relevant if it has a large PageRank value.

- ○ A document is relevant if it is included in the top 10 results in the answer of an IR system.

- ○ [No answer.]

- ○ A document is relevant if the user picks it from the results page and reads it.

Your answer is correct.

Informação

Indexing and Searching

Pergunta 5

In an Information Retrieval search tool, the concept of document is central. Select one of the following statements as true for a document in an IR system.

Selecione uma opção de resposta:

- ○ a. A document is a resource (text or otherwise) which is freely available on the Web.

- ○ b. A document is a resource (text or otherwise) that has links to and from other resources.

- ⦿ c. A document is a resource (text or otherwise) or set of resources that are indexed as a single unit in the IR system.   ✔

- ○ d. A document is a resource (text or otherwise) obtained from a reliable source.

- ○ e. [No answer.]

Your answer is correct.

In Information Retrieval, it is common to say that documents are "bags of words". Select one of the following explanations for the use of this metaphor.

Selecione uma opção de resposta:

- ○ a. [No answer.]

- ○ b. Common information retrieval models do not take into account the structure of phrases in a document.

- ○ c. In information retrieval systems the meaning of words is lost and only their position in the documents is important.

- ⦿ d. Information retrieval systems do not deal with the actual documents from the collection but with terms that may come from different documents. ✘

- ○ e. Indexes built for text collections order the terms in their vocabularies in alphabetical order.

Your answer is incorrect.

What is an inverted index?

Selecione uma opção de resposta:

- ○ An inverted index is a table with all the terms that occur in all the documents of the collection.

- ○ An inverted index is a matrix of terms versus documents where a cell value equals the weight of the term in the document if the term occurs in the document and 0 otherwise.

- ⦿ An inverted index is a data structure linking each vocabulary term to a list of occurrences of the term in the documents. ✔

- ○ An inverted index is a matrix of terms versus documents where a cell has value 1 if the term occurs in the document and 0 otherwise.

- ○ [No answer.]

Your answer is correct.

Pergunta 8

In document indexing, the different word forms occurring in documents may appear in the vocabulary as different terms, or they may be pre-processed to normalized terms. For instance, we can remove verbal forms and word variations for gender or number.

How do we expect word pre-processing to affect the search results?

Selecione uma opção de resposta:

○ If term pre-processing is used, answers have equal or better precision than if it is absent (i.e., the number of non-relevant documents in the answer is the same or lower).

○ [No answer.]

○ Term pre-processing reduces the number of non-relevant documents in the answer.

○ Term pre-processing has no effect on the quality of the answer, but makes the system more efficient.

◉ New documents may appear in the answer as a result of word pre-processing.   ✔

Your answer is correct.

Pergunta 9

Term weighting in the vector model uses the *tf* measure. What does a *tf* value represent?

Selecione uma opção de resposta:

○ The number of times the term occurs in the collection.

○ 1 or 0, depending on whether the term occurs in the document or not.

◉ The number of times the term occurs in a document.   ✔

○ [No answer.]

○ The number of documents where the term occurs.

Your answer is correct.

Pergunta 10

Term weighting in the vector model uses the *idf* measure. What does the *idf* value for a term represent?

Selecione uma opção de resposta:

○ [No answer.]

○ The percentage of documents in the collection which do not include the term.

◉ The rarity of the term in the collection.   ✔

○ The maximum frequency of the term, considering each document in the collection.

○ The frequency of the term in the collection.

Your answer is correct.

Pergunta 11

The *idf* measure is a common component of term weighting schemes. What is the *idf* of a term that occurs in every document?

Selecione uma opção de resposta:

○ This value depends on the number of documents in the collection.

○ 1

◉ 0   ✔

○ infinity

○ [No answer.]

Your answer is correct.

Pergunta 12

In the vector model, documents and queries are represented as vectors in an n-dimensional space. What is the dimension of the space?

Selecione uma opção de resposta:

○ The number of documents in the collection.

○ [No answer.]

○ The number of index terms times the number of documents in the collection.

◉ The number of index terms.   ✔

○ The number of different words extracted from the documents in the collection.

Your answer is correct.

Pergunta 13

In the vector model, documents are ranked based on a score that measures the similarity between the document and the query. How is the similarity score computed?

Selecione uma opção de resposta:

○ The similarity measure is the number of terms in common between the document and the query.

