**LAPORAN TUGAS KECIL 4**

**EKSTRAKSI INFORMASI DARI ARTIKEL BERITA DENGAN ALGORITMA PENCOCOKAN STRING**

**IF2211 Strategi Algoritma**

**­­**

Oleh:

**Daffa Pratama Putra 13518033**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**INSTITUT TEKNOLOGI BANDUNG**

**2020**

# **BAB I ALGORITMA PENCOCOKAN STRING**

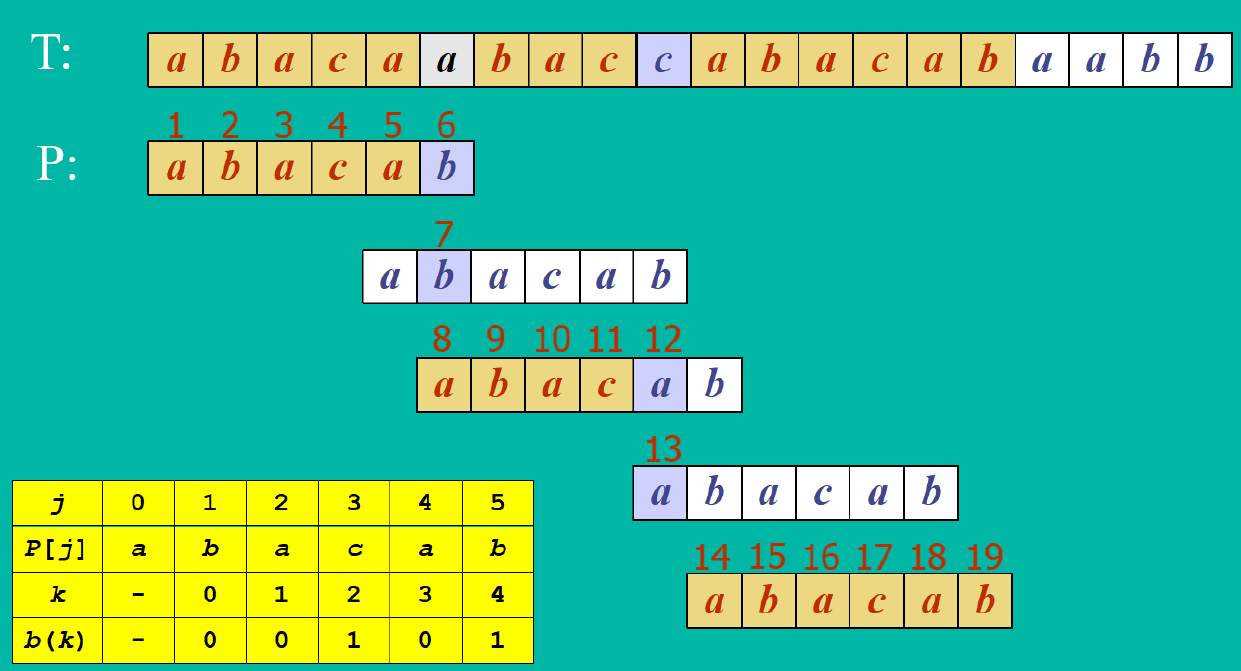
1. **Algoritma Knuth-Morris-Pratt (KMP)**

Algoritma Knuth-Morris-Pratt adalah salah satu algoritma pencocokan *string* yang dinilai lebih baik dari algoritma Brute Force. Algoritma KMP ini dikembangkan oleh salah satu *computer scientist* dari Standford University, Donald E. Knuth pada tahun 1967. Setahun sebelumnya, James H. Morris dan Vaughan R. Pratt juga mengembangkan algoritma yang sama, oleh karena itu nama dari algoritma ini diambil dari nama ketiga penciptanya, yakni Knuth, Morris, dan Pratt.

Pada dasarnya, algoritma ini mirip seperti algoritma Brute Force dalam pencocokan *string*, yaitu melakukan pencocokan dari kiri ke kanan *string*. Namun ada hal yang berbeda. Pada algoritma Brute Force, pergeseran dalam pencocokan *string* dilakukan satu persatu dari kiri ke kanan, baik karakter yang dicocokkan sama atau berbeda. Pergeseran satu persatu dilakukan hingga pencocokan karakter sudah mencapai akhir atau ditemukan *string* yang dicari oleh *pattern* pada *text*. Berbeda dengan algoritma Brute Force, algoritma KMP melakukan pergeseran dalam pencocokan *string* dengan mempertimbangkan *pattern* yang telah dibandingkan. Sehingga pada algoritma KMP dibutuhkan *pre-processing* untuk menentukan jumlah pergeseran yang dilakukan untuk setiap indeks pada *pattern* atau yang disebut dengan fungsi pinggiran.

Fungsi pinggiran atau *failure function* pada algoritma KMP adalah fungsi yang menghitung jumlah pergeseran yang dilakukan dengan menghitung panjang terbesar prefiks dari *pattern,* P[0..k], yang merupakan sufiks dari pattern tersebut, P[1..k], dengan k = (indeks *mismatch* pada pattern) - 1. Dengan demikian, proses pergeseran akan menjadi lebih efisien karena karakter yang sudah dicocokkan pada prefiks tidak dicocokkan kembali. Algoritma KMP menghasilkan kompleksitas waktu O(m+n) dengan O(m) adalah kompleksitas menghitung fungsi pinggiran dan O(n) adalah kompleksitas pencarian *string*.

Berikut adalah contoh pencarian menggunakan algoritma KMP :

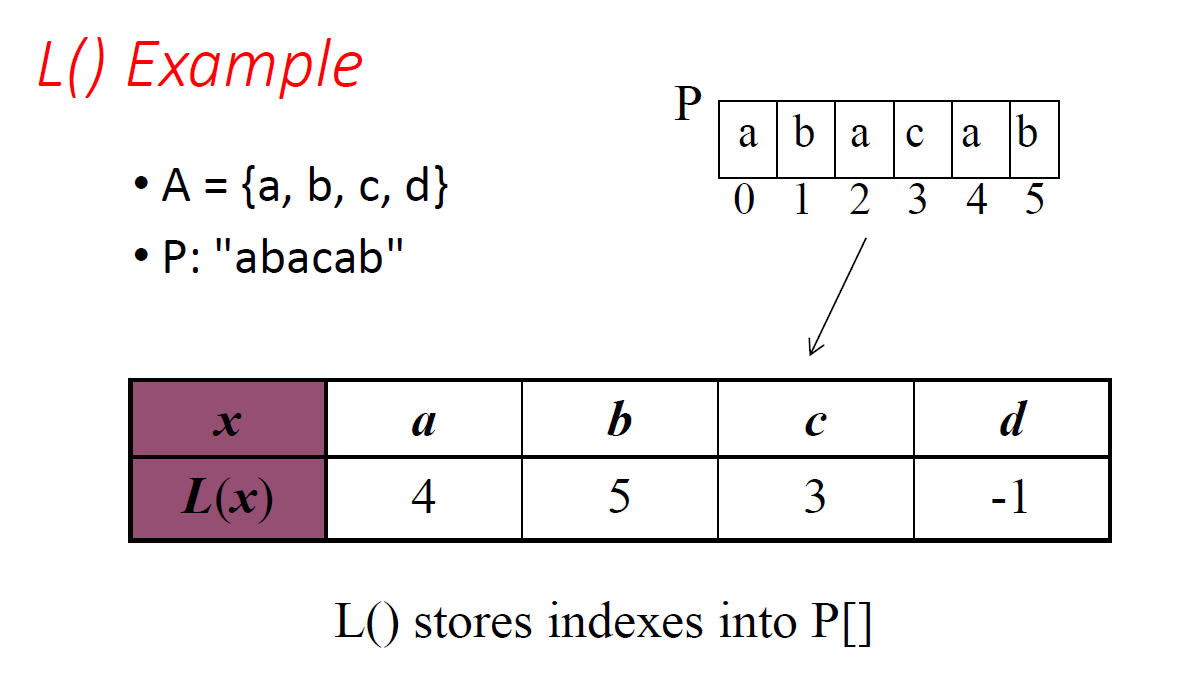


Gambar 1. Pencocokan string dengan algoritma KMP (diambil dari Bahan Kuliah IF2211 Strategi Algoritma, Pencocokan String (String/Pattern Matching)

