

Lendo documentos

Arquivos de texto

Lendo documentos

Arquivos de texto

A internet não vive apenas de HTML, aliás, desde o fim dos anos 1960, era usada apenas para troca de arquivos e e-mails, o HTML só apareceu em 1992.

Lendo documentos

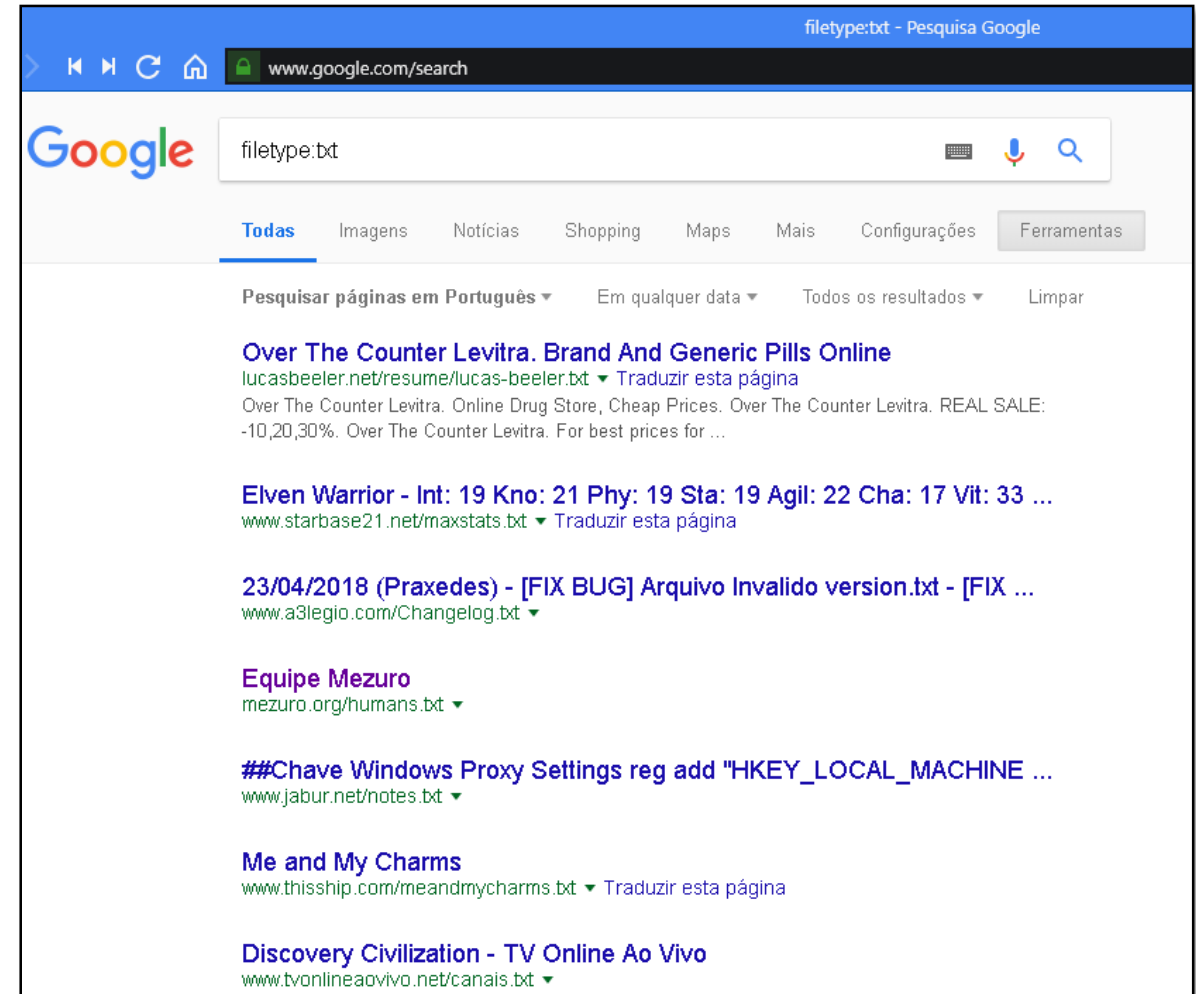
Arquivos de texto

Podemos encontrar arquivos de texto disponíveis em alguns sites e nesta aula vamos ver como tratar estes arquivos.

Lendo documentos

Arquivos de texto

Vou buscar um arquivo de texto aleatoriamente no Google informando na caixa de busca filetype:txt



Lendo documentos

Arquivos de texto

A maioria dos navegadores exibe arquivos de texto normalmente e podemos incluir arquivos de texto em nosso scraping sem problemas. Após identificar o arquivo que usaremos usar como exemplo, vamos escrever o código para tratar o arquivo.

