Group members:
Alisha Rath
Aditya Jagdishkumar Prajapati
Pruthviraj Urankar

Part 1:

Exercise 4.1:

In Section 4.3 we introduced the concept of the per-term index as a means to improve the index's random access performance. Suppose the posting lists for some term consists of 96 million postings each of which is 2 bytes long. In order to carry out single random access into the term's postings list, the search engine needs to perform two disk read operations:

1. Loading the per-term index into RAM
2. Loading a block B of postings into RAM, where B is identified by means of binary search on the list of synchronization points.

Let us call the number of postings per synchronization point the granularity of the per-term index. For the above access pattern, what is the optimal granularity (i.e., the one that minimizes the disk I/O)? What is the total number of bytes read from disk?

Answer:

The cost function for disk I/O can be expressed as the sum of two components: the per-term index length and the block length.
The per-term index length is calculated as the posting list length divided by the block length.
Therefore, the overall cost function is given by:
Cost function = (posting list length / block length) + block length

To minimize this cost function, we can first determine the optimal block length.
Let's denote the block length as "x." The cost function is then represented as:
Cost function = (96 * 10^6 * 2) / x + x df/dx

To find the minimum of this cost function, we need to find the derivative with respect to x and set it equal to zero:

df/dx = 1 - (192 * 10^6 / x^2)

Setting df/dx to zero:

0 = 1 - (192 * 10^6 / x^2)

Solving for x:
x^2 = 192 * 10^6
x = √(192 * 10^6)
x ≈ 13856.41 bytes
Since block length should be rounded to the nearest 2-byte increment, the optimal block length is approximately 13856 bytes.

Each posting list will then contain approximately 6928 postings, and each block should accommodate the same number of postings. The last block will have slightly fewer postings, around 5632 postings.

In summary, the optimal configuration for minimizing disk I/O cost includes a per-term index length and a block length of 13856 bytes. Therefore, the total number of bytes read from disk is 13856 * 2 = 27712 bytes.

Exercise 4.3:

Building an inverted index is essentially a sorting process. The lower bound for every general-purpose sorting algorithm is Ω(n log(n)). However, the merge-based index construction method from Section 4.5.3 has a running time that is linear in the size of the collection (see Table 4.7, page 130). Find at least two places where there is a hidden logarithmic factor.

Answer:

a) In the while loop at line 5 of Figure 4.13, the algorithm iteratively identifies the smallest value, following alphabetical order, within all the partitions. It commences with the first term and systematically examines each partition, selecting the smallest term and incorporating it into the final index.

(b) Allocating an optimal amount of memory space is crucial for the indexing process. Inadequate disk space allocation can severely hamper the algorithm's performance.

(c) At line 12 of Figure 4.12, there is a function call to sort an in-memory dictionary. Sorting dictionary entries in lexicographical order has a time complexity of O(n*log(n)).

Written Exercise 3:

Manish Patil, Sharma V. Thankachan, Rahul Shah, Wing-Kai Hon, Jeffrey Scott Vitter, Sabrina Chandrasekaran. Inverted indexes for phrases and strings. Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp 555--564. 2011.

It proposes a suffix tree approach to allow for exact phrase search without increasing the index size too much over single term indexes. I want you to write as the third written exercise a 1 page summary of their approach. It should explain what phrases are stored in the dictionary, how they are chosen, and how this allows for exact phrase search.

Answer:

Title: Inverted Indexes for Phrases and Strings - A Summary
Authors: Manish Patil, Sharma V. Thankachan, Rahul Shah, Wing-Kai Hon, Jeffrey Scott Vitter, Sabrina Chandrasekaran
Published in: Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011.

Summary:
The paper "Inverted Indexes for Phrases and Strings" introduces an innovative approach to efficient phrase search in text retrieval systems. It addresses the challenge of enabling exact phrase searches without significantly increasing the size of the index. This summary provides an overview of their approach, focusing on the storage of phrases in the dictionary, their selection criteria, and how this approach facilitates precise phrase retrieval.

1. ***Phrases in the Dictionary:***
The authors propose the use of a suffix tree-based approach for indexing and searching phrases within a text corpus. In this system, a phrase is defined as a sequence of contiguous words. Instead of indexing every possible phrase, which would be impractical and lead to a massive index, the approach selectively stores phrases that are deemed significant for retrieval.

2. ***Selection of Phrases:***
Phrases are selected for inclusion in the dictionary based on their significance and frequency within the text corpus. The authors employ a statistical measure, which combines the frequency of a phrase with its length, to determine the importance of a phrase. This helps in identifying meaningful and frequently occurring phrases while excluding less significant ones. This selection process serves the dual purpose of reducing index size and improving search accuracy.

