Information Retrieval (IS322) / Data Storage and Retrieval (IS313)
FINAL exam – June 2023 – Dr. Ehab Hassanein
Use the answer (bubble) sheet give the (BEST) choice
60 questions 1 mark each – submit both sheets

رقم الجلوس _____ الاسم _____

1. The problem with Boolean Search is that it _____
a. always returns too many answers b. always returns too few answers c. usually returns too few or too may answers d. returns fewer answers than ranked search e. returns more answers than ranked search.

2. In ranked retrieval models the system returns _____
a. a set of documents satisfying a query expression b. an ordering of the top documents.
c. the ranking of the different retrieval systems d. the top 25 results e. a random 25 results.

3. Feast or famine is not a problem in ranked retrieval because _____
a. The result set is always small b. The result set is always the right size. c. the result set is always big
d. It uses free text query e. the size of the result set is not an issue.

4. The biggest problem with Jaccard Coefficient _____
a. sets don't have to be the same size b. it Always assigns a number between 0 and 1 c. It does not consider term frequency in a collection d. It is difficult to compute e. sets have to be the same size.

5. Term frequency (tf) is defined as _____
a. the number of times that a term occurs in a document b. the inverse of the number of times it occurs in a document c. the number of times that a term occurs in a collection d. the inverse of the number of times it occurs in a collection e. the number of times that a term occurs in a query.

6. It is true that _____
a. Frequent terms are sure indicator of relevance. b. Frequent terms are more informative than rare terms
c. The frequency of the term is not important d. Frequent terms are less informative than rare terms
e. Term Frequency is the same as Document Frequency.

7. The collection frequency is _____
a. the number of times that a term occurs in a document b. the number of terms in the collection
c. The number of documents in the collection d. the number of times the term appears in all documents.
e. The number of documents contained the term in the collection.

8. A collection of 1,000,000 (one million) documents given a document d and two terms t1 and t2 where
t1 occurred in 100 documents and occurred 100 times in document d
t2 occurred in 100,000 documents and occurred 1000 times in document d
compute TF-IDF weight for $w_{t1,d}$ and $w_{t2,d}$
a. $w_{t1,d}=8$ & $w_{t2,d}=4$ b. $w_{t1,d}=4$ & $w_{t2,d}=8$ c. $w_{t1,d}=12$ & $w_{t2,d}=8$ d. $w_{t1,d}=12$ & $w_{t2,d}=4$ e. $w_{t1,d}=12$ & $w_{t2,d}=10$

9. Euclidean distance is a bad idea to measure similarity because the distance _____
a. does not represent the difference for equal length vectors b. represents the difference only for equal length vectors c. represents the difference for different length vectors d. the number of dimensions increases relative to the number of terms. e. is difficult to compute.

10. Choose the correct statement:
a. Semantically two documents have the same content if the angle between them is 0 b. Semantically two documents have the same content if the angle between them is 90 c. Semantically two documents have the same content if the Euclidean distance = 1 d. Semantically two documents have the same content if the Euclidean distance = 90 e. Euclidean distance has no relation to the angle between two documents.

11. if a vector lengths are 4 and 5 then L2 norm = _____

a. 9 b. 4.5 c. 10 d. $\sqrt{20}$ e. $\sqrt{41}$

The following is three terms frequency in three documents

Terms	Doc 1	Doc 2	Doc 3
Information	10	0	1000
Systems	10	1000	1000
FCI	0	1000	0

12. The cosine similarity between doc 1 and doc 2 is :

a. $\cos(doc1,doc2)=0.5$ b. $\cos(doc1,doc2)=0.75$ c. $\cos(doc1,doc2)=0.25$ d. $\cos(doc1,doc2)=1$ e. $\cos(doc1,doc2)=0$

13. The cosine similarity between doc 1 and doc 3 is :

a. $\cos(doc1,doc2)=0.5$ b. $\cos(doc1,doc2)=0.75$ c. $\cos(doc1,doc2)=0.25$ d. $\cos(doc1,doc2)=1$ e. $\cos(doc1,doc2)=0$

14. The cosine similarity between doc 2 and doc 3 is :

a. $\cos(doc1,doc2)=0.5$ b. $\cos(doc1,doc2)=0.75$ c. $\cos(doc1,doc2)=0.25$ d. $\cos(doc1,doc2)=1$ e. $\cos(doc1,doc2)=0$

15. The most similar documents is

a. doc1 & doc 2 b. doc 2 & doc 3 c. doc1 and doc 3 d. all are similar e. none is similar

16. Given one query, if you want to measure which retrieval system that returns for you result set with the most correct documents which measure would you use?

a. recall b. precision c. F-Measure d. Mean Reciprocal Rank e. Mean average precision

17. Given one query, if you want to measure which retrieval system return for you result set with that contains the biggest number of correct documents in the collection which measure would you use?

a. recall b. precision c. F-Measure d. Mean Reciprocal Rank e. Mean average precision

18. Given one query, if you want to measure which retrieval system return for you result set that contains the most balanced correct documents in the collection which measure would you use?

a. recall b. precision c. F-Measure d. Mean Reciprocal Rank e. Mean average precision

19. The number of correct documents that was not retrieved is called

a. True Positive b. True Negative c. False Positive d. False Negative e. none of them.