1. **Algoritma Boyer-Moore (BM)**

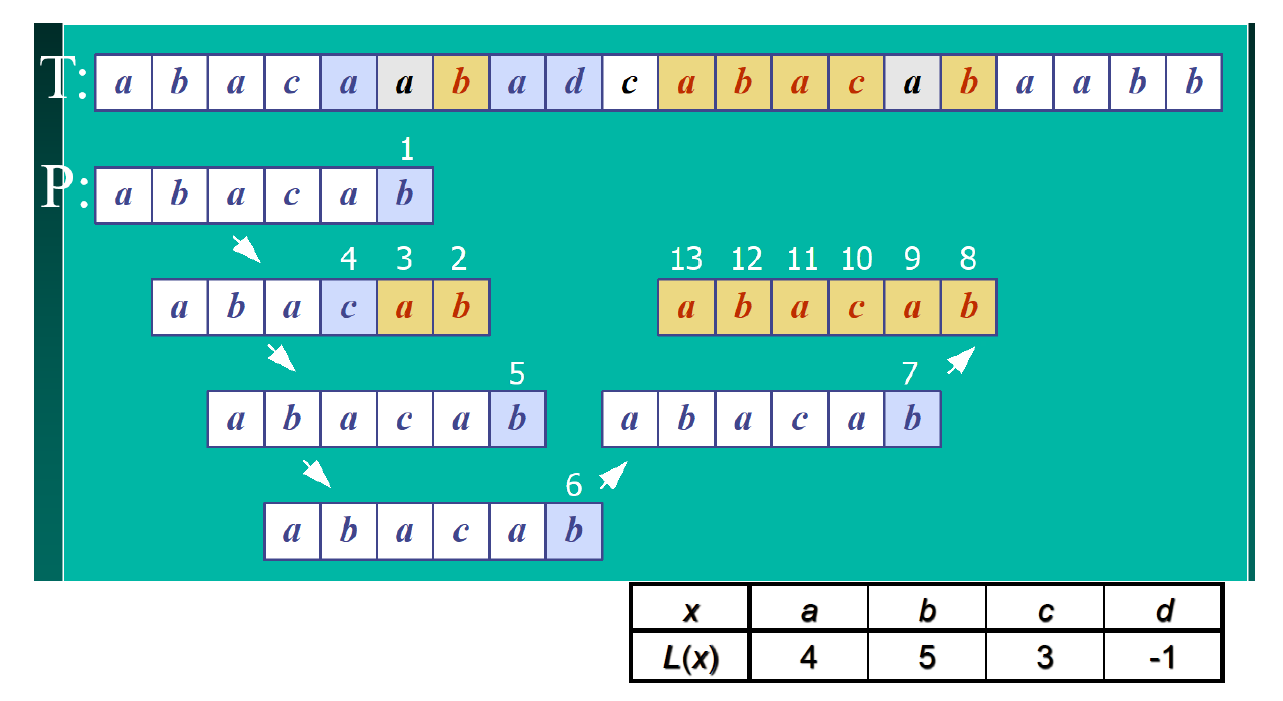
Algoritma Boyer-Moore adalah salah satu algoritma pencocokan *string* dengan teknik yang cukup unik namun dinilai paling efisien. Algoritma Boyer-Moore menggunakan 2 teknik, yaitu *looking-glass technique* dan *character-jump technique*. Teknik *looking-glass* ini akan mencocokkan *pattern* dengan *text* dengan pengecekan mundur. Pengecekan mundur seperti ini efisien karena dapat memprediksi pergeseran yang akan dilakukan selanjutnya secara optimal. Lalu teknik *character-jump* adalah teknik pergeseran yang dilakukan pada algoritma Boyer-Moore dengan mempertimbangkan lokasi kemunculan karakter yang *mismatch* antara *pattern* dengan *text*. Terdapat tiga kasus dalam teknik *character-jump* ini, yaitu karakter yang *mismatch* pada *text* ditemukan di sebelah kiri karakter *mismatch* pada *pattern*, karakter yang *mismatch* pada *text* ditemukan di sebelah kanan karakter *mismatch* pada *pattern*, dan karakter yang *mismatch* pada *text* tidak terdapat pada *pattern*.

Sama seperti algoritma Knuth-Morris-Pratt, algoritma Boyer-Moore juga membutuhkan *pre-processing* untuk mendapatkan lokasi kemunculan karakter yang ada pada *text*. *Pre-processing* pada algoritma ini adalah digunakan untuk mencari lokasi *last occurance* dari setiap karakter pada *text*. Jika karakter ada pada *pattern*, maka nilai *last occurance*nya adalah indeks terbesar kemunculan karakter tersebut pada *pattern* P, namun jika tidak ditemukan, maka nilai *last occurance*nya adalah -1. Contoh mencari nilai *last occurance* adalah sebagai berikut :



Gambar 2. Fungsi Last Occurance pada algoritma Boyer-Moore (diambil dari Bahan Kuliah IF2211 Strategi Algoritma, Pencocokan String (String/Pattern Matching)

*Last occurance* ini berguna untuk menentukan berapa banyak pergeseran yang akan dilakukan. Misalkan *mismatch* terjadi pada karakter X. Karakter X akan dicek apakah ada pada *pattern* menggunakan *last occurance*. Jika berada di sebelah kiri indeks karakter *mismatch*, maka *pattern* digeser ke kanan sehingga karakter yang *mismatch* tadi dapat benar dicocokkan dengan karakter pada *pattern*. Berikut adalah contoh pencocokan *string* dengan algoritma Boyer-Moore :



Gambar 3. Pencocokan string dengan algoritma Boyer-Moore (diambil dari Bahan Kuliah IF2211 Strategi Algoritma, Pencocokan String (String/Pattern Matching)

1. **Regular Expression (Regex)**

Regular Expression atau biasa disebut regex adalah salah satu metode pencocokan *string* yang unggul. Metode regex dinilai unggul karena dapat melakukan pencarian dengan banyak pola sekaligus. Misalkan saja ingin mencari kata berawalan “Sa”, memiliki huruf “ra”, kata berakhiran “GA”, atau bahkan mencocokkan penulisan tanggal. Pencarian dalam regex lebih menekankan pada pola pencarian yang diberikan, sehingga banyak modifikasi pola-pola yang dapat dilakukan. Beberapa bahasa pemrograman sudah menyediakan library regex, seperti pada Python, Javascript, Java, dan PHP. Bahkan ada beberapa situs yang menyediakan pencocokan *string* menggunakan metode regex, seperti regextester.com dan pythex.org. Untuk memahami regex cukup sulit, karena harus mencari pola yang benar-benar ingin dicari. Ditambah lagi penulisan pola pada regex cukup kompleks karena menggunakan simbol-simbol yang sebelumnya cukup jarang ditemukan.

Pada algoritma KMP dan BM, masukan kata atau *pattern* yang dicari haruslah eksak, sehingga hasil yang diberikan hanya kata-kata yang sesuai dengan *pattern* masukan. Namun pada metode regex, *pattern* yang dimasukkan dapat berupa *pattern* yang eksak atau pola yang ingin dicari. Misalkan pola yang dimasukkan adalah ‘^([01][0-9] | 2[0-3]):[0-5][0-9]:[0-5][0-9]’. Maka regex akan mencari semua kata yang terdapat pada *text* yang sesuai dengan pola di atas. Contoh teks yang sesuai adalah 23:59:59 atau 06:15:05. Keduanya merupakan hasil yang benar, karena masukannya berupa pola, bukan *pattern* yang eksak.

# **BAB II KODE PROGRAM**

1. **Kode Program**

Program dibuat menggunakan bahasa Python dan PHP. Python digunakan sebagai *backend* sedangkan PHP menangani tampilan *website*, pencetakan hasil, dan pemilihan *text file*. Program terdiri dari *file* main.py, kmp.py, bm.py, reg.py, textReader.py, index.php, dan submit.php. *File* kmp.py, bm.py, dan reg.py berisi algoritma pencarian *keyword* yang dimasukkan oleh pengguna melalui *website*. *File* textReader.py berisi kode program untuk membaca teks dari *file* yang di-*upload* oleh pengguna dari website. Kemudian *file* main.py berisi program utama yang memroses keseluruhan pencocokan *keyword*, ekstraksi informasi, hingga mencetak hasil ekstraksi ke layar. *File* index.php bertanggung jawab dalam penampilan *website*, termasuk pemberian *form,* sedangkan submit.php menangani *upload* *file* dan menyimpannya dalam basis data lokal. Berikut adalah kode program untuk setiap *file* di atas :

|  |
| --- |
| File kmp.py (src/kmp.py) |
| class kmp :      def \_\_init\_\_(self):  *pass*      def failureFunction(self, pattern) :          failure = [0] \* len(pattern)          m = len(pattern)          i = 1          j = 0  *while* (i < m) :  *if* (pattern[i] == pattern[j]) :                  failure[i] = j + 1                  i += 1                  j += 1  *elif* (j > 0) :                  j = failure[j-1]  *else* :                  failure[i] = 0                  i += 1  *return* failure      def match(self, text, pattern) :          m = len(pattern)          n = len(text)          i = 0          j = 0          failure = self.failureFunction(pattern)  *while* (i < n) :  *if* (text[i] == pattern[j]) :  *if* (j == (m - 1)) :  *return* i - m + 1                  i += 1                  j += 1  *elif* (j > 0) :                  j = failure[j-1]  *else* :                  i += 1    *return* -1 |

|  |
| --- |
| File bm.py (src/bm.py) |
| class bm :      def \_\_init\_\_ (self) :  *pass*      def lastOccuranceFunction (self, pattern) :          lastOccurance = [-1]\*256  *for* i in range (len(pattern)) :              lastOccurance[ord(pattern[i])] = i  *return* lastOccurance      def match(self, text, pattern) :          m = len(pattern)          n = len(text)          i = m - 1          lastOccurance = self.lastOccuranceFunction(pattern)  *if* (i > (n - 1)) :  *return* -1          j = m - 1  *while* (i <= (n - 1)):  *if* (text[i] == pattern[j]) :  *if* (j == 0) :  *return* i  *else* :                      i -= 1                      j -= 1  *else* :                  lo = lastOccurance[ord(text[i])]                  i = i + m - min(j, lo + 1)                  j = m - 1    *return* -1 |

