

MCQ:

1. What is the inverted index pipeline?
 - a. Collect documents → do linguistic models → tokenize text → indexing.
 - b. **Collect documents → tokenize text → do linguistic model → indexing.**
 - c. Collect documents → indexing → do linguistic models → tokenize text.
 - d. Collect documents → do linguistic models → indexing → tokenize text.
2. IR systems should be designed to offer choices of granularity?
 - a. **True.**
 - b. False.
3. A is an instance of a sequence of characters that are grouped together as a useful semantic unit.
 - a. **Token.**
 - b. Type.
 - c. Term.
4. is all tokens containing the same character sequence.
 - a. Token.
 - b. **Type.**
 - c. Term.
5. is normalized type that is included in the dictionary.
 - a. Token.
 - b. Type.
 - c. **Term.**
6. Which of the following is/are approach to handle tokenization issues?
 - a. Word segmentation.
 - b. Machine learning sequence models.
 - c. K-gram.
 - d. **All of them.**
7. . Normalization is the process of canonicalizing tokens so that matches occur despite differences in the character sequences.
 - a. **True.**
 - b. False.
8. Often best to normalize to a de-accented term.
 - a. **True.**
 - b. False.
9. is reducing inflectional/variant forms to base form.
 - a. Tokenization.
 - b. **Lemmatization.**
 - c. Stemming.

10. is reducing terms to their roots before indexing.
- Tokenization.
 - Lemmatization.
 - Stemming.**
11. "automate(s), automatic, automation all reduced to automat", is an example of
- Stemming.**
 - Tokenization.
 - Lemmatization.
12. "am, are, is → be" is an example of lemmatization.
- True.**
 - False.
13. Stemming is language dependent.
- True.**
 - False.
14. Porter's algorithm is commonest algorithm for stemming English.
- True.**
 - False.
15. More skips mean shorter skip spans but lots of comparison, while fewer skips mean few comparison but long skip span.
- True.**
 - False.
16. In index every consecutive pair of terms in the text as a phrase.
- Inverted.
 - Biword.**
 - Positional.
17. Using biwords index, longer phrase queries can be broken into the Boolean query on biwords.
- True.**
 - False.
18. If the index includes variable length word sequences, it is generally referred to as a phrase index.
- True.**
 - False.
19. In , postings store for each term the position(s) in which tokens of it appear.
- Inverted index.
 - Bi-words index.
 - Positional index.**
20. A positional index is 2–4 as large as a non-positional index.
- True.**
 - False.

21. Positional index size 35–50% of volume of original text.

- a. **True**.
- b. False.