# Efficient Algorithms: Suffix Array Construction

## BruteForce:

|  |
| --- |
| import re  def getAllWords():      all\_words\_cleaned=[]      with open('./TomSawyerCompleteGutenberg.txt', encoding='utf-8-sig') as f:          for line in f:              # split on space and line endings              splittet\_line=line.split()                wordIndex = 0              replacedSpecialChars = []              for wordIndex in range(len(splittet\_line)):                  # find all special chars and replace them with nothign                  newWord = re.sub("[-]+", "", splittet\_line[wordIndex])                  # make all words lowercase                  newWord = newWord.lower()                  replacedSpecialChars.append(newWord)                  wordIndex+=1              if(len(replacedSpecialChars)>0):                  for word in replacedSpecialChars:                      all\_words\_cleaned.append(word)      return all\_words\_cleaned  ## generate suffix array  def createSuffixArr(all\_words\_cleaned):      suffix\_arr = []      singleString = " ".join(all\_words\_cleaned)      #limit string      singleString = singleString[0:int(len(singleString)/4)]      #generate suffix arr      charIndex = 0      for charIndex in range(len(singleString)):              if(singleString[charIndex]== " "):                  continue              suffix = singleString[charIndex:]              suffix +="$"              suffix\_arr.append(suffix)      return suffix\_arr  ##sort suffix array  import time  i = 0  runs = 5  allTimesTogehter = 0  while i < runs:      start\_time = time.time()      allWords = getAllWords()      suffix\_array = createSuffixArr(allWords)      suffix\_array.sort()      currentTime = time.time()-start\_time      allTimesTogehter += currentTime      print("--- %s seconds ----" % (currentTime))      i +=1  print("--- %s All Runs ----" % (allTimesTogehter/5)) |

## Durchschnittliche Laufzeit bei 5 Durchläufen bei Brute Force(Bei Verwendung des ersten Viertel des Buches):

Ich konnte leider nicht den gesamten Text des Buches verwenden da dies zu einer out of memory Exception führt. Deswegen habe ich nur das erste viertel verwendet und bekomme dabei diese Zeit nach 5 Versuchen gemittelt heraus:

--- 39.259 All Runs ----

## Eigene Implementierung

|  |
| --- |
| import re  def getAllWords():      all\_words\_cleaned=[]      with open('./TomSawyerCompleteGutenberg.txt', encoding='utf-8-sig') as f:          for line in f:              # split on space and line endings              splittet\_line=line.split()                wordIndex = 0              replacedSpecialChars = []              for wordIndex in range(len(splittet\_line)):                  # find all special chars and replace them with nothign                  newWord = re.sub("[-]+", "", splittet\_line[wordIndex])                  # make all words lowercase                  newWord = newWord.lower()                  replacedSpecialChars.append(newWord)                  wordIndex+=1              if(len(replacedSpecialChars)>0):                  for word in replacedSpecialChars:                      all\_words\_cleaned.append(word)      return all\_words\_cleaned  ## generate suffix array  def createSuffixArr(all\_words\_cleaned):      suffix\_arr = []        singleString = " ".join(all\_words\_cleaned)      #limit string      singleString = singleString[0:int(len(singleString)/4)]      #generate suffix arr      charIndex = 0      for charIndex in range(len(singleString)):              if(singleString[charIndex]== " "):                  continue              suffix = singleString[charIndex:]              suffix +="$"              suffix\_arr.append(suffix)      return suffix\_arr  def quick\_sort(s):      if len(s) == 1 or len(s) == 0:         return s      else:          pivot = s[0]          i = 0          for j in range(len(s)-1):              if s[j+1] < pivot:                 s[j+1],s[i+1] = s[i+1],s[j+1]                 i += 1          s[0],s[i] = s[i],s[0]          first\_part = quick\_sort(s[:i])          second\_part = quick\_sort(s[i+1:])          first\_part.append(s[i])          return first\_part + second\_part  ##sort suffix array  import time  i = 0  runs = 5  allTimesTogehter = 0  while i < runs:      start\_time = time.time()      allWords = getAllWords()      suffix\_array = createSuffixArr(allWords)      # suffix\_array.sort()      quick\_sort(suffix\_array)      currentTime = time.time()-start\_time      allTimesTogehter += currentTime      print("--- %s seconds ----" % (currentTime))      i +=1  print("--- %s All Runs ----" % (allTimesTogehter/5)) |

## Beschreibung der Optimierungs Idee:

Da zuvor list.sort() verwendet wurde, habe ich mit etwas eingelesen und herausgefunden dass Python Timsort verwendet. Deshalb habe ich versucht einen anderen sorting algo zu verwenden. Die Wahl ist dann einfach auf Quicksort gefallen und habe diesen ausprobiert.

## Durchschnittliche Laufzeit(Bei Verwendung des ersten Viertel des Buches), 5 Versuche:

--- 28.204 All Runs ----

## Argumentation warum Ihr Algorithmus praktische schneller/langsamer ist, als Brute-Force.

Timsort liefert eine bessere Performance bei „leicht“ vorsortieren Daten. Deshalb wird Timsort auch in vielen realen Beispielen verwendet da eine große Anzahl der Daten bereits natürlich etwas vorsortiert sind. In unserem Fall sind jedoch die Suffixe wenig vorsortiert, weshalb dieser Algorithmus eine bessere Performance bei mir liefert.

Als ich die beidne Algorithmen am nächsten Tag ausprobierte wurden jedoch die Wert