

|| ARTHUR MAX C. ROCHA

VISÃO GERAL

A princípio foi pensado em um programa para a criação de perfil para interagir com um outro programa, como o modelo escolhido foi o de prototipagem, o programa mudou muito com relação ao que seria desenvolvido, se tornando um modulo para outro programa, por isso, há apenas uma única classe, que é responsável por todo o trabalho do modulo.

Como não tem muitas funções ficou mais simples e prático organizar tudo em apenas uma única função, e não foi necessário seguir os conceitos de orientação a objetos no escopo do código.

O modulo tem como objetivo criar um perfil para um sistema de recomendação de artigos.

FUNCIONAMENTO

O programa utiliza uma biblioteca livre para fazer a busca das páginas (Web Crawler UFRJ) e um programa em python para fazer o processamento de texto e gerar termos e pesos associados a esses termos obtidos das informações extraídas da página.

O programa em python utiliza uma biblioteca de processamento de texto chamada NLTK, e é chamado pelo programa java por meio de execução de script, o resultado é então retornado para a aplicação em python e salva em um arquivo de texto.

Na sua funcionalidade original, não é necessário telas, etc. Pois como é um modulo, será apenas responsável por receber entradas e retornar saídas. Mas as telas servem para uma melhor interação do usuário com o modulo independente, e para testar ele separadamente.

TELAS E TESTES

Aqui será apresentado as telas de usuário e os arquivos de resultados.

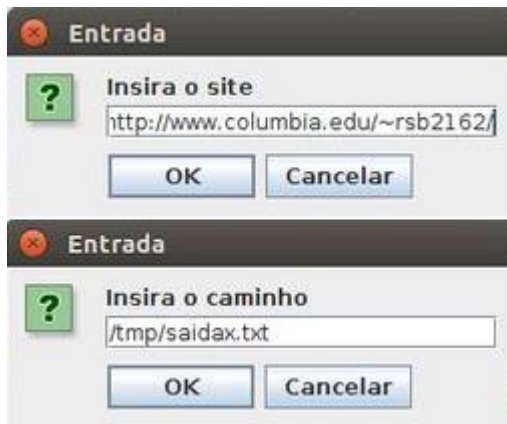
O teste foi feito de forma manual com duas páginas:

<http://www.columbia.edu/~rsb2162/>

<http://www.cs.usask.ca/faculty/julita/>

```
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
1 springerverlagbaker 8.27061450666e-05
2 four 0.0002481184352
3 membershiptmadmug 8.27061450666e-05
4 chennaiseptemb 8.27061450666e-05
5 consortiumpdfwoolf 8.27061450666e-05
6 missingscheu 8.27061450666e-05
7 auc 0.000413530725333
8 pedro 0.0034736580928
9 const 8.27061450666e-05
10 vitaequantit 8.27061450666e-05
11 concret 0.000165412290133
12 pdfnooraei 8.27061450666e-05
13 usefult 8.27061450666e-05
14 spee 8.27061450666e-05
15 spec 0.00157141675626
16 recommendedsecond 8.27061450666e-05
17 digit 8.27061450666e-05
18 classifierdocx 8.27061450666e-05
19 risk 8.27061450666e-05
20 eachaddit 8.27061450666e-05
21 pdffolcomkovarik 8.27061450666e-05
22 bm 0.000165412290133
23 microworld 0.000661649160533
24 readingsbeck 8.27061450666e-05
25 every 8.27061450666e-05
26 noncommerc 8.27061450666e-05
27 psychomet 0.000909767595732
28 affect 0.00661649160533
29 regressionpdf 8.27061450666e-05
30 quantit 0.000578943015466
31 manil 0.0002481184352
32 betterwh 8.27061450666e-05
33 school 0.00107517988587
34 rosnow 0.000165412290133
35 computerassist 8.27061450666e-05
36 upd 8.27061450666e-05
37 wednesday 8.27061450666e-05
38 correl 0.000578943015466
39 matematic 8.27061450666e-05
40 markov 8.27061450666e-05
783 stat 0.00424010641836
784 onsit 8.31393415364e-05
785 oxy 0.000166278683073
786 argu 0.000748254073828
787 pdfdabbish 8.31393415364e-05
788 detectionmonday 8.31393415364e-05
789 bayes 0.00124709012305
790 marino 8.31393415364e-05
791 agapito 0.000332557366146
792 org 0.000166278683073
793 ord 0.000332557366146
794 reflect 0.000249418024609
795 telecommun 8.31393415364e-05
796 brasil 8.31393415364e-05
797 synchron 8.31393415364e-05
798 thing 8.31393415364e-05
799 detectionwednesday 8.31393415364e-05
800 bktmonday 8.31393415364e-05
801 profess 0.00174592617226
802 mello 0.000997672098437
803 think 0.000665114732291
804 first 0.000581975390755
805 origin 8.31393415364e-05
806 degr 0.000249418024609
807 brut 8.31393415364e-05
808 reportsocumpaugh 8.31393415364e-05
809 zai 0.000166278683073
810 acquisit 0.000498836049218
811 fast 0.000249418024609
812 crit 8.31393415364e-05
813 concert 8.31393415364e-05
814 city 8.31393415364e-05
815 district 8.31393415364e-05
816 mitigatedhow 8.31393415364e-05
817 subst 0.000166278683073
818 kauffmanrogoff 0.000249418024609
819 draft 0.00257731958763
820 deploy 0.000166278683073
821 individ 0.000249418024609
822 structures 0.000581975390755
Normal text file
```

SAÍDA



ENTRADAS