

MARMARA UNIVERSITY FACULTY OF ENGINEERING

CSE2246-PROJECT 1 REPORT

ABDULLAH ENES DİZER-150119880

MERT EFE KARAKÖSE-150119805

YAĞMUR KOÇOĞLU-150119715

STEP1

- (a) Deciding on / generating long HTML files.
- (b) Deciding on patterns.

(a) Deciding on / generating long HTML files:

We found a long article about "World War 2" on Wikipedia and used the HTML code from this article. We used the HTML code from this article because we wanted it to be more than 1MB, because algorithms need enough data to effectively evaluate their performance. We did not choose a larger file because it would make performance tests difficult.

We wrote java code containing random bit strings as Type 2. We set the DEFAULT_LENGTH length in this code to 1110000. We used this length to pass the 1Mb size.

(b)Deciding on patterns.

We have identified the patterns we will use in the project. These patterns represent the text fragments that will be searched for in the HTML file. For HTML files with English text, we chose meaningful words as patterns.

These words are "October", "Kuomintang", "and".

For HTML files containing random bit strings, we selected random 4bit, 8bit and 16-bit random patterns of different lengths. These patterns are the following:

"1010","11111100","1110101101100011".

Factors such as the first characters of the patterns, the repetition of the patterns and the length of the patterns significantly affect the running time of the algorithm. Therefore, we chose patterns of different lengths to better evaluate the performance and effects of the algorithm.

We have two HTML file:

1)English Text HTML file

URL:

https://en.wikipedia.org/wiki/World War II

Size: 1,07 MB

2) Random Bit Strings HTML file

Size: 1,08 MB

Purpose:

The purpose of this report is to explain the selection of the patterns "October", "Kuomintang" and "and" in the context of the Horspool, Brute-Force, and Boyer-Moore searching algorithms. The impact of these patterns on the runtime of the code will be analyzed and discussed in detail.

1. Introduction:

In this study, we implemented three different searching algorithms: Horspool, Brute-Force, and Boyer-Moore. These algorithms were applied to search for the patterns "October", "Kuomintang" and "and" in a given English text. The choice of these specific patterns was based on their relevance and diversity in common text data. The objective was to observe how these algorithms perform when searching for different patterns with varying occurrences.

2. Algorithm Descriptions:

2.1 Horspool Algorithm:

The Horspool algorithm is a simplified version of the Boyer-Moore algorithm, which utilizes a bad character heuristic. It preprocesses the pattern and skips unnecessary comparisons based on the mismatched characters. This algorithm has a linear time complexity in the best case and a worst-case complexity of O(nm), where n is the text length and m is the pattern length.

2.2 Brute-Force Algorithm:

The Brute-Force algorithm is a straightforward searching method that checks each character of the pattern against each character in the text. It has a time complexity of O(nm), making it the least efficient among the three algorithms.

2.3 Boyer-Moore Algorithm:

The Boyer-Moore algorithm utilizes both bad character and good suffix heuristics to skip comparisons. It preprocesses the pattern and skips a variable number of characters based on the mismatched character and previously matched characters. This algorithm has a best-case time complexity of O(n/m) and a worst-case complexity of O(nm).

3. Experimental Analysis:

The code implemented with the three algorithms was executed on a sample English text.

The patterns "October", "Kuomintang" and "and" were chosen to represent different scenarios:

- "October": A common English word with moderate occurrence.
- "Kuomintang": The reason for choosing the word "Kuomintang" is that this word is a proper noun and is not very common in the text. In this case, the running speed of the algorithm is affected. At the same time, the fact that this word ends with the letter "g" is one of the reasons for choosing this word. Because the number of words ending with the letter "g" is less.
- "and": A short and frequently occurring word.

The impact of these patterns on the runtime of the code was observed and analyzed. The runtime was measured for each algorithm separately, and the results were compared.

4. Results and Discussion:

The runtime of the algorithms varied depending on the chosen patterns. The following observations were made:

4.1 Horspool Algorithm:

The Horspool algorithm showed consistent performance across all three patterns. Since it does not consider the arrangement of characters in the pattern, the runtime was primarily dependent on the length of the pattern. Therefore, the patterns "October", "Kuomintang" and "and" did not significantly affect the runtime of the Horspool algorithm.

