Answer Midterm 2024 – 03 – 26

Questions 1

1) A data structure that maps terms back to the parts of a document in which they occur is called an

a) Dictionary b) Inverted Index c) Incidence Matrix d) List

2) The model of information retrieval in which a user can search for products on an e-commerce website by entering specific criteria is called the ranked retrieval model.

a) True b) False

3) The number of times that a word or term occurs in a book is called the:

a) Proximity Operator b) Term Frequency

c) document frequency d) collection frequency

4) In a school library, common words that are excluded from the index vocabulary to help students select books efficiently are known as:

a) stop words b) tokens c) stemmed terms d) Lemmatized Words

5) Which of the following is NOT a component of an information retrieval system?

a) Query Processor b) Document Collection c) Relevance Feedback d) Text Classification

6) In tokenization language issues we may use

a) package splitter b) symmetric splitter c) compound splitter d) all of the above

7) The study of identifying the meaningful parts of words, i.e. prefixes, suffixes and roots.

a) lemmatization b) stemming

c) morphological analysis d) both a,b

8) What does the 'Precision' metric measure in information retrieval?

a) The proportion of irrelevant documents retrieved b) The proportion of true positive predictions

c) The proportion of false positive predictions d) The proportion of relevant documents retrieved

9) How should dates like '55 BC' be treated during tokenization?

a) Ignore them b) Preserve as a single token

c) Expand into multiple tokens d) Split into separate tokens

10) Term document incidence matrix is

a) Sparse b) Depends upon the data c) Dense d) Cannot predict

11) What is the term used to describe data that is not explicitly structured?

a) Unstructured data b) Semi-structured data c) Structured data d) Relational data

12) Which of the following is correct search statement for "Select all items that discuss software and not storage or hardware in the item"?

a) SOFTWARE OR (STORAGE NOT HARDWARE)

b) SOFTWARE OR STORAGE NOT HARDWARE

c) SOFTWARE AND NOT STORAGE OR HARDWARE

d) SOFTWARE OR NOT STORAGE AND HARDWARE

13) What is the role of an IR system in a hospital?

a) Tracking medical supplies

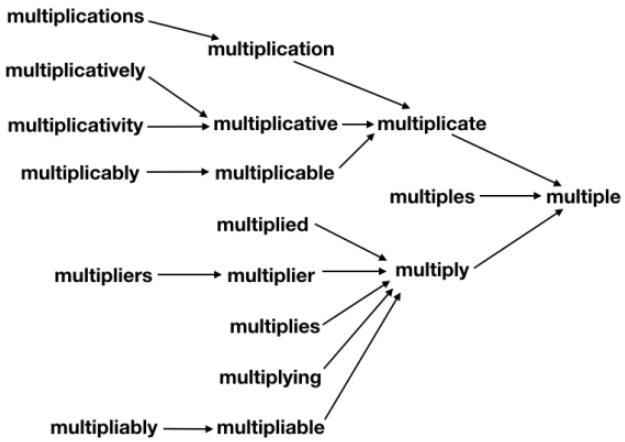
b) Diagnosing medical conditions

c) Retrieving patient information

d) Managing patient records

Questions 2:

A. Based on the figure, answer the following question:



1) The figure is an example of Lemmatization while removing "e" from the word *multiple* is an example of stemming

2) "my brother went to Finland's university. There he enjoyed studying 'how to co-educate with new friends'" — tokenize & normalize the following sentence and state how many type, terms and tokens are found.

Types → 15 **Terms → 9** **Tokens → 16**

b) What are the returned results for this query: `(happy AND live) OR dagger AND NOT `romeo`?

Happy 01000 dagger 01100

And OR And

Live 00010 Not Romeo 01011

00000 Happy 01000

Happy 01000

	d_1	d_2	d_3	d_4	d_5
romeo	1	0	1	0	0
juliet	1	1	0	0	0
happy	0	1	0	0	0
dagger	0	1	1	0	0
live	0	0	0	1	0
die	0	0	1	1	0
free	0	0	0	1	0
new-hampshire	0	0	0	1	1