|  |
| --- |
| File reg.py (src/reg.py) |
| *import* re  class reg :      def \_\_init\_\_(self) :  *pass*      def match(self, text, pattern) :          x = re.search(pattern, text)  *if* (x != None) :  *return* x.start()  *return* -1 |

|  |
| --- |
| File textReader.py (src/textReader.py) |
| class textReader :      def \_\_init\_\_(self, filepath) :          self.filepath = filepath          self.stream = []      def readText(self) :          f = open(self.filepath, 'r')          temp = f.readlines()  *for* i in range(len(temp)) :              self.stream.append(temp[i].strip("\n"))  *return* self.stream |

|  |
| --- |
| File main.py (src/main.py) |
| *from* textReader *import* \*  *from* kmp *import* \*  *from* bm *import* \*  *from* reg *import* \*  *from* nltk *import* tokenize  *import* sys  def searchIndex(pattern, text) :      pat = pattern.split(' ')  *for* i in range(len(text)) :          same = False  *if* (pat[0] == text[i]) :              j = 1  *while* (j < len(pat)) :  *if* (pat[j] != text[i+j]) :  *return* -1                  j += 1              same = True  *if* (same) :  *return* i  *return* -1  *# Init*  filename = sys.argv[1]  i = 2  pattern = ""  *while* (i < (len(sys.argv)-1)) :      pattern += sys.argv[i].lower() + " "      i += 1  algoritma = sys.argv[i]  sentenceContainer = []  newContainer = []  date = []  number = []  *# Load text*  f = textReader("uploaded/" + filename).readText()  *for* row in f :      temp = tokenize.sent\_tokenize(row)  *for* text in temp :          sentenceContainer.append(text.lower())  *# Select algorithm*  *if* (algoritma == "boyermoore") :      alg = bm()  *elif* (algoritma == "kmp") :      alg = kmp()  *else* :      alg = reg()  queryNumber = r"*^*([0-9]+([\.,:]?[0-9]\*)\*)*$*"  queryDate = r".?(([1-9]|1[0-9]|2[0-9]|3[0-1])(/|-)(0?[1-9]|1[0-2])(/|-|)([0-9]{0,4})).?"  queryMonth = r"\b(januari|februari|maret|april|mei|juni|juli|agustus|september|oktober|november|desember|jan|feb|mar|apr|jun|jul|ags|sep|okt|nov|des)\b"  queryDay = r"\b(senin|selasa|rabu|kamis|jumat|sabtu|minggu)\b"  queryTime = r"*^*(wib|wita|wit|pm|am)"  *# Sub sentence*  *for* row in sentenceContainer :      subsSentence = []      row = row.split(' ')  *for* sub in row :          subsSentence.append(sub)      newContainer.append(subsSentence)  *# Extraction + Print text*  print("Filename : " + filename, end='<br>')  print("=====================TEKS BERITA=====================", end='<br>')  *for* row in newContainer :      i = 0      dateTemp = []      numberTemp = []      res = ''  *while* i < (len(row)) :  *if* (re.match(queryNumber, row[i])) :  *# Case date*  *if* (re.match(queryMonth, row[i+1])) :  *# Case month*  *if* (re.match(queryNumber, row[i+2])) :                      dateTemp.append(row[i] + ' ' + row[i+1] + ' ' + row[i+2])                      res += '<span style="color: red">' + row[i] + ' ' + row[i+1] + ' ' + row[i+2] + '</span>' + ' '                      i += 2  *# Case without month*  *else* :                      dateTemp.append(row[i] + ' ' + row[i+1])                      res += '<span style="color: red">' + row[i] + ' ' + row[i+1] + '</span>' + ' '                      i += 1  *# Case time*  *elif* (re.match(queryTime, row[i+1])) :                  dateTemp.append(row[i] + ' ' + row[i+1])                  res += '<span style="color: red">' + row[i] + ' ' + row[i+1] + '</span>' + ' '                  i += 1  *# Case number*  *else* :                  numberTemp.append(row[i])                  res += '<span style="color: blue">' + row[i] + '</span>' + ' '  *# Case month only*  *elif* (re.match(queryMonth, row[i])) :              print("INI C: " + row[i], end="<br>")              dateTemp.append(row[i])              res += '<span style="color: red">' + row[i] + '</span>' + ' '  *# Case day*  *elif* (re.match(queryDay, row[i])) :              dateTemp.append(row[i])              res += '<span style="color: red">' + row[i] + '</span>' + ' '  *# Case date*  *elif* (re.match(queryDate, row[i])) :              dateTemp.append(row[i])              res += '<span style="color: red">' + row[i] + '</span>' + ' '  *# Another*  *else* :              res += row[i] + ' '          i = i + 1      date.append(dateTemp)      number.append(numberTemp)      print(res, end='<br>')  *# Construct date*  *for* i in range(len(date)) :      temp = ''      dateSize = len(date[i])  *for* j in range(dateSize) :          temp += date[i].pop(0)  *if* (j != dateSize-1) :              temp += ' '      date[i] = temp  *# Get news date*  *for* row in date :  *if* (len(row) != 0) :          newsDate = row  *break*  print("===================HASIL EKSTRAKSI===================", end='<br>')  *# Result + Print Result*  print("Keyword\t: " + pattern, end='<br>')  print("Hasil ekstraksi informasi\t:", end='<br>')  *for* i in range (len(sentenceContainer)) :  *if* (alg.match(sentenceContainer[i], pattern) != -1) :          patternIdx = searchIndex(pattern, newContainer[i])          numberIdx = []          tempNum = number[i].copy()  *for* word in newContainer[i] :  *if* (word in tempNum) :                  num = tempNum.pop(0)                  numberIdx.append(abs(searchIndex(num, newContainer[i]) - patternIdx))  *if* (len(numberIdx) != 0) :              nearest = min(numberIdx)              index = numberIdx.index(nearest)          print("Jumlah\t: ")  *if* (len(number[i]) != 0) :              print(number[i][index], end='<br>')  *else* :              print("-", end='<br>')          print("Tanggal\t: ")  *if* (len(date[i]) != 0) :              print(date[i], end='<br>')  *else* :              print(newsDate, end='<br>')          print(sentenceContainer[i], end='<br>')          print("=====================================================", end='<br>') |

|  |
| --- |
| File index.php (src/index.php) |
| <!DOCTYPE html>  <html lang="en">  <head>      <meta charset="UTF-8">      <meta name="viewport" content="width=device-width, initial-scale=1.0">      <title>Extractor</title>      <link rel="stylesheet" href="css/bootstrap.min.css">  </head>  <body>      <div class="container mt-3">  *<!-- PAGE TITLE -->*          <div class="text-center">              .              <h1>Informator</h1>              <p>Get your infomation here!</p>          </div>          <form action="submit.php" method="post" enctype="multipart/form-data">  *<!-- SELECT FOLDER -->*          <div class="was-validated">              <label for="file">Folder</label>              <div class="input-group custom-file mb-3">                  <input type="file" class="custom-file-input" name="file" accept=".txt" id="file" required>                  <label for="file" class="custom-file-label" id="file">Select file</label>                  <div class="invalid-feedback">Invalid file</div>              </div>          </div>  *<!-- SELECT FOLDER -->*  *<!-- INPUT KEYWORD -->*          <label for="keyword">Keyword</label>          <div class="input-group mb-3">              <input type="text" class="form-control" name="keyword" id="keyword" autofocus autocomplete="off" required>          </div>  *<!-- INPUT KEYWORD -->*  *<!-- SELECT ALGORITMA -->*          <label for="algoritma">Algoritma</label>          <div class="input-group mb-3">              <select name="algoritma" id="algoritma" class="custom-select" required>                  <option value="">Select algorithm</option>                  <option value="bm">Boyer-Moore</option>                  <option value="kmp">Knuth-Morris-Pratt</option>                      <option value="regex">Regex</option>                  </select>                  <div class="invalid-feedback">Invalid input</div>              </div>              <div class="input-group mb-3">                  <input type="submit" class="btn btn-primary" name="submit" value="Submit"></button>              </div>          </form>  *<!-- SELECT ALGORITMA -->*          <div class="mb-5">                <?php  *if* (!$\_GET) {                    }  *else* *if* (($\_GET['s'])==='success') {                      echo "<p>Result</p>";                      $fileName = $\_GET['filename'];                      $keyword = $\_GET['keyword'];                      $algoritma = $\_GET['algoritma'];                      $command = escapeshellcmd("python main.py ".$fileName." ".$keyword." ".$algoritma);                      $output = shell\_exec($command);                      echo $output;                  }                  ?>          </div>      </div>  *<!-- Footer -->*  <footer class="page-footer font-small fixed-bottom">      <div class="footer-copyright text-center py-3 bg-light footer-light">© IF2211 Strategi Algoritma : 13518033 / Daffa Pratama Putra</div>  </footer>  *<!-- Footer -->*  </body>  </html> |

|  |
| --- |
| File submit.php (src/submit.php) |
| <?php  *if* (isset($\_POST['submit'])) {          $file = $\_FILES['file'];          $fileName = $file['name'];          $fileTmp = $file['tmp\_name'];          $fileError = $file['error'];  *// echo $fileName;*  *if* ($fileError === 0) {              $fileDestination = "uploaded/".$fileName;              move\_uploaded\_file($fileTmp, $fileDestination);          } *else* {              header("Location: index.php?upload=failed");          }          $query = array(              's' => 'success',              'filename' => $fileName,              'keyword' => $\_POST['keyword'],              'algoritma' => $\_POST['algoritma']          );            $query = http\_build\_query($query);          header("Location: index.php?$query");      }  ?> |

