Web Crawler Smartphones

Eduardo Conti - ESC3
José Bezerra - JBMN2
Lucas Mendonça - LSM5

Gerenciador de Fronteira

Parte 1 - Eduardo Siqueira Conti (esc3)

ESCLARECIMENTO

Dominio

ORGANIZAÇÃO DO CÓDIGO

- Robots.txt
- A função crawl
  - As expressões regulares
- Parte recursiva
  - Busca em largura
  - Busca heurística com pesos
- Gerando arquivo fronteira

CÓDIGO

robots

```
def robots(self):
    try:
        r = requests.get(self + "/robots.txt")
        html = r.text.encode("utf8")
        string = str(html)
        string = string.split("User-agent: *\\n")
        string = string[1]
        string = string.replace("\\n", "")
        string = string.replace("*", "")
        string = string.split("Disallow: ")
        for item in string:
            if item is not '':
                exceptions.append(item)
        print("robot success")

    except:
        print("robot error")
```

CÓDIGO

crawl

```
def crawl(self):
    crawled.append(self.link)
    root = self.link
    html = self.html
    print("now crawling:", root)

    try:
      search1 = re.findall(r'href=[\'"?](http[://\w\-._]+)', html.decode("utf8"))
      for link in search1:
          if link not in exceptions:
              for item in rootNames:
                  if len(frontier) < 1000:
                      if item in link and link not in downloaded and link not in crawled:
                          if not any(excecao in link for excecao in exceptions):
                              if link not in frontier:
                                  print(link)
                                  frontier.append(link)
                                  insert(self, link)
                                  downloaded.append(link)
```

CÓDIGO

parte recursiva (largura)

```
def rec(self):
    crawl(self)
    for item in self.children:
        crawl(item)

for item in nodeList:
    frontier = []

    exceptions = ["whatsapp", "facebook", "instagram", "jpg", "png", "jpeg", "youtube", "img"]
    robots(item.link)
    crawl(item)
    for it in item.children:
        rec(it)
```

CÓDIGO

parte recursiva (heuristica)

```
roots = []
roots.append("http://www.webfones.com.br")

rootNames = ["webfones"]


frontier = []
frontierValue = {}
dicitens = ["smartphone", "motorola", "samsung", "celular", "galaxy", "s20", "selfie", "gb", "philco", "core", "nokia", "camera"]
dicvalues = [1, 0.95, 0.9, 0.85, 0.8, 0.75, 0.7, 0.65, 0.6, 0.55, 0.5, 0.45]
downloaded = []
crawled = []
exceptions = []

concat = ""
nArquivos = 0
exceptions = []
```

CÓDIGO

arquivo fronteira

```
concat = ''
for item in frontier:
    concat = concat + "\n" + item
f = open("concat.txt", "w")
f.write(concat)
f.close()

print("Fim do", item, " e todos seus filhos. Tamanho da fronteira: ", len(frontier))
```