20. The number of wrong documents that was not retrieved is called

a. True Positive b. True Negative c. False Positive d. False Negative e. none of them.

Given

 = the relevant documents

and the following two rankings

Ranking #1

Ranking #2

21. The **recall** for each item ranking #1
a. 0.33. 0.67, 1, 1, 1, 1 b. 0, 0, 0.33,0.67,1,1 c. 0.33.0.67, 0.67, 0.67, 0.67, 0.67
d. 1, 1, 0.67, 0.5, 0.4, 0.33 e. 0, 0, 0, 0.33, 0.67, 1

22. The **precision** for each item ranking #1
a. 0.33. 0.67, 1, 1, 1, 1 b. 0, 0, 0.33, 0.67, 1 c.0.33. 0.67, 0.67, 0.67, 0.67
d. 1, 1, 0.67, 0.5, 0.4, 0.33 e. 0. 0, 0.2, 0.25, 0.35

23. The **recall** for each item ranking #2
a. 0.33. 0.67, 1, 1, 1, 1 b. 0, 0, 0, 0.33, 0.67, 0.67 c.0.33. 0.67, 0.67, 0.67, 0.67
d. 1, 1, 0.67, 0.5, 0.4, 0.33 e. 0, 0, 0, 0.33, 0.67, 1

24. The **precision** for each item ranking #2
a.0.33. 0.67, 1, 1, 1 b. 0. 0, 0.33, 0.67, 1 c.0.33. 0.67, 0.67, 0.67, 0.67
d. 0. 0, 0, 0.2, 0.25, 0.35 e. 0, 0, 0, 0.25, 0.4,0.5

25. The Average Precision for ranking #1
a. 100% b. 80% c. 60% d. 40% e. 20%

26. The Average Precision for ranking #2
a. 100% b. 76% c. 72% d. 38% e. 19.1%

27. When building a crawler we need to fetch and parse each URL in order to:
a. extract the images b. extract the links c. extract information for a query
d. avoid spider traps e. filter the URL

28. A URL that leads to a set of web pages that may be used to cause a crawler to crash is called:
a. URL normalization b. URL filter c. URL frontier d. a spider traps e. URL elimination

29. BERT, is a deep learning model that is based on _____.
a. GPT b. BART c. decoders d. encoders e. Transformers

30. BERT uses the surrounding text to provide _____(1)_____in order to help computers understand the meaning of_____(2)_____ in text
a. (1) filter, (2) long sentences b. (1) context, (2) ambiguous words
c. (1)filter, (2) short sentences d. (1) context, (2) clear words e. (1)explanation, (2) conflicting words

31- The process that involves retrieval of data from various sources in order to process it further is called:
a- Data Extraction b-Information retrieval c- Web Mining d- Data Mining e-Data Analysis

32- The automated retrieval of specific information related to a selected topic from bodies of text is called
a- Data Analysis b- Crawling c-Data Extraction d-Data Mining e-Information Extraction

33- The Goldberg machine is a ---------- Machine that searched for a pattern of dots or letters across catalog entries stored on a roll of microfilm.
a- Magnetic Tape b-Mechanical c-Electronic d-Laser e-Digital

34- The fraction of the returned results are relevant to the information need is called -----
a. recall b. precision c. f-measure d. relevance e. soundness

35- Consider Grepping: It is NOT true that:
a. It is a very effective process b. grep is a UNIX command c. Impractical for near queries
d. good for ranked retrieval e. allows useful possibilities for wildcard pattern matching

36- The Boolean Retrieval model is a ---------
a. model for information retrieval b. model that views a document as a set of sentences
c. data model d. good model for ranked retrieval e. a model for ranked retrieval

37- ---------- is the topic about which the user desires to know more
a- A query b- An information need c-A user task d- A misconception e. A misformulation

38- ---------- is what the user conveys to the computer in an attempt to communicate the information need.
a- A query b- An information need c-A user task d- A misconception e. A misformulation

39- if the result is called ---------- that means the user perceives as containing information of value with respect to his information need.
a. valid b. complete c. reasonable d. relevant e. incomplete