○ [No answer.]

◉ The similarity measure is the cosine of the angle formed by the document and the query vectors.   ✔

○ The similarity measure is the ratio between the Euclidean norms for the query and the document.

○ The similarity measure is the difference between the Euclidean norms of the document and the query vectors.

Your answer is correct.
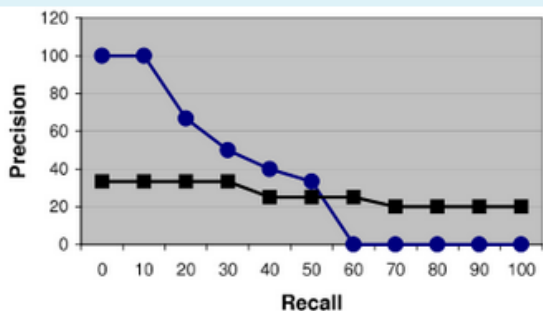
The graphic above has 2 precision versus recall curves. They are the result of an experiment where 2 retrieval systems (A and B) were compared using a set of information needs and a document collection. A is the system corresponding to the blue dots in the graphic and B the one with the black squares.

Based on the information needs, queries were submitted to each system. For each information need, the set of relevant documents in the collection had been previously identified.

The ranked results list for each query in each system was analysed and the positions of the relevant documents were determined.

For each information need and each system, the precision at the standard recall levels (0%, 10%, 20%, ..., 100%) was calculated.

The points in the curves are average precision values, at the standard recall levels, for A and B.

Pergunta 14   Correta   Pontuou 5,0 de 5,0   ⚐ Retirar destaque

A user is choosing between systems A and B, for a task that requires some relevant documents at the top of the results list. Which system would you recommend?

Selecione uma opção de resposta:

○ I recommend system B, because (on average) precision has little variation with the recall level.

○ [No answer.]

○ I recommend system B, because it can guarantee a relevant result at the top of the results list.

○ I recommend system A, because it has a better average ordering of results in the answer.

◉ I recommend system A, because it exhibits 100% precision at the first recall level.   ✔

Your answer is correct.

Pergunta 15

Consider a user who is comparing systems A and B for information needs that require all relevant documents in the collection to be retrieved. Which of the following statements is true?

Selecione uma opção de resposta:

○ The user is expected to find all relevant documents in both systems; with system A the total number of documents to be analysed will be smaller.

○ [No answer.]

○ The user is not expected to retrieve all relevant documents in any of the systems.

○ The user can select any of the systems and will find all relevant documents; we cannot tell which of the systems will make the task easier.

◉ With system A the user is not expected to find all relevant documents.   ✔

Your answer is correct.

Pergunta 16

The *recall* ratio is used in the evaluation of information retrieval systems.

How do we define recall for an information need IN and a results list?

Selecione uma opção de resposta:

○ [No answer.]

○ Recall is the percentage of the documents relevant for IN in the collection which are present in the top 10 results in the results list.

○ Recall is the percentage of documents in the results list which are relevant for IN.

○ Recall is the number of documents relevant for IN in the collection which are present in the top 10 results in the results list.

◉ Recall is the percentage of the documents relevant for IN in the collection which are present in the results list.   ✔

Your answer is correct.

The *precision* ratio is used in the evaluation of information retrieval systems.

How do we define precision for an information need IN and a results list?

Selecione uma opção de resposta:

◉ Precision is the percentage of documents in the results list which are relevant for IN. ✔

○ Precision is the percentage of the documents relevant for IN in the collection which are present in the results list.

○ Precision is the percentage of the documents relevant for IN in the collection which are present in the top 10 results in the results list.

○ [No answer.]

○ Precision is the number of documents relevant for IN in the collection which are present in the top 10 results in the results list.

Your answer is correct.

The *mean average precision* (MAP) metric is used in the evaluation of information retrieval systems; it is based on the ranked lists of results for a set of information needs.

Which of the following is true for MAP?

Selecione uma opção de resposta:

◉ MAP takes into account, for each information need, the precision calculated at each new relevant result. ✔

○ MAP takes into account, for each information need, the precision calculated at each standard recall level.

○ [No answer.]

○ MAP is an mean precision for a set of information needs, and for each information need it averages the precision at the standard recall levels.

