

EECS 498 System Design of a Search Engine
Winter 2026 Final Exam Solution

This exam is closed book, closed notes, no cellphones, computers or other electronic devices. The standard amount of time for this exam is two hours.

You are to abide by the University of Michigan College of Engineering honor code. To receive a grade, you must sign below, agreeing to abide by that code.

I have neither given nor received aid on this exam, nor have I concealed any violations of the Honor Code.

Signature _____

Printed name _____

Uniqname _____

This exam has **8** pages. Please be sure you have them all and that you have put your uniqname on every page.

Section 1. True/False and multiple choice. 25 questions, each worth 2 points.
Circle your choices.

1. Copy-on-write happens when a process writes to a memory-mapped file.
True **False**
2. Stemming will increase the average posting list length.
True False
3. Span of control is typically about 18 direct reports.
True **False** (It's more like 5 to 12.)
4. You're selling a house rumored to be haunted. Do you have to disclose this?
Yes **In some places** No, ghosts aren't real
5. BNF stands for Brown Norris Format, named for the inventors, Drs. Emmett Brown and Charles "Chuck" Norris.
True **False**
6. Each rule in BNF is called a production.
True False
7. Which of the following are true about Allow or Disallow directives that appear before any User-agent?
☐ They apply to all robots.
☐ Robots.txt is completely voluntary. Every crawler does something different.
☐ **They are ignored.**
☐ They are treated as an error, causing most crawlers to ignore the rest of the file.

8. Which of the following are true about a Sitemap statement?

- ☐ It applies only to the preceding group of user-agents.
- ☒ It applies to all robots no matter where it appears in the file.
- ☐ It must appear at the top of the file.
- ☐ You can only have one.
- ☐ You can have lots but only the last applies.

9. A Bloom filter can return false positives but never false negatives when checking if an element is in a set.

True False

10. What is the biggest problem with using a simple FIFO queue for your crawler's frontier?

- ☐ It cannot store enough URLs in memory.
- ☒ It doesn't consider prioritization or politeness.
- ☐ It cannot handle HTTP requests properly.
- ☐ It prevents the crawler from discovering new URLs.

11. In a distributed web crawler, machines often send discovered URLs to other machines responsible for those URLs. Why is it better to send URLs in batches rather than streaming them one at a time?

- ☒ Batching reduces network overhead and improves efficiency.
- ☐ Streaming URLs individually does not lead to good load balancing.
- ☐ Batching ensures URLs are processed in strict FIFO order.
- ☐ Streaming prevents duplicate URLs from being sent.

12. Stemming will not change the total number of posts in the index.
- True False
13. An EndDoc ISR is used to ensure that all the words in a phrase are in the same document.
- True False
14. You're selling a house you know has carpenter ants. The buyers hired a licensed and bonded home inspector who wandered around the house for hours but did not report any ants. He's an expert, and they're relying him. Do you have to disclose the ants?
- Yes No
15. BNF allows a grammar to be defined for a compiled language. Recursion was not allowed in early versions of BNF and was added later.
- True False
16. $\langle x \rangle ::= \langle y \rangle [\langle z \rangle]$ matches a y followed by any number of z's.
- True False
17. Businesses make a distinction between line managers who have "people responsibility" and staff employees who do not.
- True False
18. In the Hawthorne studies, they found it didn't matter if the lights were turned up or down. So, they turned them down to save electricity.
- True False
19. In 1943, Abraham Maslov proposed a hierarchy of needs, arguing that self-respect and family were among the most basic across all societies.
- True False
20. You just got fired because they didn't like the red tie you wore last week. But you have never worn a red tie and don't even own one. Was that legal?
- Yes No (The reason doesn't have to be true, just not protected.)

21. Various federal laws make it illegal for employers to discriminate based on protected status. They also cannot retaliate because someone complained of discrimination unless the complaint was untrue and made in bad faith.
- True False
22. The incidence of antisocial personality disorder and psychopathy in the general population is around:
- 0.1% 0.5% 1% 5%
23. Generally speaking, the more words in the query, the fewer matching pages, but the more time it takes to rank each one.
- True False
24. Pathnames in an Allow or Disallow directive are case-sensitive.
- True False
25. Matches in the URL are more important than matches in the body text, but less important than matches in the title.
- True False.

Section 2. Short answer. 10 questions, each worth 5 points.
Please write clearly. You have plenty of time.

26. Define serialization and deserialization.

Serialization is the process of converting in-memory objects into a format that can be stored on disk or transmitted over the wire. Deserialization is the process of reconstructing that data back into an active object.

27. Why is anchor text hard to do?

Any good reason, e.g.,

1. The instructor never explained how to do it.
2. The anchor text isn't on the page you're indexing.
3. Anchor text comes in with every link to a page, before and after it's crawled and indexed.

28. What is the main downside to tf-idf and BM25 as ranking algorithms? Why aren't they sufficient by themselves?

They're both "bag of word" techniques that only consider term frequency, not where the words on the page or their proximity to each other.

29. What is the purpose of a Container ISR?

It allows for queries that contain excluded terms, i.e., terms that shouldn't appear in a matching page.

30. If someone wants to disallow all crawlers from their site, what two lines of text are needed and where do they go?

Put this into robots.txt in the root directory:
user-agent: *
disallow: /

31. You've been asked to implement a top-down recursive descent parser for a grammar containing the following rule. Is it correct? If not, how would you fix it?

`<Add_Expr> ::= <Add_Expr> { '+' <Multiply_Expr> }`

This production is left recursive. Written directly as a TDRD routine, it would recurse indefinitely. It should be rewritten as either

```
<Add_Expr> ::= <Multiply_Expr> { + <Multiply_Expr> }  
<Add_Expr> ::= <Multiply_Expr> { + <Add_Expr> }
```

32. What problems do codes of professional ethics typically address? Why do they focus on these problems and how do they typically address them?

Professional ethics are generally designed to address problems of information asymmetry and reliance. Engineers, lawyers, doctors, and other professionals have special expertise that normal people don't and clients rely on their advice. It would be easy for a professional to mislead a client. Professional ethics, e.g. published by IEEE and other professional organizations, address these problems with ethics rules which require disclosing possible conflicts of interest, being honest/realistic in claims and estimates, etc.

33. What is the simplest way you might guarantee your engine never takes longer than 1 second to return results?

Set a timer. When one second has elapsed, return what you have.

34. Consider an Phrase ISR currently pointing at a match. Outline an algorithm by which the constraint solver would find the next matching document.

1. Seek all ISRs to the first occurrence beginning just after the end of the current document.
2. Pick the furthest term and attempt to seek all the other terms to the first location beginning where they should appear relative to the furthest term.
3. If any term is past the desired location, return to step 2.
4. If any ISR reaches the end, there is no match.

35. Matches in the URL are significant when found but it can be hard to search on them with the constraint solver. Why is that? How would you solve this?

Words are often run together in the domain name and elsewhere in the URL without any punctuation between them and it's not easy to split them up for indexing. This makes it difficult to match a search term against the URL using the inverted word index and the constraint solver. To fully solve this, you need a word-breaker, typically using a dictionary approach find likely breaks between the words.