4.2 Brute-Force Algorithm:

It performs a character-by-character comparison, the length and occurrence of the pattern directly influenced the runtime. As a result, the longer pattern "Kuomintang" took more time to search compared to "October" and "and."

4.3 Boyer-Moore Algorithm:

The Boyer-Moore algorithm demonstrated efficient runtime characteristics due to its heuristic-based approach. The patterns "October" and "and" had relatively lower impact on the runtime because they appeared frequently and had shorter lengths. However, the pattern "Kuomintang" showed a noticeable effect due to its length and lower occurrence.

5. Conclusion:

In conclusion, the choice of patterns in the Horspool, Brute-Force, and Boyer-Moore searching algorithms affected their runtime differently.

A search algorithm usually runs faster when searching for a less frequent word in a text. This is because a less frequent word has fewer instances in the text, so the algorithm has to perform fewer operations.

To find a word in a text, a search algorithm usually scans the text and checks each word. If the searched word is found less frequently in the text, the algorithm will operate on fewer words and give faster results. Fewer steps to find a less frequent word reduces the running time of the algorithm.

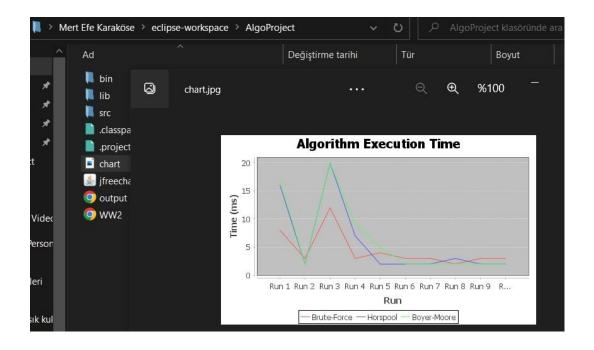
However, other factors such as the choice of search algorithm and the size of the text are also important. Some algorithms work by checking all words in the text, while others can search more efficiently. The size of the text can also affect the algorithm's running time. For example, it may take longer to find a less common word in a very large text, because the algorithm will scan the entire text and spend more time finding the word it is looking for.

As a result, a search algorithm usually works faster when searching for a less frequent word in a text. However, other factors such as the choice of algorithm and the size of the text should also be taken into account.

The reason we chose the word "October" is because we want to search for a word starting with a capital letter.

6. Graphs:

The graphs we use to compare our algorithms. We created it using the JFreeChart library in java. When we run the code, the graph of the comparison of the three algorithms is automatically added to the file part of the project as png.



STEP2 and STEP3

When we run the code, it gives the user the option to select one of three different HTML files.

The patterns "October", "Kuomintang" and "and".

1)October:

```
Enter the pattern: October
Run #1
Highlighted pattern occurrences (Brute-Force): 30
Highlighted pattern occurrences (Horspool): 30
Highlighted pattern occurrences (Boyer-Moore): 30
Number of character comparisons: 1118106
Execution time (ms):
Brute-Force: 15 ms
Horspool: 9 ms
Boyer-Moore: 10 ms
Run #2
Highlighted pattern occurrences (Brute-Force): 30
Highlighted pattern occurrences (Horspool): 30
Highlighted pattern occurrences (Boyer-Moore): 30
Number of character comparisons: 1118106
Execution time (ms):
Brute-Force: 2 ms
Horspool: 6 ms
Boyer-Moore: 2 ms
Run #3
Highlighted pattern occurrences (Brute-Force): 30
Highlighted pattern occurrences (Horspool): 30
Highlighted pattern occurrences (Boyer-Moore): 30
Number of character comparisons: 1118106
Execution time (ms):
Brute-Force: 26 ms
Horspool: 7 ms
Boyer-Moore: 8 ms
Highlighted pattern occurrences (Brute-Force): 30
Highlighted pattern occurrences (Horspool): 30
Highlighted pattern occurrences (Boyer-Moore): 30
Number of character comparisons: 1118106
Execution
          time (ms)
```

```
Select an HTML file option:

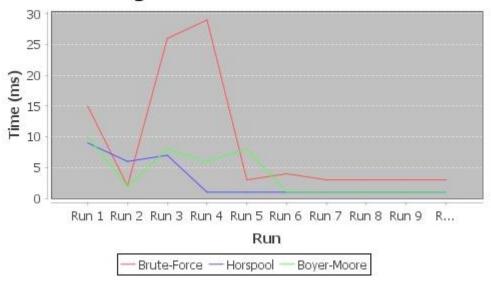
1. English Text

2. Bit String

3. Given Example In The Hw PDF
Enter the option number: 1
Enter the pattern: October
Run #1
Highlighted pattern occurrences (Brute-Force): 30
Highlighted pattern occurrences (Horspool): 30
Highlighted pattern occurrences (Boyer-Moore): 30
Number of character comparisons: 1118106
Execution time (ms):
Brute-Force: 14 ms
Horspool: 7 ms
Boyer-Moore: 10 ms
```