# **BAB III UJI COBA**

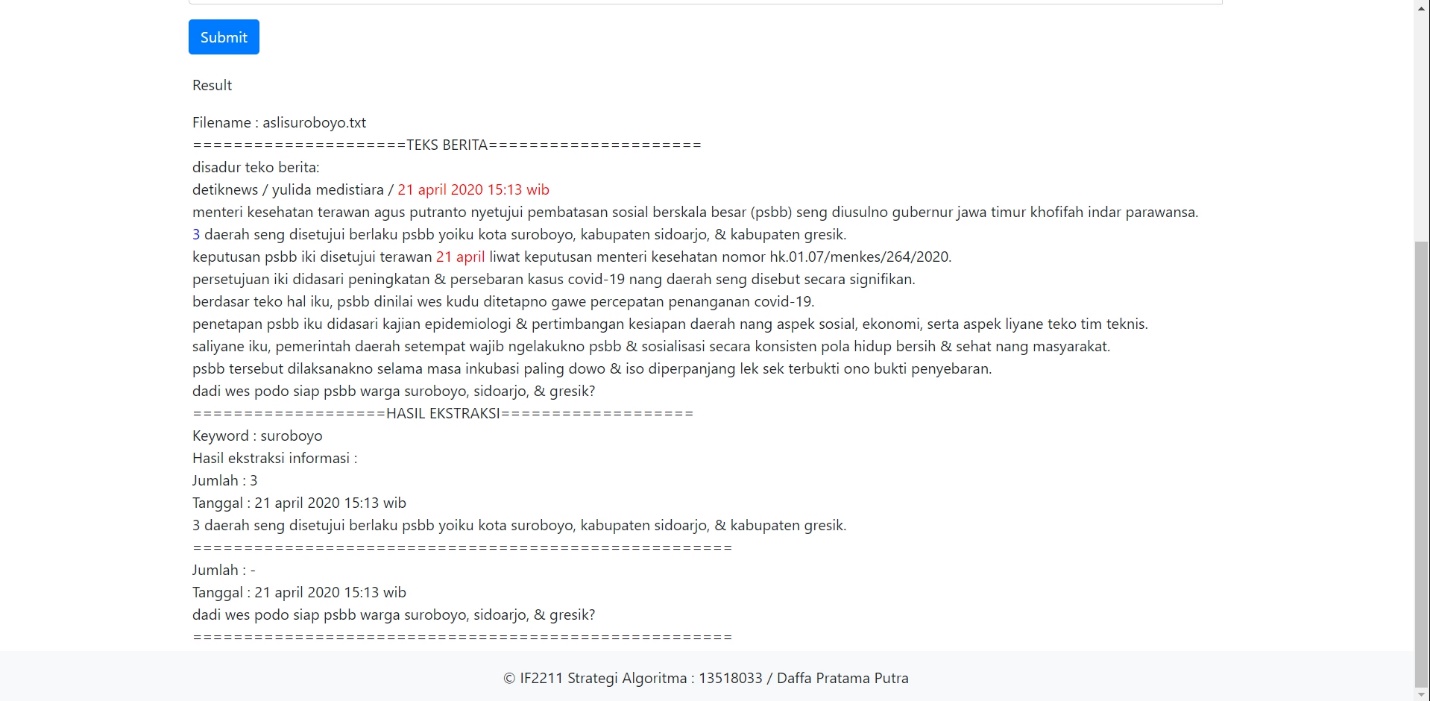
1. **Berita “aslisuroboyo.txt”**

Berikut adalah hasil uji coba dengan masukan :

Filename : “aslisuroboyo.txt”

Keyword : suroboyo

Algoritma : boyer-moore



Gambar 4. Hasil ekstraksi pada teks “aslisuroboyo”

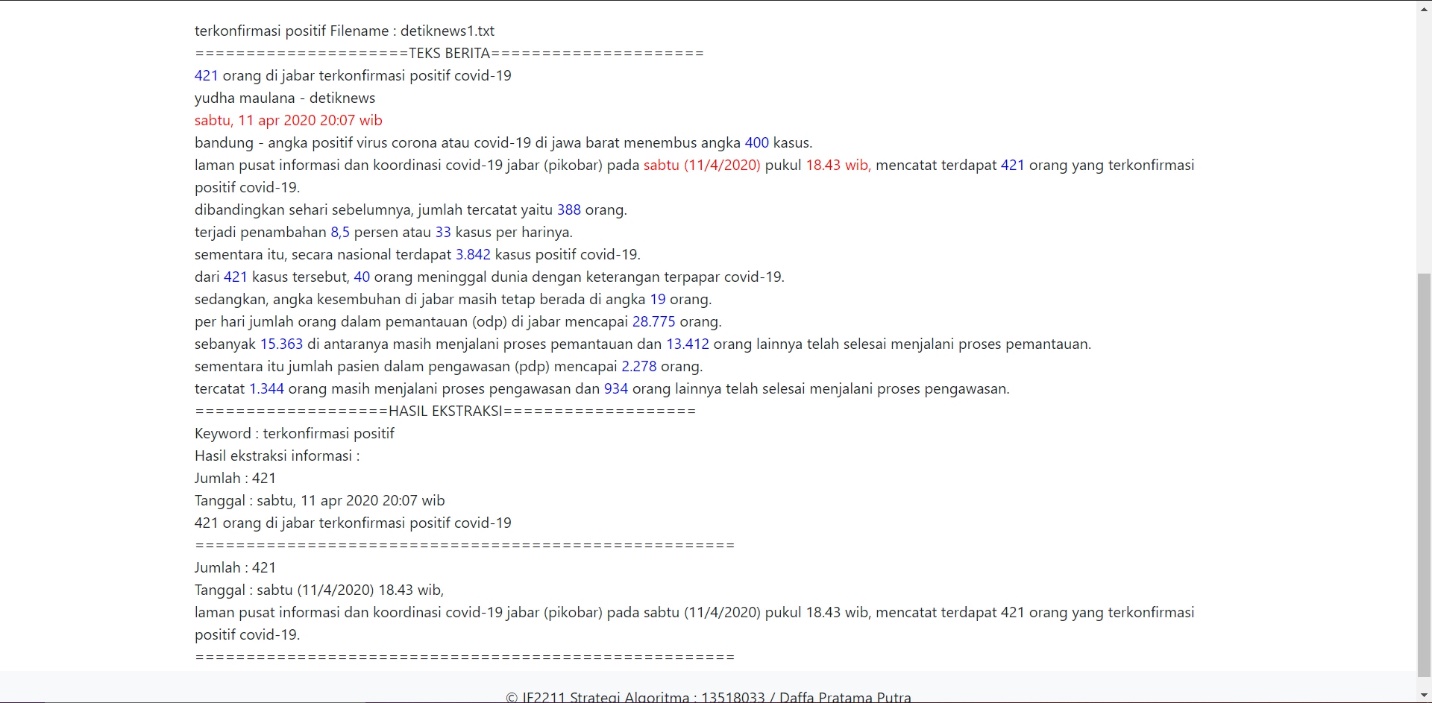
1. **Berita “detiknews1.txt”**

Berikut adalah hasil uji coba dengan masukan :

Filename : “detiknews1.txt”

Keyword : terkonfirmasi positif

Algoritma : boyer-moore



Gambar 5. Hasil ekstraksi pada teks “detiknews1.txt”

