CS5154/6054 Midterm Exam Key

1. Suppose there is only one relevant document in a collection of 1000. A very bad IR program simply retrieves all the documents. In this case, the harmonic mean of precision and recall, or the F1 measure, 2PR/(P + R), is closest to P=1/1000, R=1, 1/F1=(1/P+1/R)/2=1001/2, F1=2/1001
   a. 0.02.
   b. 0.2.
   c. 0.5.
   d. 0.002.

2. How many zeros are there in a term-document matrix with M terms and N documents and an average of L terms for each document? area=MN, number of 1's=LN
   a. (N-L)M
   b. (M-N)L
   c. (L-N)M
   d. (M-L)N

3. Which text emoticon is not matched by the regular expression r'[:;<]\-?[\)\(3]'? no \\
   a. :-\
   b. :3
   c. :-)
   d. ;-(

4. When all the documents are retrieved, all relevant are retrieved
   a. precision becomes 1.
   b. recall becomes 1.
   c. accuracy becomes 1.
   d. specificity becomes 1.

5. The regular expression r'<[^<>]*>' matches all except no < or > between < and >
   a. <12345>.
   b. <[^<>]*>.
   c. <>.
   d. <^^^>.

6. The "l" variant of term frequency (Fig. 6.15) replaces tf with wf. When tf = 1, we have 1+log(tf)
   a. wf = 0.
   b. wf = 0.1.
   c. wf = 10.
   d. wf = 1.

7. RNNRRNNRNN are the top ten results with four relevant documents (the leftmost is the top ranked one). There are four precisions at the four recall levels (1, 2, 3, 4 from left to right every time there is an R). At which recall level the precision gets interpolated on the precision-recall curve? precisions at these recall levels are 1/1, 2/4, 3/5, and 4/8. 2/4 is interpolated with 3/5
   a. 2
   b. 1
   c. 3
   d. 4

8. Most times the documents and queries are represented as vectors with non-negative elements. In this case, the cosine similarity between two of them has the range of
   a. [0, 1].
   b. [0, 2].
   c. [-2, 2].
   d. [-1, 1].

9. In the vector space where documents and queries are represented, inverse document frequency (idf) acts as a weight on
    a. the angle of a vector.
    b. a direction.
    c. the magnitude of a vector.
    d. a dimension.

10. Boolean retrieval presents
    a. all and only the relevant documents.
    b. all the relevant documents.
    c. only the relevant documents.
    d. top ranked relevant documents.

11. An information need is satisfied by
    a. retrieved documents.
    b. relevant documents retrieved.
    c. a query.
    d. a bag of words.

12. Ad hoc retrieval is only for
    a. standing queries.
    b. static document collections.
    c. dynamic document collections.
    d. arbitrary user information needs.

13. Documents that are nonrelevant and not retrieved are
    a. true negatives.
    b. true positives.
    c. false positives.
    d. false negatives.

14. Hamming distance between sets ABC and ABDE is
    a. 1.
    b. 3.
    c. 4.
    d. 2.

15. re.findall(r'\w\w+', text) may return tokens like
    a. I.
    b. remove_accents.
    c. 9:30am.
    d. 2021-10-14.

16. Jaccard coefficient between sets ABC and ABDE is
    a. 2/7.
    b. 2/5.
    c. 5/7.
    d. 3/4.

17. Suppose there is only one relevant document in a collection of 1000. A very bad IR program simply retrieves all the documents. The value "true negatives" (tn) is
    a. 999.
    b. 1.
    c. 1000.
    d. 0.

18. Stop words are often those with
    a. very small postings lists.
    b. very high document frequencies.
    c. very low document frequencies.
    d. very high tfidf weights.

19. Suppose there is only one relevant document in a collection of 1000. A very bad IR program simply retrieves all the documents. In this case, the arithmetic mean of precision and recall (P + R)/2 is closest to P=1/1000, R=1, (P+R)/2 = 1.0001/2 = 0.5005
    a. 0.2.
    b. 0.5.
    c. 0.002.
    d. 0.02.

20. Which value is context-dependent/collection-dependent?
    a. document frequency.
    b. term frequency.
    c. Jaccard coefficient.
    d. Hamming distance.

21. Without normalization, cosine similarities are dot products (SMART notation "n"). Dot product between vectors [1, 2, 0] and [0, 2, 1] is iir 6.3.1
    a. 4.
    b. 1.
    c. 3.
    d. 2.

22. When half of the retrieved documents are relevant,
    a. accuracy = 0.5.
    b. precision = 0.5.
    c. F1 = 0.5.
    d. recall = 0.5.

23. From sklearn, we have tfidf = TfidfVectorizer(); dt = tfidf.fit_transform(tweets["text"]). A call to cosine_similarity(dt.T, dt.T) generates a similarity matrix between IR9D.py and slide 9/23/25
    a. documents and terms.
    b. terms.
    c. documents.
    d. documents and queries.

24. Suppose that half of the documents are relevant and half are retrieved. Which combination is not possible? slide 9/7/22
    a. precision = recall = 1.
    b. precision = recall = 0.5.
    c. precision != recall.
    d. precision = recall = 0.