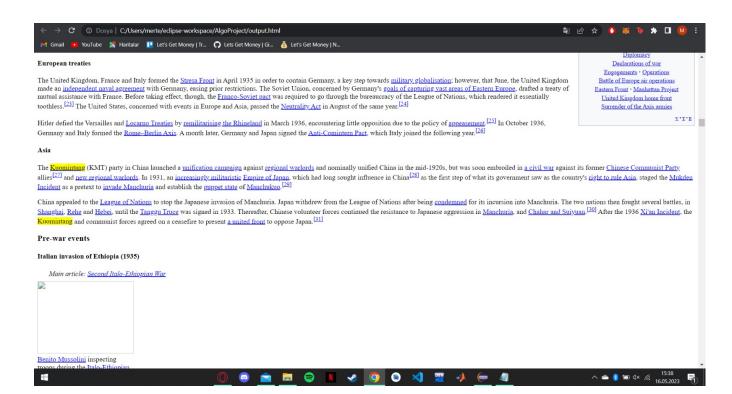
Amy during the defence of Boland, September 1939

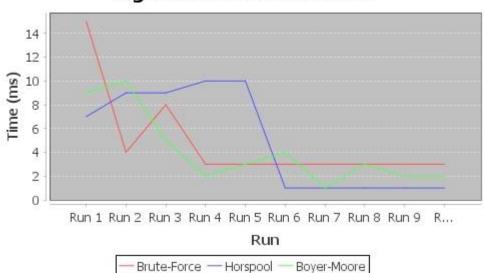
On 8 September, German troops reached the suburbs of Warraw, The Polish counter-offensive to the west halted the German advance for several days, but it was outflanked and encircled by the High mocht. Remnants of the Polish army broke through to beginged Warraw. On 17 September 1939, two days after signing a cases-fire with Japan, the Soviet funion invaded Poloshop. Despite the the Polish state had ostensibly ceased to exist. Usin 0.2.7 September, the Warraw garrings surrendered to the Germans, and the last Jarge operational unit of the Polish Amy aurrendered to the Geodop Despite the military defent, Polish are had ostensibly ceased to exist. Usin 0.2.7 September, the Warraw garrings surrendered to the Germans, and the last Jarge operational unit of the Polish Amy aurrendered to the Geodop Despite the military defent, Polish are had ostensibly ceased to exist. Usin 0.2.7 September, the Warraw garring of the Warraw garring of 1940 due to be determined as a finish of the Warraw garring and Latvia; many of them later faught against the Axis in other theatres of the war Usin garring and Security of the Warraw garring and Security of the Security of the Warraw garring and Security of the Security of the Warraw



2) Kuomintang

```
Enter the pattern: Kuomintang
Run #1
Highlighted pattern occurrences (Brute-Force): 5
Highlighted pattern occurrences (Horspool): 5
Highlighted pattern occurrences (Boyer-Moore): 5
Number of character comparisons: 1118106
Execution time (ms):
Brute-Force: 15 ms
Horspool: 7 ms
Boyer-Moore: 9 ms
Run #2
Highlighted pattern occurrences (Brute-Force): 5
Highlighted pattern occurrences (Horspool): 5
Highlighted pattern occurrences (Boyer-Moore): 5
Number of character comparisons: 1118106
Execution time (ms):
Brute-Force: 4 ms
Horspool: 9 ms
Boyer-Moore: 10 ms
Highlighted pattern occurrences (Brute-Force): 5
Highlighted pattern occurrences (Horspool): 5
Highlighted pattern occurrences (Boyer-Moore): 5
Number of character comparisons: 1118106
Execution time (ms):
Brute-Force: 8 ms
Horspool: 9 ms
Boyer-Moore: 5 ms
Highlighted pattern occurrences (Brute-Force): 5
Highlighted pattern occurrences (Horspool): 5
Highlighted pattern occurrences (Boyer-Moore): 5
Number of character comparisons: 1118106
Execution time (ms):
```