Lendo documentos

Arquivos de texto

```
from urllib.request import urlopen
paginaTexto = urlopen("http://mezuro.org/humans.txt")
print(paginaTexto.read())
```

Neste caso não podemos utilizar o BeautifulSoup por exemplo, porque em arquivos de texto não existem as marcações HTML para tratarmos cada parte do texto.

Lendo documentos

Arquivos de texto

Ler um arquivo de texto não é um grande problema, porém, podemos nos deparar com idiomas e codificações que teremos que tratar.

Existem alguns padrões para definição de caracteres que veremos a seguir.

Lendo documentos

Arquivos de texto

ASCII: é um padrão de codificação de texto usado desde os anos 1960, emprega 7 bits para codificar cada um de seus caracteres, tendo um total de 2^7 , ou seja, 128 caracteres. O que atende perfeitamente o alfabeto latino, incluindo maiúsculas, minúsculas, pontuação e basicamente todos os caracteres encontrados no teclado de uma pessoa que fala inglês.

Lendo documentos

Arquivos de texto

UTF-8: criado por uma entidade sem fins lucrativos, denominada The Unicode Consortium, sua proposta era ser um codificador de textos universal, estabelecendo codificações para cada caractere que precisasse ser usado em documentos de texto, em qualquer idioma. O objetivo era incluir tudo que fosse proveniente do alfabeto latino, o alfabeto cirílico, os pictogramas chineses, símbolos matemáticos e lógicos e até mesmo emoticons e símbolos variados.

Lendo documentos

Arquivos de texto

O UTF-8 significa “Universal Character Set – Transformation Format 8 bit”, onde 8 bits é o menor tamanho que um caractere requer para ser exibido. O UTF-8 permite caracteres de até 4 bytes (32 bits), porém, alguns bytes são utilizados para codificar o caractere, sendo assim, normalmente temos 21 bits de informações sendo utilizados, possibilitando 2.097.152 caracteres possíveis, dos quais 1.114.112 são usados atualmente.

Lendo documentos

Arquivos de texto

Existem outros padrões Unicode como o UTF-16 e UTF-32, mas raramente encontramos documentos com estas codificações.

Um dos problemas do Unicode é que qualquer documento escrito em um único idioma estrangeiro é muito maior do que deveria. Seu idioma pode requerer apenas cerca de 100 caracteres, mas você precisará de no mínimo 16 bits para cada caractere em vez de apenas 8, como é o caso do padrão ASCII do idioma inglês. Isso faz com que textos em outros idiomas (que não usem o idioma latino) sejam no mínimo duas vezes maiores que textos no idioma inglês.

Lendo documentos

Arquivos de texto

A codificação ISO tenta resolver este problema criando uma codificação específica para cada idioma. Como o Unicode, possui as mesmas codificações que o ASCII, criando caracteres especiais para todos idiomas que precisem deles. O ISO-8859-1 projetado para o alfabeto latino possui, além dos caracteres existentes na codificação ASCII por exemplo, símbolos de frações ou o sinal de direito autoral “©”.