○ MAP is the average Precision-Recall graph, based on the P-R graphs at the standard recall levels for each information need.

Your answer is correct.

In the evaluation of a retrieval system based on a document collection C, we have examined the collection and know that the number of relevant documents for information need IN1 is 20.

Consider the following representation for a ranked results list, where N stands for a non-relevant document and R stands for a relevant one. The leftmost document is the top of the list.

N N R N N R R N N R N N N N R N N N N N R N N N

Pergunta 19   Respondida   Pontuou 10,0 de 10,0   ⚑ Destacar pergunta

Calculate pairs of recall-precision values for the list, considering the sublists up to each relevant document (6 pairs). Represent values as percentages.

There are 20 relevant docs, thus the recall value will "change" 20 times, with an incremental step of $1/20 = 5\%$.

Precision(recall = 5%) = 1/3 = 33.33%
Precision(recall = 10%) = 2/6 = 33.33%
Precision(recall = 15%) = 3/7 = 42.86%
Precision(recall = 20%) = 4/10 = 40.00%
Precision(recall = 25%) = 5/15 = 33.33%
Precision(recall = 30%) = 6/22 = 27.27%

Comentário:

Pergunta 20   Respondida   Pontuou 10,0 de 10,0   ⚑ Destacar pergunta

Calculate the average precision for this results list, as defined for MAP.

MaP = (33.33% + 33.33% + 42.86% + 40.00% + 33.33% + 27.27%)/6 = 35.02%

Comentário:

Pergunta 21

Information needs on the web have been classified in three categories: "informational", "transactional", and "navigational". According to this distinction, select the FALSE statement from the list.

Selecione uma opção de resposta:

○ [No answer.]

◉ When responding to an "informational" search the retrieval tool does not use the authority information of the candidate results.   ✔

○ The results presented by search tools in response to popular "navigational" queries are pre-computed.

○ A web user evaluates differently a search result depending on the "informational", "transactional", and "navigational" nature of the information need.

○ Detecting "transactional" information needs is specially important for the comercial activity of a search engine.

Your answer is correct.

Pergunta 22

Web search and search in institutional collections ("enterprise search") use similar tools but also exhibit substantial differences. In the following list of differences, select the one which has stronger implications in the configuration of the search tool.

Selecione uma opção de resposta:

◉ A controlled collection versus a non-controlled one.   ✔

○ [No answer.]

○ Long documents versus short documents.

○ Documents in the same language versus documents in multiple languages.

○ Specialized users versus lay users.

Your answer is correct.

Pergunta 23

Multimedia resources (images, video) are now regularly included in generalist search, in what has come to be called "universal search". Including images in the search engine index requires their association to search terms. Which of the following strategies is NOT used to include multimedia in a document index?

Selecione uma opção de resposta:

○ The text content of the document which includes the multimedia resource is used for describing it.

○ The anchor text in links to the multimedia resource is associated to it.

◉ Multimedia resources are manually described using crowdsourcing.   ✔

○ The tags assigned to a multimedia resource are associated to it.

○ [No answer.]

Your answer is correct.

Pergunta 24

The anchor text for a hyperlink is used by search engines. Select a reason for this text to be especially interesting when indexing web pages.

Selecione uma opção de resposta:

○ The existence of the hyperlink allows the use of terms in the target page to index the source page.

○ [No answer.]

○ Terms in the anchor text are carefully chosen by page authors.

◉ Text in the anchor provides a description of the target page by the author of another page.   ✔

○ Pages with many links have few text of their own and so anchor text is used instead.

Your answer is correct.

Pergunta 25

Ranking functions in web search engines take into account a large number of features (signals) obtained from documents, their context, and the user. Select from the following features one that is NOT possible to use for including in a ranking function.

Selecione uma opção de resposta:

- ○ a. Data from the user (e.g. web browser version, location, time of day, etc).

- ○ b. [No answer.]

- ○ c. Date of the page.

- ⊙ d. Total number of visits to the page (e.g. direct visits, visits from other search engines, etc).   ✔

- ○ e. Number of in-links (links to the page) and out-links (links from the page).

Your answer is correct.

Pergunta 26

Several studies of the web have shown an overall "bowtie" structure with large components that have been named SCC (strongly connected component, the "knot"), IN and OUT (the "lobes"), as well as several additional smaller sets of nodes. The following relation holds between the SCC, IN and OUT components:

Selecione uma opção de resposta:

- ○ a. Following page links, one can get from any page in SCC to any page in IN.