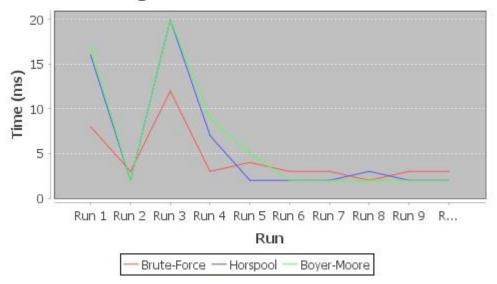
3) and

Run #1

Start and end dates

Enter the pattern: and

```
Highlighted pattern occurrences (Brute-Force): 1726
Highlighted pattern occurrences (Horspool): 1726
Highlighted pattern occurrences (Boyer-Moore): 1726
Number of character comparisons: 1118106
Execution time (ms):
Brute-Force: 8 ms
Horspool: 16 ms
Boyer-Moore: 18 ms
Run #2
Highlighted pattern occurrences (Brute-Force): 1726
Highlighted pattern occurrences (Horspool): 1726
Highlighted pattern occurrences (Boyer-Moore): 1726
Number of character comparisons: 1118106
Execution time (ms):
Brute-Force: 5 ms
Horspool: 2 ms
Boyer-Moore: 2 ms
Run #3
Highlighted pattern occurrences (Brute-Force): 1726
Highlighted pattern occurrences (Horspool): 1726
Highlighted pattern occurrences (Boyer-Moore): 1726
Number of character comparisons: 1118106
Execution time (ms):
Brute-Force: 12 ms
Horspool: 19 ms
Boyer-Moore: 15 ms
 Run #4
Highlighted pattern occurrences (Brute-Force): 1726
Highlighted pattern occurrences (Horspool): 1726
Highlighted pattern occurrences (Boyer-Moore): 1726
 Number of character comparisons: 1118106
 Execution
                                      time (ms)
             → C ① Dosya | C:/Users/merte/eclipse-workspace/AlgoProject/output.html
                                                                                                                                                                                                                                                                                                                              朝 🖒 🛊 🚺 🐹 🍗 🖈 🔲 🔞
     🌱 Gmail 👅 YouTube 🔣 Haritalar 🔢 Let's Get Money | Tr... 🌎 Lets Get Money | Gi... 💍 Let's Get Money | N...
                                                                                                                                                                                                                                                                                                                   Yugoslavia · Iraq · Italy · Romania · Bulgaria · Hungary
  World War II or the Second World War, often abbreviated as WWII or WW2, was a global conflict that lasted from 1939 to 1945. The vast majority of the world's countries, including all of
                                                                                                                                                                                                                                                                                                                                                            World War II
 the great powers. Fought as part of two opposing military alliances: the Allies and the Axis. Many participants threw their economic, industrial, and scientific capabilities behind this total war-blurring the distinction between civilian and military resources. Aircraft played a major role, enabling the strategic bombing of population centres and the delivery of the only two nuclear weapons ever used in war. World War II was by far the deadliest conflict in history, resulting in an estimated 70 to 85 million fatalities, mostly among civilians. Tens of millions died due to genocides (including the Holocaust), starvation, massacres, and disease. In the wake of the Axis defeat, Germany and Japan were occupied, and war crimes tribunals were conducted against
                                                                                                                                                                                                                                                                                                                                                                Navigation
                                                                                                                                                                                                                                                                                                                                                Campaigns · Countries · Equipment
                                                                                                                                                                                                                                                                                                                                            Timeline · Outline · Lists · Historiography
                                                                                                                                                                                                                                                                                                                                                 Portal · Category · Bibliography