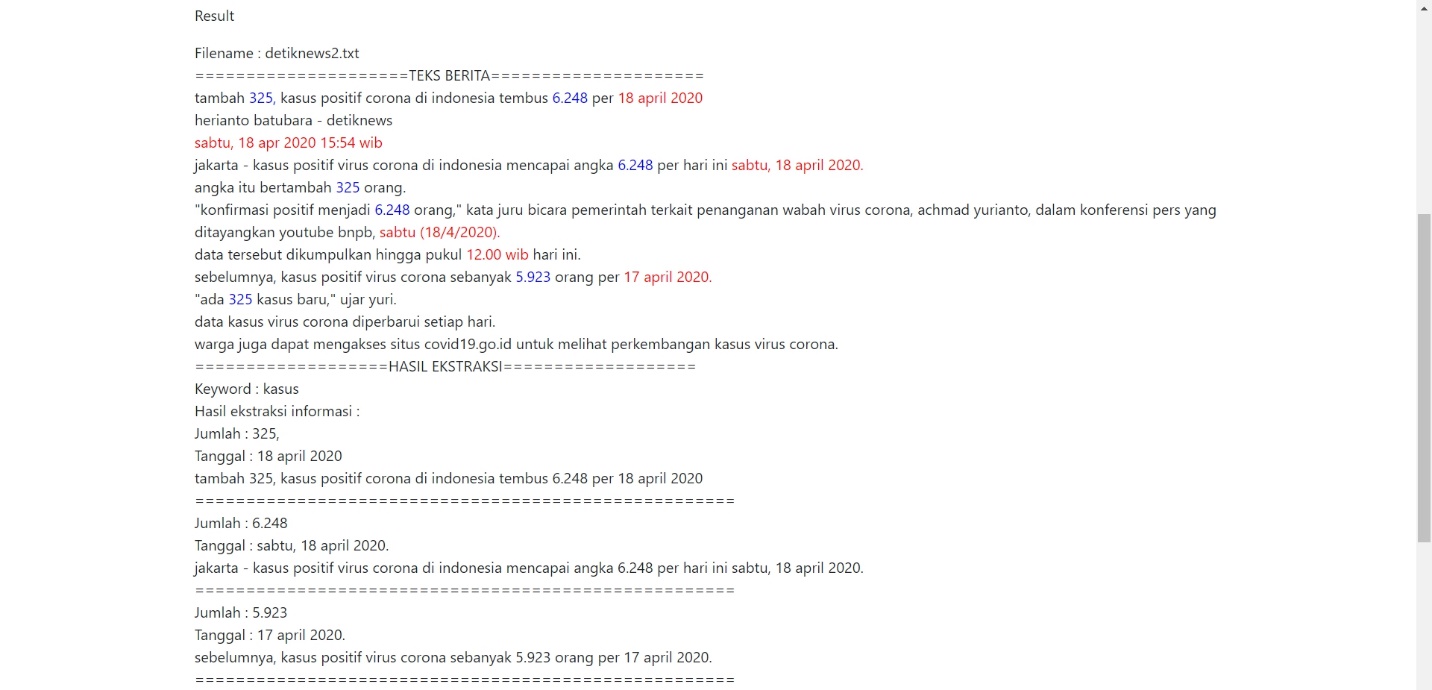
1. **Berita “detiknews2.txt”**

Berikut adalah hasil uji coba dengan masukan :

Filename : “detiknews2.txt”

Keyword : kasus

Algoritma : kmp



Gambar 6. Hasil ekstraksi pada teks “detiknews2.txt”



Gambar 7. Hasil ekstraksi pada teks “detiknews2.txt”

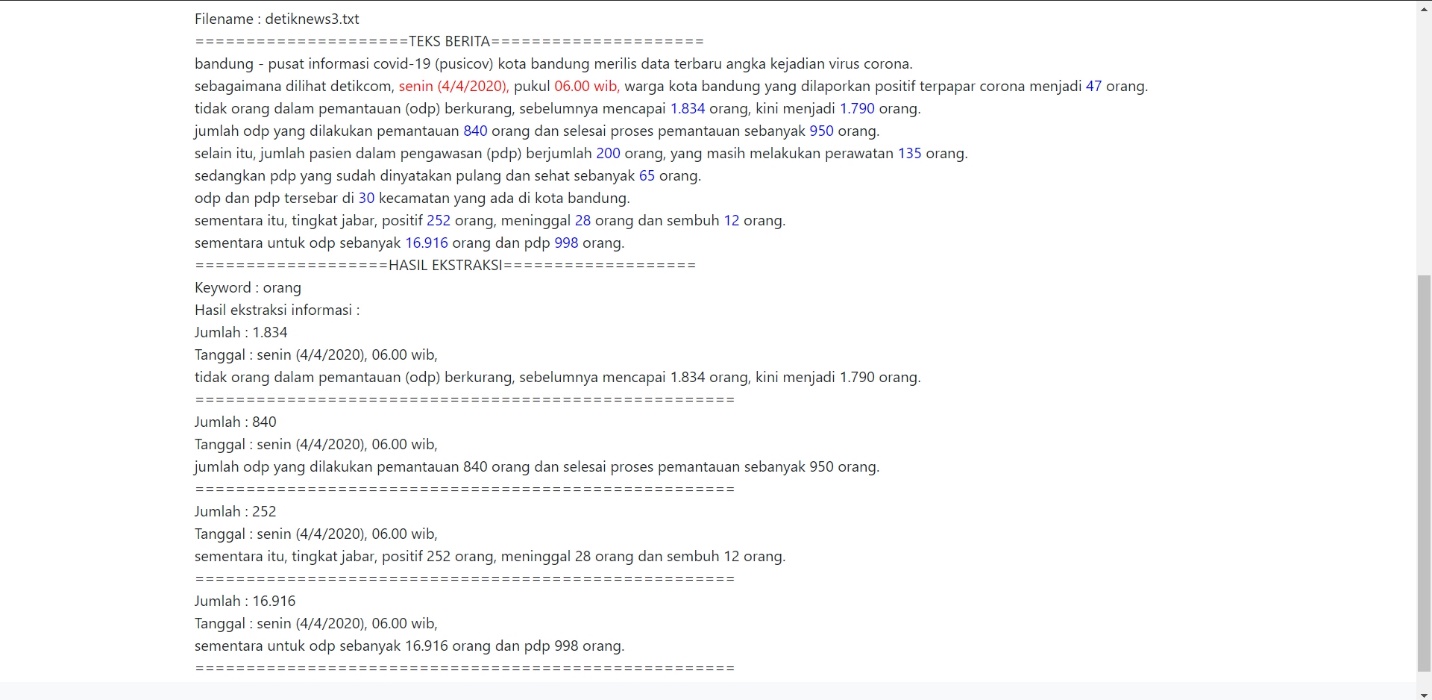
1. **Berita “detiknews3.txt”**

Berikut adalah hasil uji coba dengan masukan :

Filename : “detiknews3.txt”

Keyword : orang

Algoritma : kmp



Gambar 8. Hasil ekstraksi pada teks “detiknews3.txt”

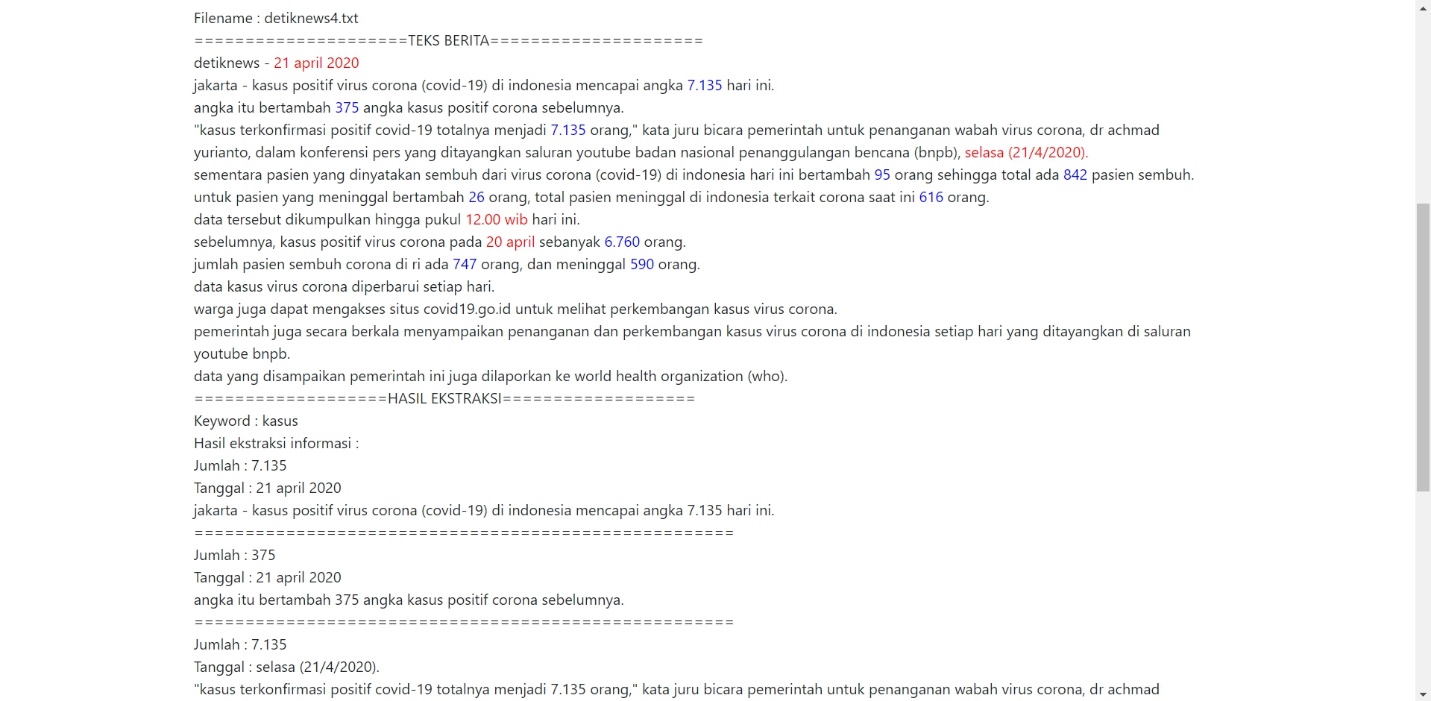
1. **Berita “detiknews4.txt”**

Berikut adalah hasil uji coba dengan masukan :