3. ***Facilitating Exact Phrase Search:***
The key innovation in this approach lies in the use of a suffix tree to efficiently support exact phrase searches. A suffix tree is a data structure that captures all suffixes of a text in a highly compressed form. By constructing a suffix tree for the text corpus and indexing significant phrases within it, the authors achieve two major advantages:

- **Improved Retrieval Precision:** With only significant phrases stored in the index, exact phrase searches become more precise. Retrieving documents containing the exact phrase is efficient because the search directly leverages the suffix tree structure to locate and verify phrase occurrences.
- **Index Size Control:** By judiciously selecting and storing only important phrases, the authors keep the index size manageable. This is a critical achievement, as traditional methods that index all possible phrases can lead to unwieldy and resource-intensive indexes.

In conclusion, the approach presented in this paper offers an elegant solution to the challenge of efficient and accurate phrase search in large text corpora. By selectively storing and indexing significant phrases based on their importance and frequency, and harnessing the power of suffix trees, the authors manage to strike a balance between index size and retrieval accuracy. This work contributes significantly to the field of information retrieval by enabling more precise and scalable phrase searching, which is invaluable in various applications, including search engines and document retrieval systems.

Coding part:
First, I want you to go to a website that has statistics for popular web search engine queries such as: Mondovo. From these pick a list of 5 multi-word queries you find interesting.

1. Business Process Outsourcing
2. Business Plan Templates
3. Business Software Alliance
4. International Business Machines
5. Better Business Bureau

Next on at least Google and Bing search on these queries and obtain the urls of the top five results for each. The web pages associated with these urls will be your corpus.

Google

1. Business Process Outsourcing

    a. https://www.forbes.com/advisor/business/business-process-outsourcing/#:~:text=Business%20process%20outsourcing%20(BPO)%20happens,and%20supply%20chain%20management%20functions.
    b. https://www.investopedia.com/terms/b/business-process-outsourcing.asp
    c. https://www.accenture.com/us-en/services/business-process-outsourcing-index
    d. https://en.wikipedia.org/wiki/Business_process_outsourcing
    e. https://www.techtarget.com/searchcio/definition/business-process-outsourcing
2. Business Plan Templates
    a. https://www.sba.gov/business-guide/plan-your-business/write-your-business-plan
    b. https://www.smartsheet.com/content/simple-business-plan-templates
    c. https://www.hubspot.com/business-templates/business-plans
    d. https://www.forbes.com/advisor/business/simple-business-plan-template/
    e. https://www.score.org/resource/template/business-plan-template-a-startup-business
3. Business Software Alliance
    a. https://www.bsa.org/
    b. https://en.wikipedia.org/wiki/Software_Alliance
    c. https://twitter.com/BSAnews?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor
    d. https://www.reddit.com/r/msp/comments/v7mka8/bsa_software_alliance_audit_1_user_msp/
    e. https://www.linkedin.com/company/bsa-the-software-alliance
4. International Business Machines
    a. https://www.ibm.com/
    b. https://en.wikipedia.org/wiki/IBM
    c. https://finance.yahoo.com/quote/IBM/
    d. https://www.wsj.com/market-data/quotes/IBM
    e. https://www.cnbc.com/quotes/IBM
5. Better Business Bureau
    a. https://www.bbb.org/
    b. https://www.winston.com/en/legal-glossary/better-business-bureau
    c. https://www.dummies.com/article/business-careers-money/business/better-business-bureau/check-business-better-business-bureau-239870/
    d. https://www.thebalancemoney.com/filing-a-complaint-with-the-better-business-bureau-1794757
    e. https://en.wikipedia.org/wiki/Better_Business_Bureau