Lendo documentos

Arquivos de texto

ISO-8859-9 -> Turco

ISO-8859-2 -> Alemão

ISO-8859-15 -> Francês

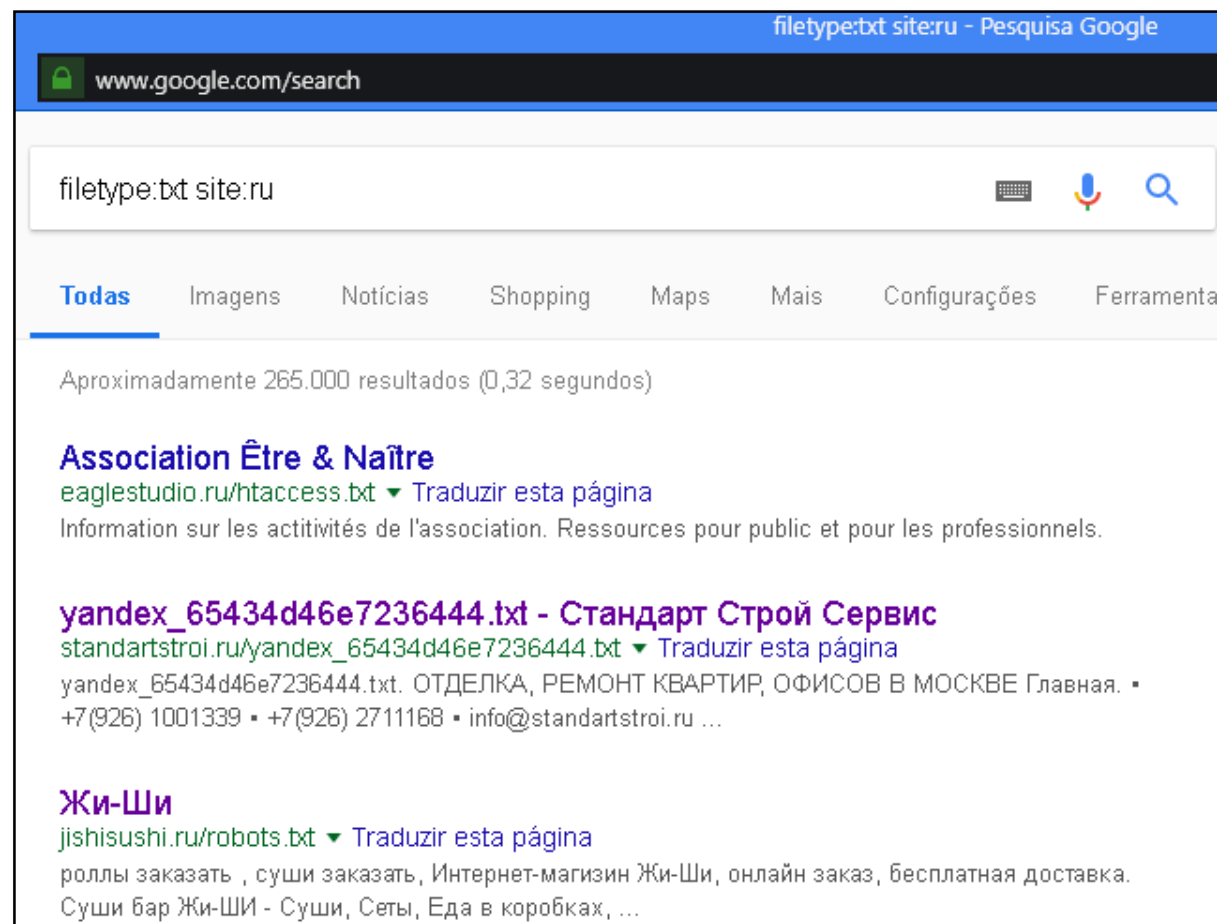
Embora a popularidade dos documentos codificados em ISO tenha diminuído nos últimos anos, cerca de 8% dos sites na Internet ainda são codificados com alguma versão padrão*, o que torna essencial conhecê-lo e verificar as codificações antes de fazer o scraping de um site.

*Fonte: https://w3techs.com/technologies/history_overview/character_encoding

Lendo documentos

Arquivos de texto

Vou novamente ao Google fazer uma busca por arquivos txt em sites da Rússia usando site:ru



Lendo documentos

Arquivos de texto

Vamos trabalhar com o txt
encontrado em:
<http://bgcrm.ru/changes.txt>

Lendo documentos

Arquivos de texto

Executando sem especificar o encode (ASCII):

```
from urllib.request import urlopen
paginaTexto = urlopen("http://bgcrm.ru/changes.txt")
print(paginaTexto.read())
```