- ⊙ b. Following page links, one can get from any page in SCC to any page in OUT.   ✔

- ○ c. [No answer.]

- ○ d. Following page links, one can get from any page in OUT to any page in SCC.

- ○ e. Following page links, one can get from any page in IN to any page in OUT.

Your answer is correct.

Pergunta 27

In web retrieval, the quality of documents is quite diverse, and therefore content analysis alone is not enough to provide a good ranking of search engine results. What is the general approach to solve this?

Selecione uma opção de resposta:

○ Document content is used to select results which are then ranked according to a commercial model.

◉ The score for a document is calculated as a combination of signals, including content-based metrics, the position of query words in documents, and the domain where the documents are from.   ✔

○ [No answer.]

○ Content-based scores are replaced by PageRank.

○ Documents are ranked by popularity, as measured by click-through.

Your answer is correct.

Informação   ⚐ Destacar pergunta

Link Analysis

Pergunta 28

The use of link analysis in web information retrieval has its origin in the study of the impact of scientific publications based on their citation record. In the following, mark the statement that captures an assumption in web link analysis.

Selecione uma opção de resposta:

◉ A link from page P1 to page P2 (an in-link in page P2) represents a vote cast by page P1 on page P2.   ✔

○ A link from page P1 to page P2 (an in-link in page P2) represents a vote cast by page P2 on page P1.

○ A link from page P1 to page P2 means that the authors of P1 agree with the content of P2.

○ [No answer.]

○ If there is a link between page P1 and page P2, their scores in the web graph are similar.

Your answer is correct.
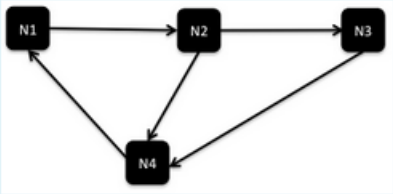
The *PageRank* algorithm uses links between web pages to generate a measure of importance for each page in a web graph.

In the graph above, you are calculating the PageRank for page N4, with no damping factor. Select the correct formula.

Selecione uma opção de resposta:

○ PR(N4) = PR(N1)

○ PR(N4) = PR(N2) + PR(N3)

○ PR(N4) = PR(N2) + PR(N3) - PR(N1)

○ [No answer.]

⊙ PR(N4) = 1/2 PR(N2) + PR(N3)   ✔

Your answer is correct.

Using the same graph, we now calculate the *authority* values according to the HITS algorithm. Start with initial values of 1 for *authority* and 1 for *hub* at each node. What are the *authority* values after 1 iteration?

Selecione uma opção de resposta:

○ a(N1)= 1; a(N2)=2; a(N3)= 1; a(N4)=1;

○ a(N1)= 1; a(N2)=1; a(N3)= 1; a(N4)=1;

⊙ a(N1)= 1; a(N2)=1; a(N3)= 1; a(N4)=2;   ✔

○ a(N1)= 2; a(N2)=3; a(N3)= 2; a(N4)=3;

○ [No answer.]

Your answer is correct.

Pergunta 31

Using the same graph, we now calculate the *hub* values according to the HITS algorithm. Start with initial values of 1 for *authority* and 1 for *hub* at each node. What are the *hub* values after 1 iteration?

Selecione uma opção de resposta:

- ○ [No answer.]
- ○ h(N1)= 2; h(N2)=3; h(N3)= 2; h(N4)=3;
- ○ h(N1)= 1; h(N2)=1; h(N3)= 1; h(N4)=2;
- ○ h(N1)= 1; h(N2)=1; h(N3)= 1; h(N4)=1;
- ⦿ h(N1)= 1; h(N2)=2; h(N3)= 1; h(N4)=1;   ✔

Your answer is correct.

Informação

Course Project

Pergunta 32

In your project, you selected datasets and considered search functionalities for the data.

What is your dataset(s)? Give a name and a very short description of its/their contents.

