1. What are the main problem of Boolean search *[2 marks]*

gives too few; OR gives too many.

2. Compare Information retrieval (IR) with Information extraction (IE) *[2 marks]*

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections

Information extraction (IE) is the automated retrieval of specific information related to a selected topic from a body or bodies of text.
So in IR we find the materals that may include the information but in IE we find the specific information themseves.

3. Compare web crawler with web scraper. *[2 marks]*

A web crawler sometimes called a "spider," is a standalone bot that systematically scans the Internet for indexing and searching for content, following internal links on web pages.
A web scraper is a process of extracting specific data. Unlike web crawling, a web scraper searches for specific information on specific websites or pages.
Web crawler scan and index whereas web scraper scan and find and return the relevant information

4. What are the main issues for biword indexes?
Why it may yield a false positive? *[2 marks]*

False positives, for phrases larger tha 2 words

Index blowup due to bigger dictionary

5. Write a query using Westlaw syntax which would find any of the words:
professor , teacher , or lecturer in the same sentence as a form of the verb explain .
[2 marks]

professor teacher lecturer /s explain!

6. Given the following part of a positional index *[2 marks]*

– to → 2:1,17,74,222,551; 4:8,16,190,429,433; 7:13,23,191,...

– be → 1:17,19; 4:17,191,291,430,434; 5:14,19,101,…

which document(s) has the phrase to be

4

7. Compare Data mining with Web Mining *[2 marks]*

Data mining is the method of analyzing expansive sums of data in an exertion to discover relationships, designs, and insights.

Web Mining is the method of utilizing data mining strategies and calculations to extract information specifically from the net by extracting it from web documents and services, substance, hyperlinks and server logs.

So web Mining is using data mining techniques to extract information from the web.

8. Given *[3 marks]*

Doc 1: feature engineering used in software engineering

Doc 2: software engineering is fun

Doc 3: computer engineering used hardware

Doc 4: software testing is fun

i. Draw the posting list for: software and engineering

ii. Draw the term-document incidence matrix

i. engineering(3) → 1, 2, 3

software (3) → 1, 2, 4

ii.	d1	d2	d3	d3
feature	1	0	0	0
engineering	1	1	1	0
used	1	0	1	0
in	1	0	0	0
software	1	1	0	1
is	0	1	0	1
fun	0	1	0	1
computer	0	0	1	0
hardware	0	0	1	0
testing	0	0	0	1

9. Write an algorithm for the merging of two postings lists *[3 marks]*

```
INTERSECT(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```