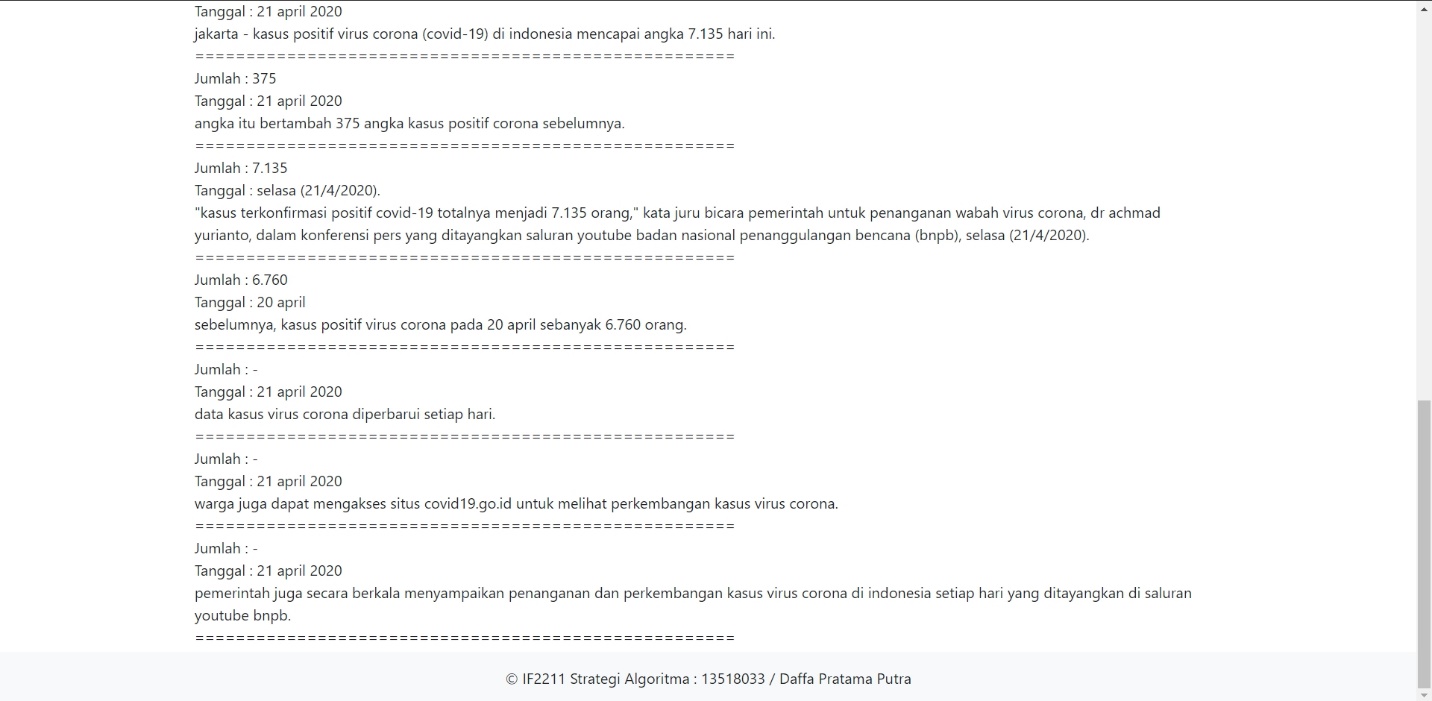
Filename : “detiknews4.txt”

Keyword : kasus

Algoritma : kmp



Gambar 9. Hasil ekstraksi pada teks “detiknews4.txt”



Gambar 10. Hasil ekstraksi pada teks “detiknews4.txt”

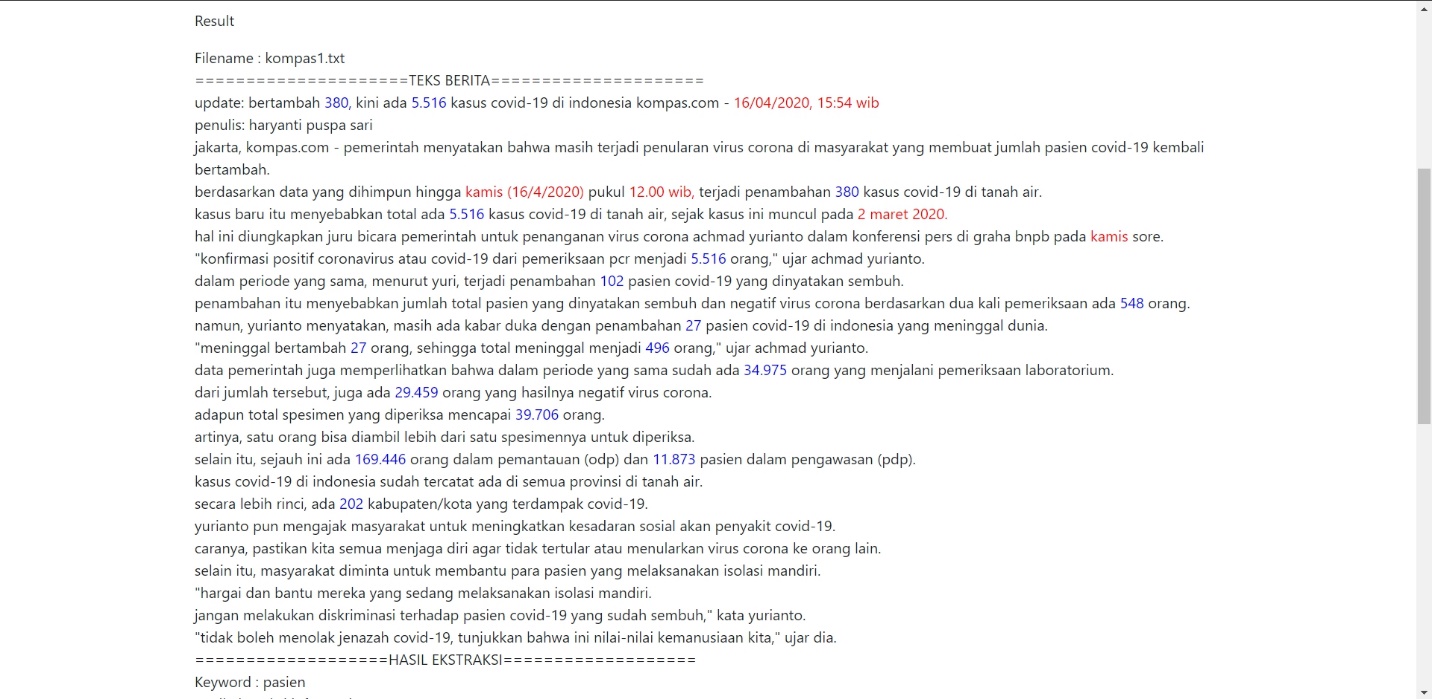
1. **Berita “kompas1.txt”**

Berikut adalah hasil uji coba dengan masukan :

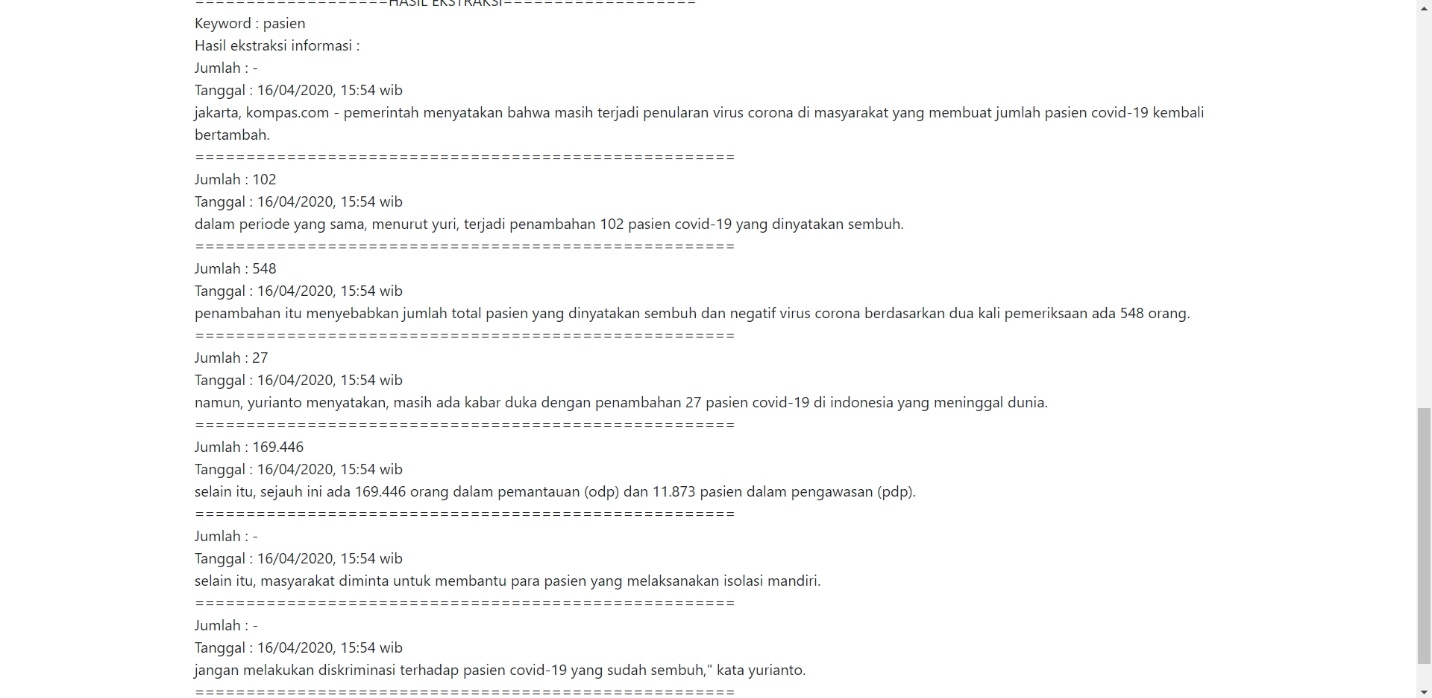
Filename : “kompas1.txt”