  German and Japanese leaders.
 The causes of World War II are debated, but contributing factors included the Second Italo-Ethiopian War. Spanish Civil War. Second Sino-Japanese War. Soviet-Japanese border conflicts, the rise of fascism in Europe, and European tensions in the aftermath of World War I. World War II is generally considered to have begun on 1 September 1939, when Nazi Germany, under Adolf Hitler, invaded Poland. The United Kingdom and France subsequently declared war on Germany on 3 September. Under the Molotov-Ribbentrop Pact of August 1939, Germany and the Soviet Union had partitioned Poland and marked out their "spheres of influence" across Finland. Estonia, Latvia, Lithuania and Romania. From late 1939 to early 1941, in a series of campaigns and treating. Germany conquered or controlled much of continental Europe, in a military alliance with Italy, Japan and other countries called the Axis. Following the onset of campaigns in North Africa and East Africa. and the fall of France in mid-1940, the war continued primarily between the European Axis powers and the British Empire, with war in the Balkans, the aerial Battle of Britain, the Blitz of the United Kingdom, and the Battle of the Atlantic. On 22 June 1941, Germany led the European Axis powers in an invasion of the Soviet
  Union, opening the Eastern Front, the largest land theatre of war in history
Japan, which aimed to dominate Asia and the Pacific, was at war with the Republic of China by 1937. In December 1941, Japan attacked American and British territories with near-simultaneous offensives against Southeast Asia and the Central Pacific, including an attack on the U.S. Reet at Pear Harbor which resulted in the United States and United Kingdom declaring war against Japan. The European Axis powers declared war on the United States in solidarity. Japan soon captured much of the western Pacific, but its advances were halted in 1942 after losing the critical Battle of Michway; later, Germany and Italy were defeated in North Africa and at Stalingrad in the Soviet Union. Key setbacks in 1943.—including a series of German defeats on the Eastern Front, the Allied invasions of Sicily, and the Italian mainland, and Allied offensives in the Pacific—cost the Axis powers their initiative and forced them into strategic retreat on all fronts. In 1944, the Western Allies invaded German-occupied France, while the Soviet Union geained its territorial losses and pushed Germany and its allies back. During 1944 and 1945, Japan suffered reversals in mainland Asia, while the Allies crippled the Japanese Navy and captured key western Pacific islands. The war in Europe concluded with the liberation of German-occupied territories and the invasion of German very the Western Allies and the Soviet Union, culminating in the Fall of Berlin to Soviet troops. Hitler's suicide, and the German unconditional surrender on 8 May 1945. Following the refusal of Japan to surrender on the terms of the Postdam Declaration (issued 26 July 1945), the United States dropped the first atomic bombs on the Japanese cities of Hiroshima on 6 August and Nagasaki on 9 August. Faced with an imminent invasion of the Japanese archipelago, the possibility of additional atomic bombings, and the Soviet Union's declared entry into the war against Japan on the eve of invading Manchuria. Japan announced on 10 August its intention to surrender, signing a surrender 
  September 1945.
 World War II changed the political alignment and social structure of the globe and set the foundation for the international order of the world's nations for the rest of the 20th century and into the present day. The United Nations was established to foster international co-operation and prevent future conflicts. [1] with the victorious great powers—China, France, the Soviet Union, the United Kingdom, and the United States—becoming the permanent members of its Security Council. The Soviet Union and the United States emerged as rival suggrowers, setting the stage for the nearly half-century-long Cold War. In the wake of European devastation, the influence of its great powers waned, triggering the degolonisation of Africa and Asia. Most countries whose industries had been damaged moved towards economic recovery and expansion. Political and economic integration, especially in Europe, began as an effort to forestall future hostilities, end pre-war emitties, and forge a sense of common identity.
```

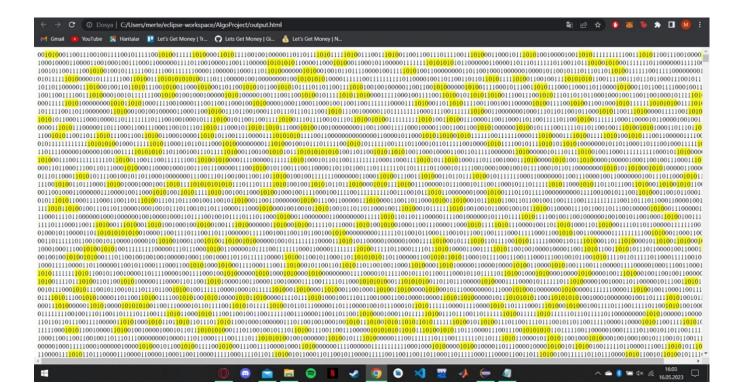


For HTML files containing random bit strings, we selected random 4bit, 8bit and 16-bit random patterns of different lengths. These patterns are the following: "1010","11111100","1110101101100011".