Bing

1. Business Process Outsourcing
    a. https://us.nttdata.com/en/services/bpo-and-bpaas?mkwid=s_dc&pcrid=&pkw=business%20process%20outsourcing&pmt=e&msclkid=306ff5747fcd1656a740904ed7f4a6e0&utm_source=bing&utm_medium=cpc&utm_campaign=DigitalOps_FY22_High&utm_term=business%20process%20outsourcing&utm_content=DigitalOps_BPOBPaaS_Exact
    b. https://go.paychex.com/smb-payroll-services?keyword=payroll%20software%20for%20small%20business&targetid=kwd-77034854921014:loc-4084&feeditemid=&loc_physical_ms=86421&adposition=&placement=&network=o&device=c&matchtype=b&campaignid=328637746&adgroupid=1232552750550759&campaign_id=7010g000000mkuTAAQ&campaign_name=PD_Sml_Bus_Pyrl&kpid=bi_cmp-328637746_adg-1232552750550759_ad-77034726130422_kwd-77034854921014:loc-4084_dev-c_ext-_sig-94fc53e515301d75048736dfbd09581a&utm_id=bing_328637746_1232552750550759_77034726130422_kwd-77034854921014:loc-4084_c&k_ignore=&kenibpid=p.405._k_94fc53e515301d75048736dfbd09581a_k_.cr617826&msclkid=94fc53e515301d75048736dfbd09581a&utm_source=bing&utm_medium=cpc&utm_campaign=seer_sea_nb_leadgen_nonbrand_paychex_multi_all_c-us_p_lead_bi_en_txt_small-business_smallbusinesspayroll&utm_term=payroll%20software%20for%20small%20business&utm_content=payroll-software___multi_txt_leadgen_nonbrand__
    c. https://www.ibm.com/consulting/outsourcing?utm_content=SRCWW&p1=Search&p4=43700075921553088&p5=e&&msclkid=924c06dd2f7a1c520b6963e7874380e0&gclid=924c06dd2f7a1c520b6963e7874380e0&gclsrc=3p.ds
    d. https://www.nice.com/solutions/business-process-outsourcers?utm_source=bing&utm_medium=cpc&utm_term=cxone&msclkid=f33e6f553f6917c770f30a6f9750aacf&utm_campaign=IP%20%7C%20US%20%7C%20EN%20%7C%20Search%20%7C%20Consolidated%20-%20CXone%20%7C%20Generic&utm_content=BPO
    e. https://www.investopedia.com/terms/b/business-process-outsourcing.asp
2. Business Plan Templates
    a. https://www.sba.gov/business-guide/plan-your-business/write-your-business-plan
    b. https://create.microsoft.com/en-us/templates/business-plans
    c. https://www.canva.com/documents/templates/business-plan/
    d. https://www.forbes.com/advisor/business/simple-business-plan-template/

    e. https://www.bplans.com/downloads/business-plan-template/
3. Business Software Alliance
    a. https://www.bsa.org/
    b. https://en.wikipedia.org/wiki/Software_Alliance
    c. https://bsadefense.com/about-the-bsa/
    d. https://hypertecsp.com/knowledge-base/business-software-alliance-bsa/
    e. https://www.vondranlegal.com/what-is-the-business-software-alliance
4. International Business Machines
    a. https://www.ibm.com/us-en?utm_content=SRCWW&p1=Search&p4=43700055662457278&p5=e&&msclkid=93d73efb01001c3d01b55bc1a8e761a0&gclid=93d73efb01001c3d01b55bc1a8e761a0&gclsrc=3p.ds
    b. https://en.wikipedia.org/wiki/IBM
    c. https://www.bloomberg.com/news/articles/2023-10-25/ibm-posts-better-than-expected-sales-affirms-cash-flow-outlook
    d. https://www.britannica.com/topic/International-Business-Machines-Corporation
    e. https://www.ibm.com/annualreport/
5. Better Business Bureau
    a. https://www.bbb.org/
    b. https://www.bbb.org/all/consumer-resources
    c. https://www.dummies.com/article/business-careers-money/business/better-business-bureau/check-business-better-business-bureau-239870/
    d. https://www.thebalancemoney.com/filing-a-complaint-with-the-better-business-bureau-1794757
    e. https://en.wikipedia.org/wiki/Better_Business_Bureau