Keyword : pasien

Algoritma : bm



Gambar 11. Hasil ekstraksi pada teks “kompas1.txt”



Gambar 12. Hasil ekstraksi pada teks “kompas1.txt”

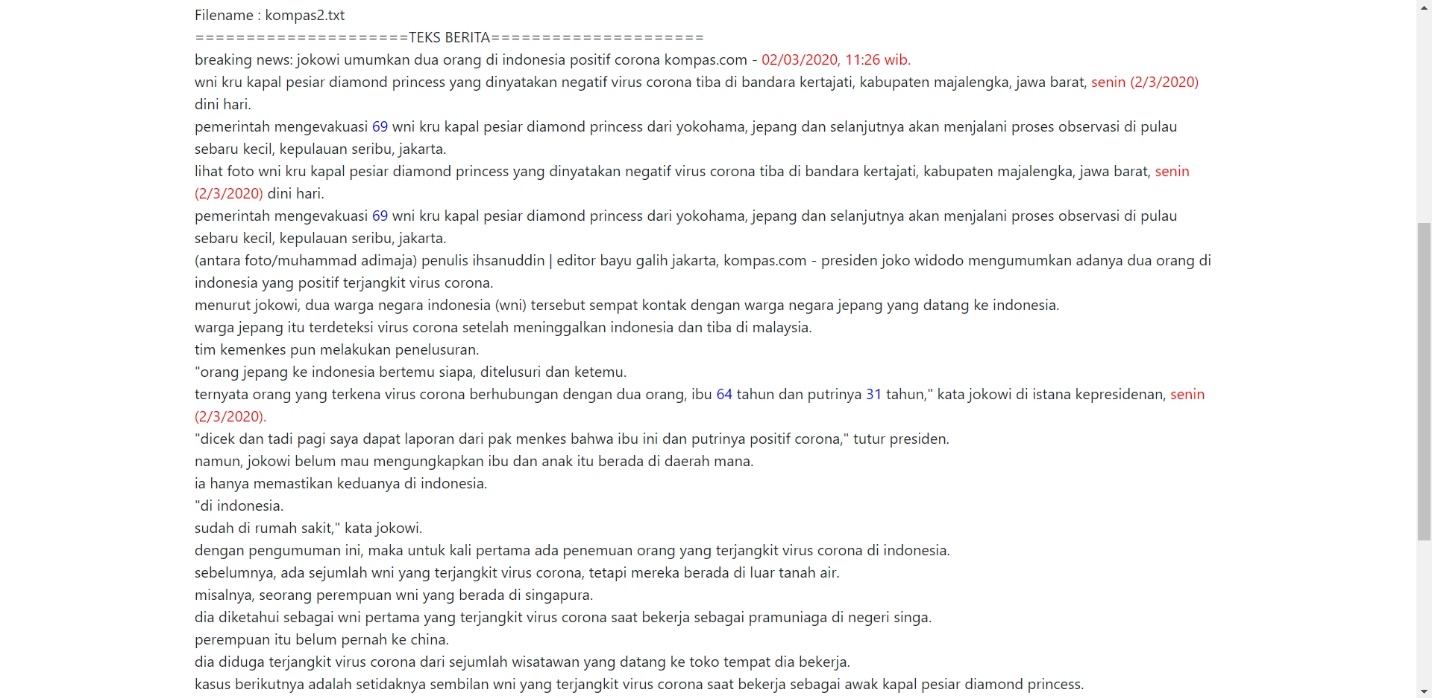
1. **Berita “kompas2.txt”**

Berikut adalah hasil uji coba dengan masukan :

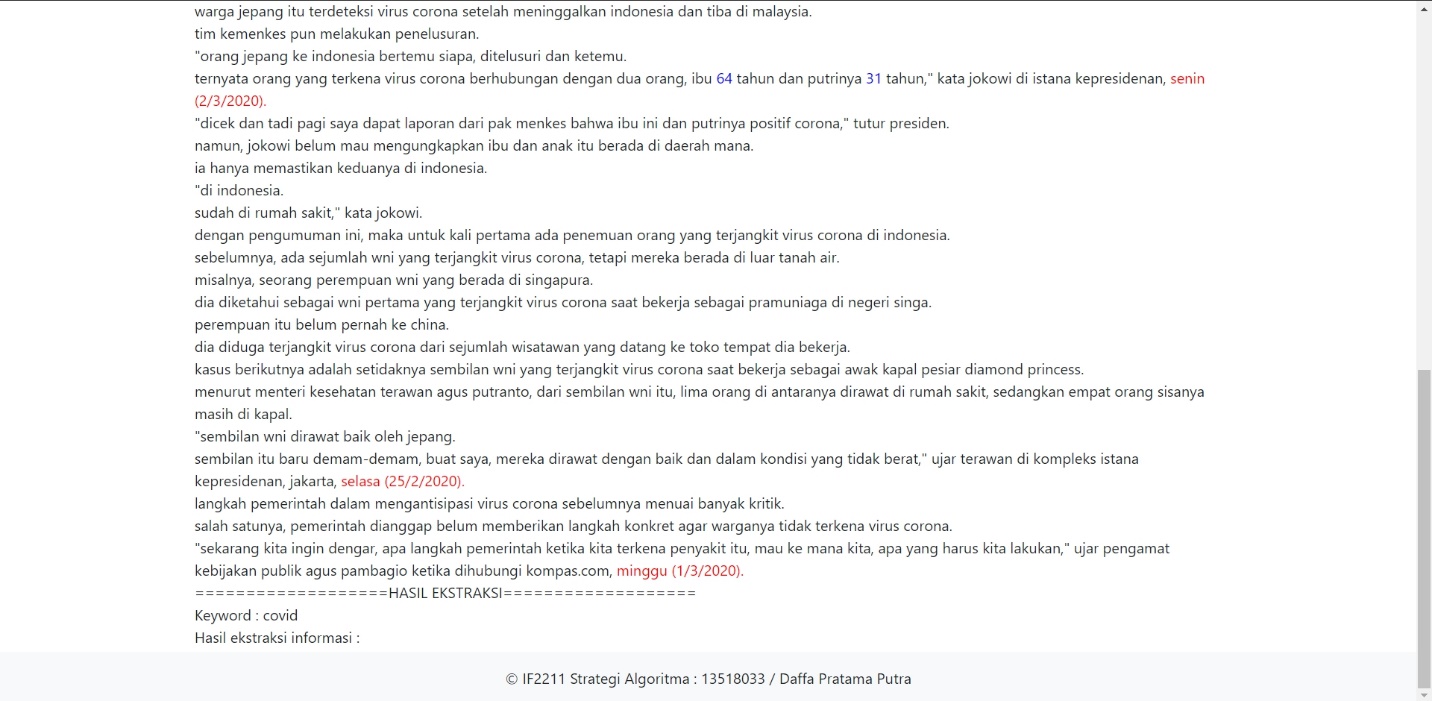
Filename : “kompas2.txt”

Keyword : covid

Algoritma : regex



Gambar 13. Hasil ekstraksi pada teks “kompas2.txt”



Gambar 13. Hasil ekstraksi pada teks “kompas2.txt”