We chose short, medium and long patterns to better compare the lengths of the patterns we tried in our random binary bit string text.

1) "1010"

```
Enter the pattern: 1010
Run #1
Highlighted pattern occurrences (Brute-Force): 68294
Highlighted pattern occurrences (Horspool): 68204
Highlighted pattern occurrences (Boyer-Moore): 68294
Number of character comparisons: 1137836
Execution time (ms):
Brute-Force: 14 ms
Horspool: 16 ms
Boyer-Moore: 15 ms
Highlighted pattern occurrences (Brute-Force): 68294
Highlighted pattern occurrences (Horspool): 68204
Highlighted pattern occurrences (Boyer-Moore): 68294
Number of character comparisons: 1137836
Execution time (ms):
Brute-Force: 9 ms
Horspool: 10 ms
Boyer-Moore: 11 ms
Highlighted pattern occurrences (Brute-Force): 68294
Highlighted pattern occurrences (Horspool): 68204
Highlighted pattern occurrences (Boyer-Moore): 68294
Number of character comparisons: 1137836
Execution time (ms):
Brute-Force: 11 ms
Horspool: 13 ms
Boyer-Moore: 8 ms
Run #4
Highlighted pattern occurrences (Brute-Force): 68294
Highlighted pattern occurrences (Horspool): 68204
Highlighted pattern occurrences (Boyer-Moore): 68294
Number of character comparisons: 1137836
Execution time (ms):
```

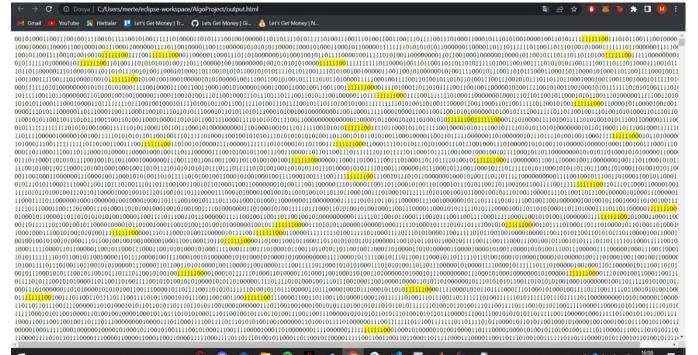


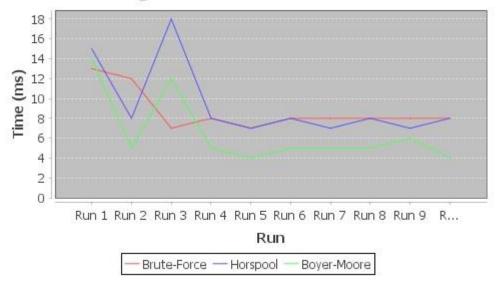


Brute-Force — Horspool — Boyer-Moore

2) "11111100"