40- The fraction of the relevant documents in the collection were returned by the IR system is called -----
a. recall b. precision c. f-measure d. relevance e. soundness

Given the following Term-Document Incidence Matrix for questions (11-15)

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6
Egypt	1	1	1	0	1	1
Syria	0	0	1	1	0	1
Russia	1	0	1	0	1	0
France	0	1	1	0	0	0
Iraq	1	1	0	1	1	1

41- The query Russia and Egypt and France will result to
a. 110110 b. 111011 c. 100010 d. 001001 e. 001000

42- The query Russia and Egypt not France will result to
a. 110110 b. 111011 c. 100010 d. 001001 e. 001000

43- Which document has Syria and Iraq but not Egypt
a. 1 b. 2 c. 3 d. 4 e. 5

44- The posting list 1,3,5 is for
a. Doc 1 b. Doc 2 c. Egypt d. Syria e. Russia

45- The given matrix is not typical because it
a. has a big collection b. has too many terms c. is sparse d. is not sparse e. has a lot of zero's

46- If the Term-Document Incidence Matrix is sparse then the equivalent inverted index
a. contains fewer terms b. contains more terms
c. contains shorter posting lists d. contains longer posting lists e. use more memory

47- **In a Boolean retrieval system, stemming** ----------
a. increase the size of the vocabulary b. never lowers precision. c. can increase the retrieved set
d. increase the number of relevant documents e. should not be invoked at indexing

48- In a Boolean retrieval system, stemming never lowers recall because stemming-------
a. will decrease the retrieved set b. can increase the retrieved set c. increase the size of the vocabulary
d. decrease the size of the vocabulary e. increase the number of relevant documents

49- If the collection is 1,000,000 and the number of terms is 100,000 and the number of terms in a query is 5, what is the maximum size of any posting list
a. 500,000 b. 5,000,000 c.1,000,000 d. 100,000 e. 20,000

50- If the collection is 1,000,000 and the number of terms is 100,000 and the number of terms in a query is 5, what is the maximum number of posting lists. (assume no phrases)
a. 500,000 b. 5,000,000 c.1,000,000 d. 100,000 e. 20,000

51- In the initial stages of text processing Tokenization is the process of:
a. cut character sequence into words b. mapping text and query terms to the same form
c. omitting very common words d. matching different forms of a root e. authorization

52- In the initial stages of text processing Stemming is the process of:
a. cut character sequence into words b. mapping text and query terms to the same form
c. omitting very common words d. matching different forms of a root e. authorization

53- The goal of the Extended Boolean model is to overcome the drawbacks of the Boolean model that has been used in information retrieval which mainly was -----
a. always too few results b. always too much results c. always wrong results
d. bad ranking of the result set e. the result set is is often too small or too big

54- WestLaw is NOT _____
a- an example of Extended Retrieval Model b. a type of a Boolean model c. legal search service
d. a model that require special query language e. for western Diplomacy

55- Not Knowing what to search for in order to get your information need is called
a. False information b. miscommunication c. misformulation d. Misconception e. fake information

56- The main issues for biword indexes
a. slower than positional indexes b. famous names such as "Mohamed Ali"
c. complicated inverted index d. False positives e. stop words

57- Not Knowing how to write suitable query for your information need is called
a. False information b. miscommunication c. misformulation d. Misconception e. fake information

Given the following portion of a positional index (FOR 58,59=)
angels: 2: (36,174,252,651); 4: (12,22,102,432}; 7: (17) ;
fools: 2: (1,17,74,222); 4: (2, 18,78,108,458); 7: {3,13,23,193) ;
fear : 2: (87,704,722,901) ; 4: (13,43,113,433); 7: (18,328,528) ;
in: 2: (3,37,76,444,851); 4: (3,10,20,110,470); 6: (5,15,25,195);
rush: 2: (2,66,194,321,702); 4: (6, 9, 19,69,114,429,569); 7: (4,14,404);

58- Which document(s) if any meet the positional query "fools rush in"
a. 2, 4, 7 b. 2,4, 6 c. 4, 7 d. 2. 4 e. none of them

59- Which document(s) if any meet the positional query "angels fear rush"
a. 2, 4, 7 b. 2,4, 6 c. 4, 7 d. 2. 4 e. none of them

60 - which of the following westlaw queries will find the following sentence
happiness is an emotional state characterized by feelings of joy
a. happ! /s emot! /p joy satis! b. happ! /s emot! /2 joy satis! c. happ! /p emot! /2 joy satis!
d. happy /s emot! /p joy satis! e. happ! /s emot! /p satis!