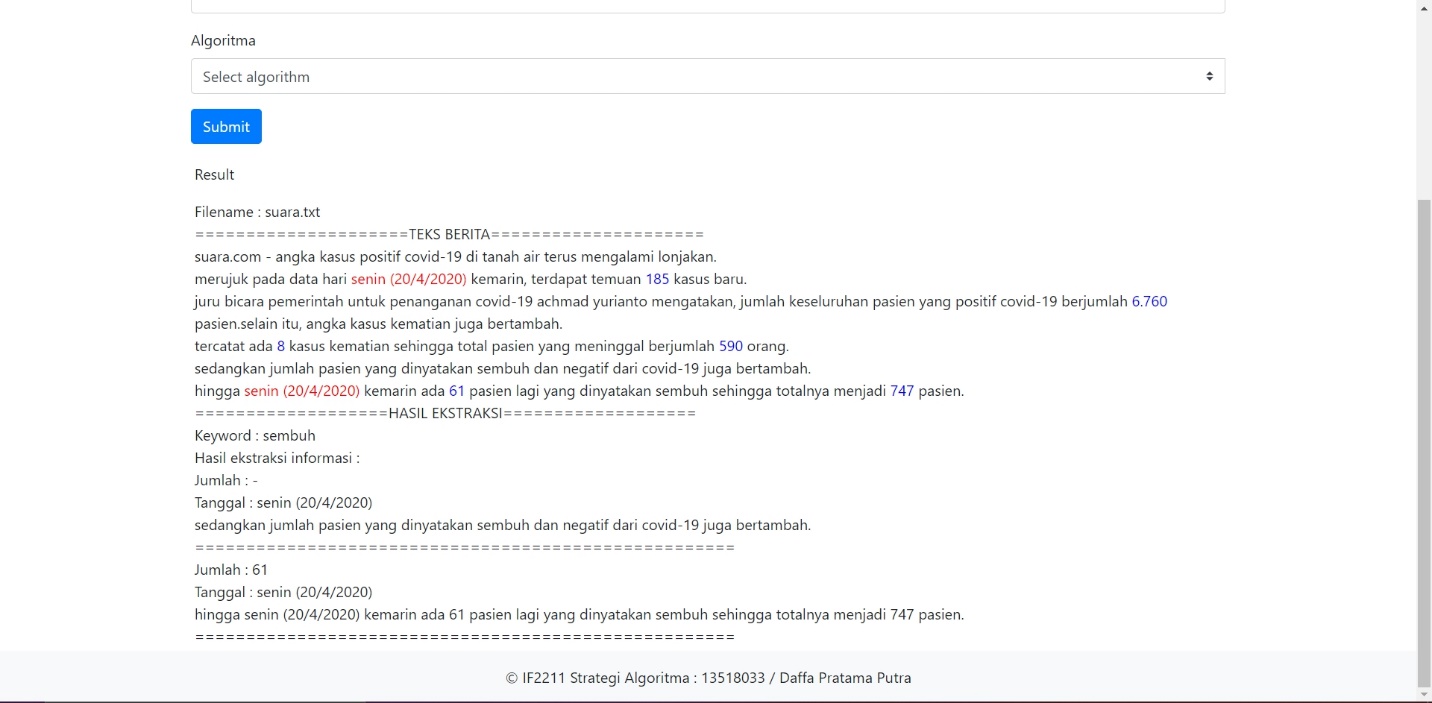
1. **Berita “suara.txt”**

Berikut adalah hasil uji coba dengan masukan :

Filename : “suara.txt”

Keyword : sembuh

Algoritma : regex



Gambar 14. Hasil ekstraksi pada teks “suara.txt”

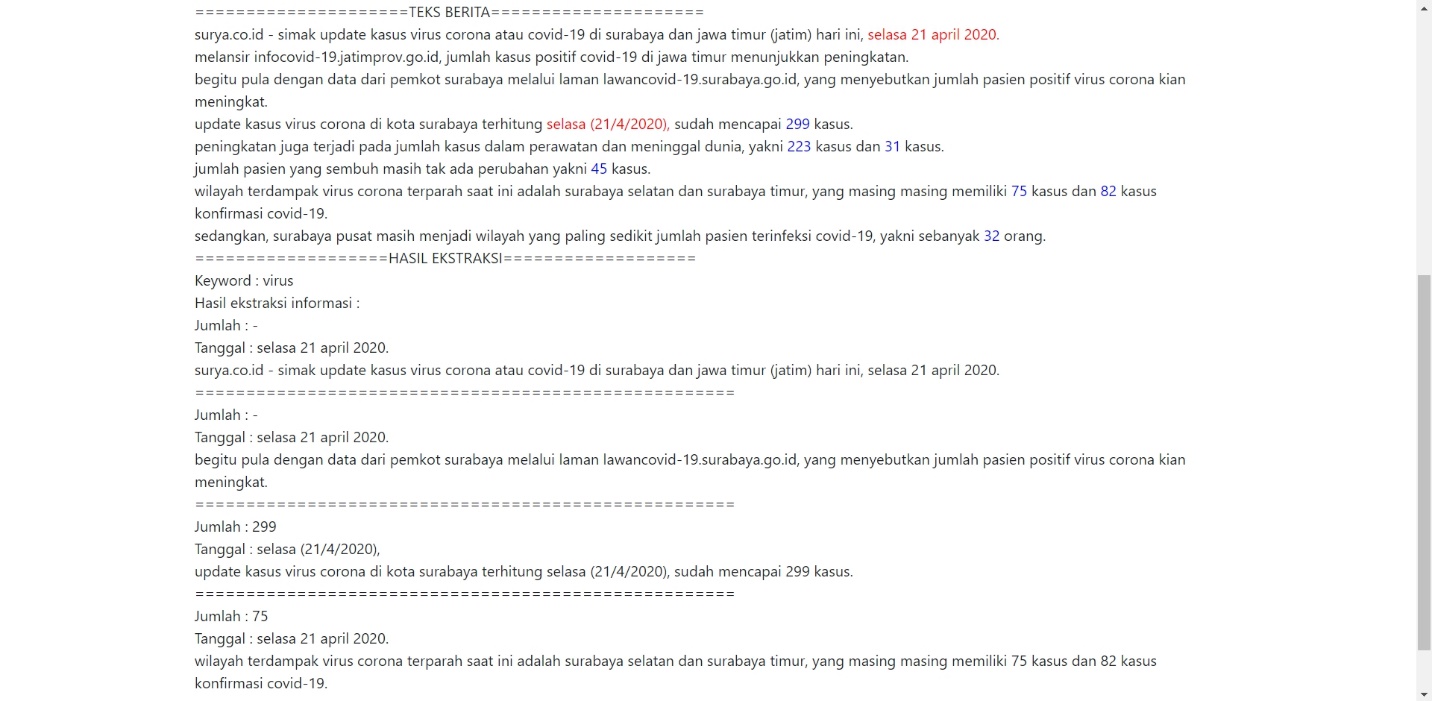
1. **Berita “surya.txt”**

Berikut adalah hasil uji coba dengan masukan :

Filename : “surya.txt”

Keyword : virus

Algoritma : regex



Gambar 15. Hasil ekstraksi pada teks “surya.txt”

1. **Berita “tirto.txt”**

Berikut adalah hasil uji coba dengan masukan :

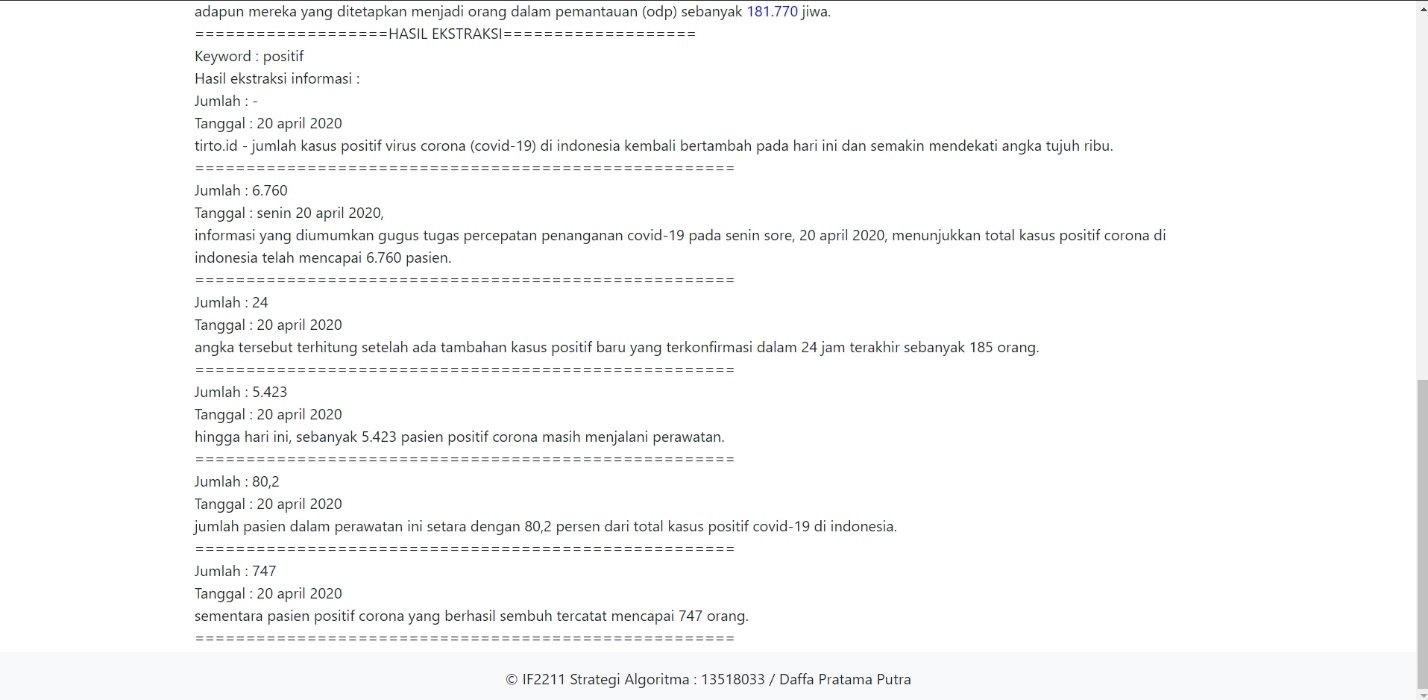
Filename : “tirto.txt”

Keyword : positif

Algoritma : regex



Gambar 16. Hasil ekstraksi pada teks “tirto.txt”



Gambar 17s. Hasil ekstraksi pada teks “tirto.txt”