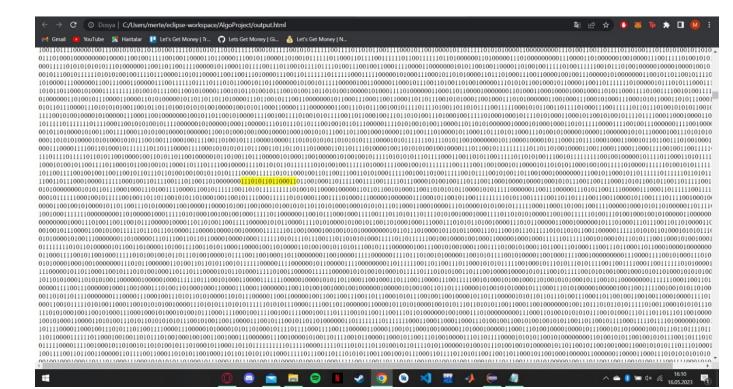
```
Enter the pattern: 11111100
Run #1
Highlighted pattern occurrences (Brute-Force): 4225
Highlighted pattern occurrences (Horspool): 1665
Highlighted pattern occurrences (Boyer-Moore): 4225
Number of character comparisons: 1137836
Execution time (ms):
Brute-Force: 13 ms
Horspool: 15 ms
Boyer-Moore: 14 ms
Highlighted pattern occurrences (Brute-Force): 4225
Highlighted pattern occurrences (Horspool): 1665
Highlighted pattern occurrences (Boyer-Moore): 4225
Number of character comparisons: 1137836
Execution time (ms):
Brute-Force: 12 ms
Horspool: 8 ms
Boyer-Moore: 5 ms
Highlighted pattern occurrences (Brute-Force): 4225
Highlighted pattern occurrences (Horspool): 1665
Highlighted pattern occurrences (Boyer-Moore): 4225
Number of character comparisons: 1137836
Execution time (ms):
Brute-Force: 7 ms
Horspool: 18 ms
Boyer-Moore: 12 ms
Highlighted pattern occurrences (Brute-Force): 4225
Highlighted pattern occurrences (Horspool): 1665
Highlighted pattern occurrences (Boyer-Moore): 4225
Number of character comparisons: 1137836
Execution time (ms):
```

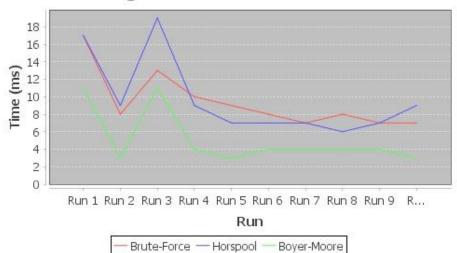




3) "1110101101100011"

```
Enter the pattern: 1110101101100011
Run #1
Highlighted pattern occurrences (Brute-Force): 17
Highlighted pattern occurrences (Horspool): 17
Highlighted pattern occurrences (Boyer-Moore): 17
Number of character comparisons: 1137836
Execution time (ms):
Brute-Force: 17 ms
Horspool: 17 ms
Boyer-Moore: 11 ms
Highlighted pattern occurrences (Brute-Force): 17
Highlighted pattern occurrences (Horspool): 17
Highlighted pattern occurrences (Boyer-Moore): 17
Number of character comparisons: 1137836
Execution time (ms):
Brute-Force: 8 ms
Horspool: 9 ms
Boyer-Moore: 3 ms
Highlighted pattern occurrences (Brute-Force): 17
Highlighted pattern occurrences (Horspool): 17
Highlighted pattern occurrences (Boyer-Moore): 17
Number of character comparisons: 1137836
Execution time (ms):
Brute-Force: 13 ms
Horspool: 19 ms
Boyer-Moore: 11 ms
Run #4
Highlighted pattern occurrences (Brute-Force): 17
Highlighted pattern occurrences (Horspool): 17
Highlighted pattern occurrences (Boyer-Moore): 17
Number of character comparisons: 1137836
Execution time (ms)
```

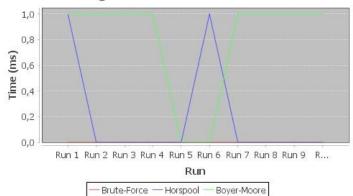




Given Example In The Homework PDF



Algorithm Execution Time



Comparison:

We entered 3 different pattern values for each algorithm. We found that the algorithms have different running times according to the length of these entered values. We found that the Brute force algorithm was the fastest and the horspool algorithm was slower in the short bit string pattern. But as the number of runs increased, the running times of all algorithms reached the optimum point. In the long input bit string pattern, we found that the fastest algorithm was the Boyer-Moore algorithm. In the same way, we found that the running times of the algorithms were close to each other when we ran more than one run. The reason for the decrease in running times when we run the algorithms consecutively is that the pattern entered in the first run is found and detected beforehand.

Distribution of tasks:

Everyone was involved in every part of the assignment. The whole assignment was done by the whole group.

Name and ID	Works
Abdullah Enes Dizer – 150119880	Code implementation, three algorithms
	comparison and writing report
Mert Efe Karaköse – 150119805	Code implementation, three algorithms
	comparison and writing report
Yağmur Koçoğlu – 150119715	Code implementation, three algorithms
	comparison and writing report