Part 4:
corpus:
https://www.forbes.com/advisor/business/business-process-outsourcing/#:~:text=Business%20process%20outsourcing%20(BPO)%20happens,and%20supply%20chain%20management%20functions
https://www.accenture.com/us-en/services/business-process-outsourcing-index
https://www.investopedia.com/terms/b/business-process-outsourcing.asp
https://en.wikipedia.org/wiki/Business_process_outsourcing
https://www.techtarget.com/searchcio/definition/business-process-outsourcing
https://www.sba.gov/business-guide/plan-your-business/write-your-business-plan
https://www.smartsheet.com/content/simple-business-plan-templates
https://www.hubspot.com/business-templates/business-plans
https://www.forbes.com/advisor/business/simple-business-plan-template/
https://www.score.org/resource/template/business-plan-template-a-startup-business
https://www.bsa.org/
https://en.wikipedia.org/wiki/Software_Alliance
https://www.reddit.com/r/msp/comments/v7mka8/bsa_software_alliance_audit_1_user_msp/
https://www.linkedin.com/company/bsa-the-software-alliance
https://en.wikipedia.org/wiki/IBM
https://finance.yahoo.com/quote/IBM/
https://www.wsj.com/market-data/quotes/IBM
https://www.bbb.org/
https://www.winston.com/en/legal-glossary/better-business-bureau
https://www.dummies.com/article/business-careers-money/business/better-business-bureau/check-business-better-business-bureau-239870/
https://www.thebalancemoney.com/filing-a-complaint-with-the-better-business-bureau-1794757
https://en.wikipedia.org/wiki/Better_Business_Bureau
https://us.nttdata.com/en/services/bpo-and-bpaas?mkwid=s_dc&pcrid=&pkw=business%20process%20outsourcing&pmt=e&msclkid=306ff5747fcd1656a740904ed7f4a6e0&utm_source=bing&utm_medium=cpc&utm_campaign=DigitalOps_FY22_High&utm_term=business%20process%20outsourcing&utm_content=DigitalOps_BPOBPaaS_Exact

https://go.paychex.com/smb-payroll-services?keyword=payroll%20software%20for%20small%20business&targetid=kwd-77034854921014:loc-4084&feeditemid=&loc_physical_ms=86421&adposition=&placement=&network=o&device=c&matchtype=b&campaignid=328637746&adgroupid=1232552750550759&campaign_id=7010g000000mkuTAAQ&campaign_name=PD_Sml_Bus_Pyrl&kpid=bi_cmp-328637746_adg-1232552750550759_ad-77034726130422_kwd-77034854921014:loc-4084_dev-c_ext-_sig-94fc53e515301d75048736dfbd09581a&utm_id=bing_328637746_1232552750550759_77034726130422_kwd-77034854921014:loc-4084_c&k_ignore=&kenibpid=p.405._k_94fc53e515301d75048736dfbd09581a_k_.cr617826&msclkid=94fc53e515301d75048736dfbd09581a&utm_source=bing&utm_medium=cpc&utm_campaign=seer_sea_nb_leadgen_nonbrand_paychex_multi_all_c-us_p_lead_bi_en_txt_small-business_smallbusinesspayroll&utm_term=payroll%20software%20for%20small%20business&utm_content=payroll-software__multi_txt_leadgen_nonbrand__

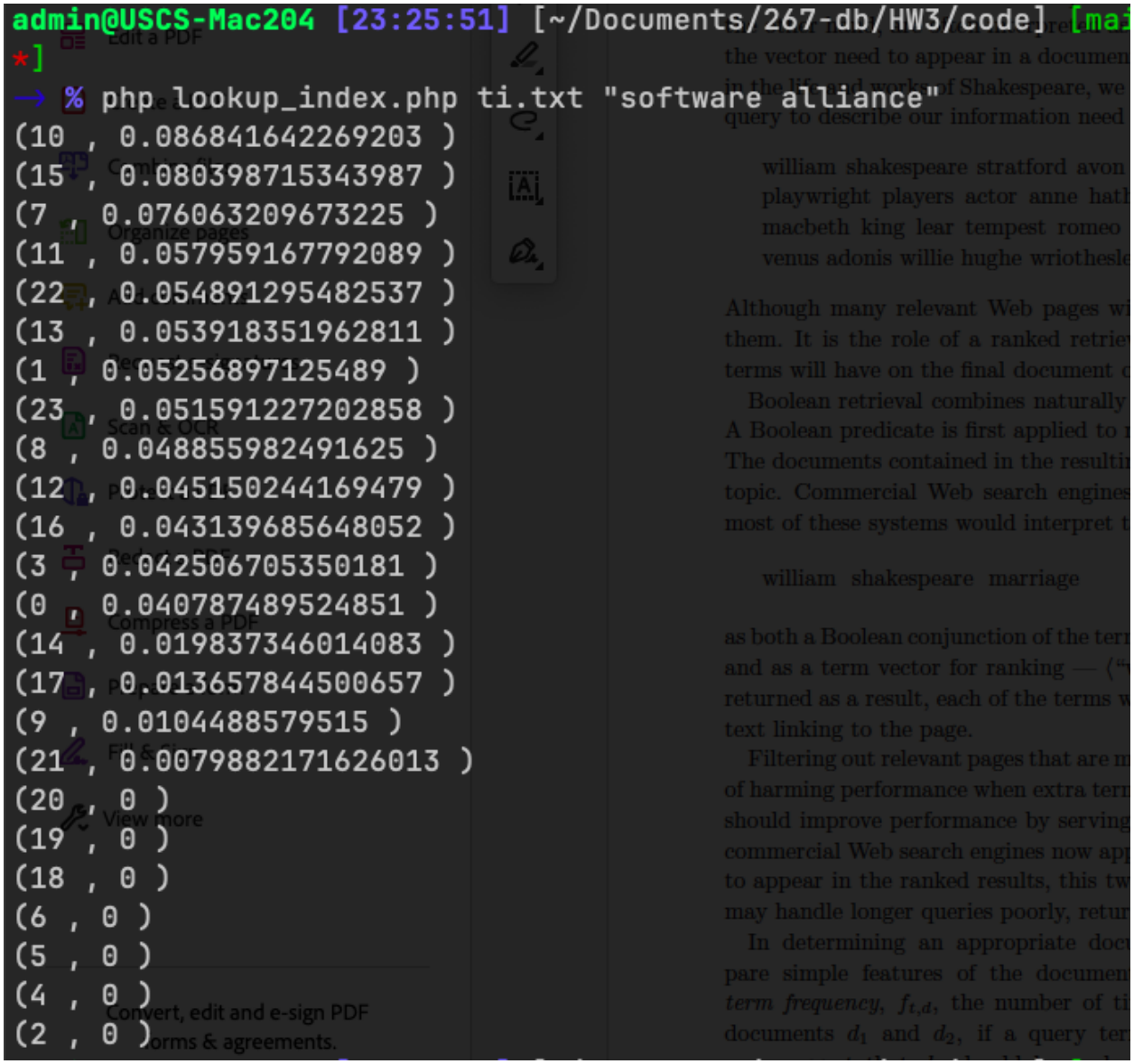
Output of lookup_index.php:

Query: business software alliance
Total number of documents: 25
Number of relevant documents: 25

Precision at 1 = 1/1 = 1
Precision at 2 = 2/2 = 1
….
Precision at 25 = 25/25 = 1
Thus, mean average precision = (sum of precision over all documents)/number of documents

= 1

Query #2: software alliance
Output of lookup_index.php:



Total number of documents: 25
Number of relevant documents: 17
Average precision @0: 1/1 = 1
Average precision @1: 2/2 = 1
Average precision @3: 3/4
Average precision @7:4/8 = 0.5

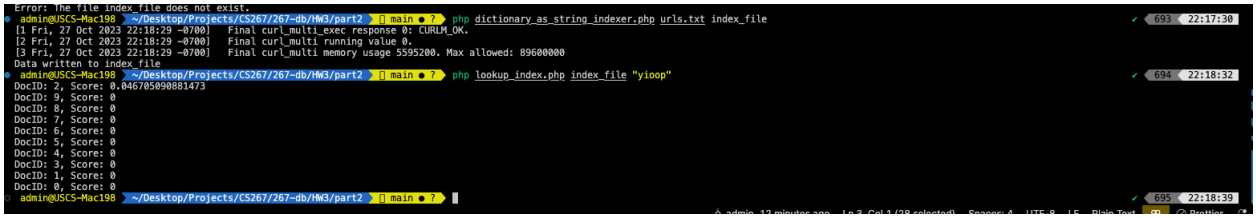
Average precision @8: 5/9
Average precision @9: 6/10
Average precision @10:7/11
Average precision @11:8/12
Average precision @12:9/13
Average precision @13:10/14
Average precision @14:11/15
Average precision @15:12/16
Average precision @16:13/17
Average precision @17:14/18
Average precision @21:15/22
Average precision @22:16/23
Average precision @23:17/24
Average precision @24:18/25

Mean Average Precision: Sum of these average precisions/total number of documents
= 11.9467999477/17
=0.7027529381

Similarly, we can calculate MAP for all the queries

https://www.business.att.com/products/business-wifi.html
https://bigdata.ieee.org/
And for the corpus:

https://www.yioop.com/s/news
https://en.wikipedia.org/wiki/Beothuk_language
https://www.economist.com/business/2021/04/14/ceo-activism-in-america-is-risky-business
https://www.wired.com/story/opinion-hackers-used-to-be-humans-soon-ais-will-hack-humanity/
https://www.creativebloq.com/how-to/delete-an-instagram-account
https://www.theverge.com/22786305/how-to-delete-restore-instagram-account-web-app-temporarily-suspend
https://www.adamenfroy.com/how-to-download-youtube-videos
https://www.pcmag.com/how-to/how-to-download-youtube-videos



We can see that yioop is only found in https://www.yioop.com/s/news, which makes sense since no other document has it, which implies that the code computes the similarity correctly