Coding Tech Life

Information Retrieval - Quiz 2 - BITS PILANI WILP

Information Retrieval - SSZG537 - Quiz 2

BITS PILANI WILP

1. Which is true about clustering algorithms?
Select one:
a. Flat algorithms are those which create unrelated partitioning of documents into clusters.
b. Soft clustering algorithms are those in which a document belongs to exactly one cluster.
c. All of the above.
d. Hard clustering algorithms are those in which a document can belong to more than one cluster.

Ans: a. Flat algorithms are those which create unrelated partitioning of documents into clusters.

2. Rule-based machine translation in Cross-language Information retrieval involves:
Select one:
a. Involves very less semantic analysis.
b. Inter-lingua representation.
c. All of the above.
d. Involves very less syntactic analysis

Ans: c. All of the above.

3. The termination criteria for k-means algorithm is:
Select one:
a. Centroid positions don't change
b. Terminate when the Residual Sum of Squares distance falls below a threshold.
c. All of the above.
d. A fixed number of iterations

Ans: c. All of the above.

4. Which is true about the Bernoulli model of text classification?
Select one:
a. It does not consider the probability of non-occurrence of the terms of the vocabulary in the test document.
b. It estimates P(t|c) as the fraction of documents of class c that contain term t
c. It considers the number of occurrences of the term in the test document.
d. It estimates P(t|c) as the fraction of tokens or fraction of positions in documents of class c that contain term t

Ans: d. It estimates P(t|c) as the fraction of tokens or fraction of positions in documents of class c that contain term t

5. Group-average agglomerative clustering (GAAC) is determined by:
Select one:
a. Average similarity of all document pairs including those from the same cluster but self-similarities are not included in the average.
b. Average similarity of all document pairs including those from the same cluster.
c. Average similarity of all document pairs excluding those from the same cluster.
d. Average similarity of all document pairs excluding those from the same cluster but self-similarities are not included in the average.

Ans: a. Average similarity of all document pairs including those from the same cluster but self-similarities are not included in the average.

Ref: http://nlp.stanford.edu/IR-book/html/htmledition/group-average-agglomerative-clustering-1.html

6. Document frequency of a term is the:
Select one:
a. Number of documents that contain the term.
b. None of the above.
c. Number of times the term appears in the document
d. Number of times the term appears in the collection.

Ans: a. Number of documents that contain the term.

Search This Blog

Search

Blog Archive

► 2023 (7)
► 2022 (6)
► 2021 (6)
► 2020 (6)
▼ 2017 (66)
► December (2)
► November (8)
► October (8)
► September (4)
► August (2)
► May (2)
► April (20)
▼ March (8)
Machine Learning (ISZC464) Comprehensive Question ...
Information Retrieval - SSZG537 - Comprehensive Q...
Information Retrieval - Quiz 2 - BITS PILANI WILP
Usability Engineering- SZG547 Mid-Semester Questio...
Data Mining - ISZC415 - Quiz 2 - BITS-WILP
Data Mining - ZC415 - Mid-Semester Question Paper ...
Information Retrieval - SSZG537 - Mid-Semester Que...
Artificial Intelligence ZC444 Mid-Semester Questio...
► February (10)
► January (2)
► 2016 (39)
► 2015 (38)
► 2014 (9)

Labels

Advanced Data Mining Android AngularJs Artificial Intelligence Automation BITS Comprehensive Test BITS Mid-Semester Test BITS Previous Question Paper BITS Quiz BITS-WILP Bootstrap ClassDiagram Cloud Computing Cooking CSS d3 Data Mining Data Storage Technologies and Networks Data

7. Which is true about the IBM models?
Select one:
a. IBM model 5 adds a fertility factor.
b. IBM model 2 is an absolute reordering model whereas IBM model 4 is a relative reordering model.
c. IBM model 3 keeps track of available positions for output words
d. IBM model 4 is an absolute reordering model whereas IBM model 2 is a relative reordering model.

Ans: b. IBM model 2 is an absolute reordering model whereas IBM model 4 is a relative reordering model.

8. The idf-weight of a rare term is:
Select one:
a. Lower than frequent term.
b. No relation.
c. Higher than frequent term.
d. Same as frequent term.

Ans: c. Higher than frequent term.