My project features 3 datasets:

- A books dataset, that was downloaded from the goodreads10k github repository, which was scrapped from the Goodreads website pages; It features content information from books, such as their name, goodreads id, rating, date, etc

- A book reviews dataset, that was downloaded from a website that extracted book reviews from the Goodreads website; If features the id of the book the review is reviewing (goodreads book id), the review content, publish date, rating, etc

- An authors dataset, which was built by doing a set of API requests to the WikiData website, based on the unique author names extracted from the aforementioned book dataset. It features the name of the author, their birth date, gender, nationality, etc

Pergunta 33

What is a document in your dataset? Give a short description of a document and of the relations between documents in the dataset.

There are 3 different types of documents in the system (that consist of retrievable units of the system).
Books, which feature a name, publish date, rating, written language, ...
Authors, which feature a name, birth year, gender, nationality, ...
Book Reviews, which feature a content, publish date, rating, ...

A **book** is associated to *multiple* (1 to many) **book reviews** (the book reviews consist of a textually expressed opinion of a Goodreads user of the target book).
A **book** is written by *one or multiple authors* (1 to many) **authors** (the book may be written by a single author or by multiple authors).

A user in the system may have information needs regarding either specific books, specific authors or specific reviews. Thus, all of these 3 entities should be regarded as retrievable documents in the context of the system.

There are other "abstract entities" that are not regarded as documents, such as book sagas or book genres.

Comentário:

Pergunta 34

Describe a scenario where a user is interacting with your tool and performing a retrieval task. What is the user profile you have in mind? How does it affect the decisions concerning relevance?

The user profile that my group had in mind when conceiving the system was an individual that is used to reading books and with interest in finding interesting information about their favorite books, as well as facts from their authors (such as their nationality and birth date), while at the same time being able to find out about other people's opinion regarding the books they are searching (book reviews).

A possible retrieval task would be a user finding out about a fantasy child-friendly book for their child. They would search about something as "Fantasy child friendly book -mature -violence". They would be interested in finding out about books that are suited for children and that do not feature mature or violent content, on the fantasy-related content. The presented results should be books that respect those restrictions, as well as reviews that confirm that the book is child-friendly. Moreover, it should present information about other books that the author has written (to check if they have written other child-friendly / non-child-friendly books).

Finally, it is worth mentioning that the user's profile should have an impact on the relevance of the results. For example, in the aforementioned retrieval task example, books that are considered as child friendly should be ranked above books that are not child friendly. However, it may also be of interest to include results of reviews that indicate that a given book may not be child friendly, to warn the user about the fact that the target book is not of interest.

Comentário:

The current paradigm for search in heterogeneous collections, using full-text indexes on document contents, is a good solution for a large number of situations where people interact with information.

The other well-established paradigm is database search, where the user knows the data model and queries the database with a formal language.

Consider the case of a relational database with all the information for a company. You want to set up an information retrieval tool to explore the same data.

Outline an approach to design a prototype for this tool:

(1) Suggest a search technology.

(2) A search tool indexes documents; where are the documents in this case?

(3) Your search tool provides list of results; What does a line in the list look like?

(4) Can you connect a search result to the information in the original database?

An IR (Information Retrieval) systems features 3 modules: a (a) crawling module, which crawls through a set of documents to gather information (in this case, the tuples of the tables of the database), a (b) indexing module, that based on the collection documents will index them and (c) a ranking and retrieval module, that will, based on a user's information need (expressed as a query) will evaluate and rank possible results and retrieve them to the user. The user interface for interacting with the IR system may also be considered a module.

Since the system's existing data is stored in a structured format (in a relational database), most of the information should already be normalized (white-space trimmed, in the same format (e.g. utf-16), having null fields for missing data instead of empty strings, etc). Thus, it should be easy to use a technology such as **Solr** to build a schema for the documents of the system's collection - by importing a dataset using Solr, a possible schema is automatically assigned (which can and should be later refined). **Solr** also allows querying the information and specifying how different fields should be indexed, so it wou

The documents, originally stored in the databased, are now stored in Solr (after pre-proceded and indexed).

For example, a document of the system could be an Employee's personal data (featuring their name, address, position, birth date, id, etc). When searching for an employee (let's say someone searches for a "computer security expert" within the company), a results list would be returned. Each line would feature a result, featuring some comprised information of a security expert employee - once clicked on, that results line could redirect the user to the entire document, featuring all the information of the Employee document.

The search result could easily be connected to the information in the original database (e.g. using the employer ID that would be featured in the document) - this is important to ensure that information is both traceable and not lost with the usage of the IR system.

Comentário: