1. **Pengenalan Information Retrieval**

   Information Retrieval adalah proses menemukan informasi relevan dari kumpulan dokumen berdasarkan kebutuhan pengguna. Fokus utama adalah efisiensi pencarian dan akurasi relevansi.

2. **Model Boolean Retrieval**

   Model ini menggunakan logika boolean AND, OR, NOT untuk menentukan dokumen relevan. Cocok untuk dataset kecil dan struktur sederhana.

3. **Vector Space Model**

   Model ini menggunakan representasi vektor untuk dokumen dan query. Skor relevansi dihitung menggunakan cos similarity.

4. **Tokenisasi dalam IR**

   Tokenisasi memecah teks menjadi unit kata. Tahap ini penting dalam indexing dokumen sebelum pencarian dilakukan.

5. **Stopword Removal**

   Stopword seperti "dan", "atau", "adalah" dihapus agar pencarian lebih akurat. Mengurangi noise pada data.

6. **Stemming Bahasa Indonesia**

   Proses mengembalikan kata menjadi bentuk dasar, misalnya “mengajar”, “pengajar”, “ajarannya” → “ajar”.

7. **Lemmatization**

   Lebih kompleks dari stemming, mempertimbangkan konteks gramatikal untuk mendapatkan bentuk dasar kata.

8. **Inverted Index**

   Struktur data yang menyimpan daftar kata dan lokasi kemunculannya di dokumen. Dasar dari search engine modern.

9. **Term Frequency (TF)**

   Menghitung seberapa sering kata muncul dalam dokumen. Menjadi salah satu elemen utama bobot kata

10. **Inverse Document Frequency (IDF)**

    Mengurangi bobot kata yang terlalu umum. Kata yang jarang muncul dianggap lebih informatif.

11. **TF-IDF Weighting**

    Kombinasi TF dan IDF untuk menghasilkan bobot kata lebih akurat dalam model pencarian.

12. **Query Expansion**

    Mengembangkan query dengan sinonim untuk meningkatkan recall dalam pencarian.

13. **Spell Correction Pada Search Engine**

    Mesin pencari memperbaiki kesalahan ketik dengan teknik edit distance.

14. **Stopword Bahasa Indonesia dalam IR**

    Contoh stopword: yang, kepada, namun, ketika, karena. Berfungsi mengurangi beban komputasi

15. **Document Ranking**

    Proses mengurutkan dokumen berdasarkan skor relevansi pada query.

16. **Cosine Similarity**

    Pengukuran kesamaan antara vektor dokumen dan vektor query.

17. **Jaccard Similarity**

    Menghitung kesamaan berdasarkan irisan dan gabungan kata antara dua teks.

18. **Precision dalam Evaluasi IR**

    Precision mengukur ketepatan hasil: jumlah dokumen relevan yang benar-benar muncul.

19. **Recall dalam Evaluasi IR**

    Recall mengukur kelengkapan pencarian: berapa banyak dokumen relevan yang berhasil ditemukan.

20. **F1-Score pada IR**

    Harmonic mean antara precision dan recall, memberikan nilai evaluasi menyeluruh.

21. **Query-By-Example**

    Pengguna memberikan contoh dokumen untuk mencari dokumen serupa.

22. **Relevance Feedback**

    Sistem belajar dari dokumen yang dipilih pengguna untuk meningkatkan hasil pencarian berikutnya.

23. **Index Compression**

    Teknik untuk memperkecil ukuran index menggunakan gamma coding atau variable byte

24. **Search Engine Pipeline**

    Meliputi crawling, indexing, ranking, dan returning results.

25. **Web Crawling**

    Proses mengumpulkan halaman web secara otomatis untuk disimpan dalam basis data search engine.

26. **Penggunaan Bigram dan N-gram**

    Membantu memahami konteks antar kata, misalnya "data mining".

27. **Statistical Language Model**

    Menggunakan probabilitas urutan kata untuk memperbaiki hasil pencarian.

28. **Query Log Mining**

    Menganalisis log pencarian pengguna untuk personalisasi search engine.

29. **Personalized Search**

    Hasil pencarian disesuaikan dengan profil pengguna.

30. **Semantic Search**

    Mesin pencari memahami makna kata, bukan hanya mencocokkan teks literal.

31. **Named Entity Recognition (NER)**

    Mengetahui entitas seperti nama orang, lokasi, organisasi dalam teks.

32. **Document Clustering dalam IR**

    Dokumen dikelompokkan berdasarkan kemiripan untuk mempercepat proses pencarian.

33. **Topic Modeling untuk Klasifikasi Dokumen**

    LDA digunakan untuk menemukan topik tersembunyi.

34. **Query Suggestion**

    Sistem memberikan saran pencarian berdasarkan tren.

35. **Evaluation Benchmark IR**

    Dataset seperti TREC dan Cranfield digunakan untuk mengukur akurasi mesin pencari.