9. Optimal clustering in k-means depends upon:
Select one:
a. None of the above.
b. No. of iterations
c. Seed choice
d. Choice of objective function

Ans: c. Seed choice

10. The criteria to determine the cuts in the dendrogram is:
Select one:
a. Cut the dendrogram where the gap between two successive combination similarities is largest.
b. Cut at a pre-specified level of similarity.
c. All of the above.
d. Cut the dendrogram to obtain a pre-specified number of clusters.

Ans: c. All of the above.

11. The most common hierarchical clustering algorithms have a complexity that is:
Select one:
a. At least linear in the number of documents
b. At most linear in the number of documents
c. At most quadratic in the number of documents
d. At least quadratic in the number of documents

Ans: d. At least quadratic in the number of documents

12. Boolean queries often result in:

Select one:
a. Too many or too few results
b. None of the above.
c. Too few results
d. Too many results.

Ans: a. Too many or too few results

13. Purity of clustering is 1 when:
Select one:
a. None of the above.
b. Each document gets its own cluster.
c. Each document gets atleast one cluster.
d. The number of clusters is large.

Ans: b. Each document gets its own cluster.

14. The decision boundary between 2 clusters in Rocchio classification is found by:
Select one:
a. Line at which all points are equidistant from the centroids of the 2 clusters.
b. Line at which atleast one point is equidistant from the centroids of the 2 clusters.
c. Line at which atmost 1 point is equidistant from the centroids of the 2 clusters.
d. None of the above.

Ans: a. Line at which all points are equidistant from the centroids of the 2 clusters.

15. The more frequent the query term in the document is:

Select one:

a. The lesser the score of the document.
b. Does not make any affect.

c. The higher the score of the document.
d. None of the above.

Ans: c. The higher the score of the document.

16. The objective or the partitioning criterion in k-means text clustering algorithm is to:
Select one:
a. Minimize the average squared difference from the centroid
b. Maximize the average squared difference from the centroid
c. Maximize the residual sum of squares distance for all the clusters.
d. Minimize the residual sum of squares distance for all the clusters.

Ans: a. Minimize the average squared difference from the centroid

17. Issues with the Jaccard coefficient are:

Select one:
a. It doesn't consider term frequency.
b. It does not consider the fact that rare terms in a collection are more informative than frequent terms.
c. It is biased towards shorter documents.
d. All of the above.

Ans: d. All of the above.

18. The tf-idf weight of a term increases with:
Select one:
a. The length of the document.
b. The rarity of the term in the collection
c. The number of occurrences within a document
d. Both number of occurrences and rarity of the term.

Ans: d. Both number of occurrences and rarity of the term.

19. The best measure that is used to rank the documents is:
Select one:
a. Jaccard coefficient
b. Cosine similarity
c. Euclidean distance
d. N-gram overlap

Ans: b. Cosine similarity

20. Benefits of doing text clustering are:
Select one:
a. To improve retrieval recall
b. All of the above.
c. To compute better similarity scores.
d. To improve retrieval speed

Ans: b. All of the above.

21. kNN classification rule for k > 1 is:
Select one:
a. Assign each test document to the class of its nearest neighbour in the training set.
b. Assign each test document to the minority class of its k nearest neighbours in the training set.
c. Assign each test document to the majority class of its k nearest neighbours in the training set.
d. Assign each test document to a random class of its k nearest neighbours in the training set.

Ans: c. Assign each test document to the majority class of its k nearest neighbours in the training set.

22. Ranked retrieval models take as input:

Select one:
a. None of the above
b. Boolean queries
c. Logical queries
d. Free text queries

Ans: d. Free text queries

23. What is contiguity hypothesis in vector space classification?
Select one:
a. Documents from different classes don't overlap
b. Documents in the same class form a contiguous region of space.
c. All of the above.
d. Intra-cluster similarity is higher than inter-cluster similarity

Ans: c. All of the above.

24. A document with 10 occurrences of the term is more relevant than a document with 1

occurrence of the term. What is the degree of this relevance?

Select one:
a. Same relevance.
b. 10 Times more relevant.
c. None of the above.
d. Log of term frequency.

Ans: d. Log of term frequency.

25. Which one is true about the Bag of words model?

Select one:
a. It considers a document as a collection of term frequencies.
b. It considers a document as a collection of terms
c. Vector representation doesn't consider the ordering of words in a document.
d. All of the above.

Ans: d. All of the above.

Labels: BITS Quiz, BITS-WILP, Information Retrieval, Information Retrieval Quiz

No comments:

Post a Comment

To leave a comment, click the button below to sign in with Google.

SIGN IN WITH GOOGLE

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)