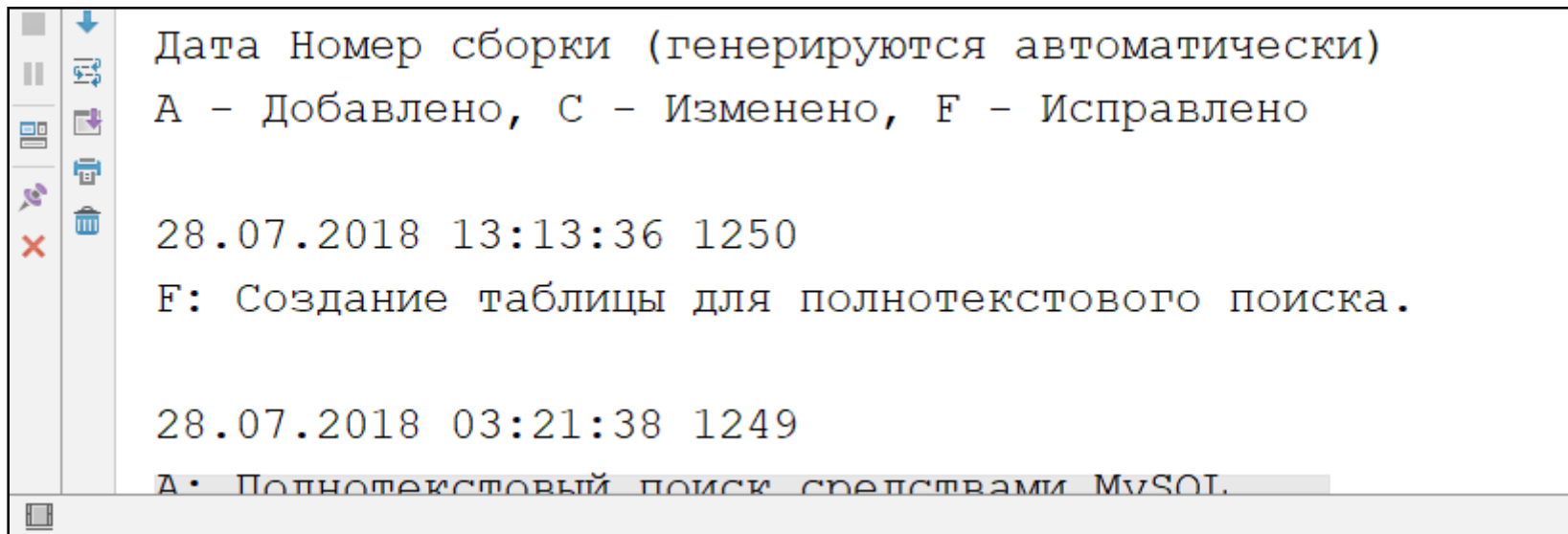
```
4
5 b'\xd0\x94\xd0\xb0\xd1\x82\xd0\xb0 \xd0\x9d\xd0\xbe\xd0\xbc\xd0\xb5\xd1\x80 \xd1\x81\xd0\xb
6 (\xd0\xb3\xd0\xb5\xd0xbd\xd0\xb5\xd1\x80\xd0\xb8\xd1\x80\xd1\x83\xd1\x8e\xd1\x82\xd1\x81\x
7 \xd0\xb0\xd0\xb2\xd1\x82\xd0\xbe\xd0\xbc\xd0\xb0\xd1\x82\xd0\xb8\xd1\x87\xd0\xb5\xd1\x81\xd
8 \xd0\x94\xd0\xbe\xd0\xb1\xd0\xb0\xd0\xb2\xd0\xbb\xd0\xb5\xd0xbd\xd0\xbe, C - \xd0\x98\xd0\
9 F - \xd0\x98\xd1\x81\xd0\xbf\xd1\x80\xd0\xb0\xd0\xb2\xd0\xbb\xd0\xb5\xd0xbd\xd0\xbe \n\n
10 28.07.2018 13:13:36 1250\nF: \xd0\xa1\xd0\xbe\xd0\xb7\xd0\xb4\xd0\xb0\xd0xbd\xd0\xb8\xd0\x
11
```


Lendo documentos

Arquivos de texto

Especificando o encode UTF-8:

```
from urllib.request import urlopen
paginaTexto = urlopen("http://bgcrm.ru/changes.txt")
print(str(paginaTexto.read(), "utf-8"))
```

A screenshot of a text file's content, likely from a web browser or a text editor. The text is in Russian and contains two entries, each with a date, time, and a number. The first entry is dated 28.07.2018 at 13:13:36 with ID 1250, and the second is dated 28.07.2018 at 03:21:38 with ID 1249. The text is displayed in a monospaced font within a window that has a standard OS-style title bar and a toolbar on the left.

Дата Номер сборки (генерируются автоматически)
А - Добавлено, С - Изменено, F - Исправлено

28.07.2018 13:13:36 1250
F: Создание таблицы для полнотекстового поиска.

28.07.2018 03:21:38 1249
А • Полнотекстовый поиск средствами MySQL.

Lendo documentos

Arquivos de texto

Na maioria das vezes você pode usar UTF-8, até porque esse padrão também manipula caracteres ASCII, o problema é quando você se depara com um arquivo no padrão ISO que são 8% dos sites na internet atualmente.

Em se tratando de arquivos de texto não é possível identificar sua codificação. Existem bibliotecas para tentar deduzir a codificação, mas nem sempre acertam.

Em se tratando de sites, podemos procurar a tag meta que apresenta a codificação da página.

```
<meta charset="utf-8"/>
```

Ao realizar scraping em sites, ao encontrar esta tag, use a codificação que ela recomenda para ler o conteúdo da página.

FIM