Case Crawling (Java)

Objetivo

Você deve implementar um programa para fazer o *crawling* (coleta) de 3 tipos de fonte diferente: Arquivos numa pasta do sistema operacional; Páginas web; Mensagens do Gmail.

Restrições

O programa precisa ser todo implementado em Java utilizando qualquer biblioteca (jar) desde que ela seja de domínio público.

Requisitos funcionais

1) O programa deve ter uma classe chamada Main.java onde o método “public static void main(String[] args)” executa o crawler.
2) O programa deve ler como argumentos uma string indicando o tipo de fonte que será “*crawleado”* e em seguida configurações específicas de cada fonte.
3) Para todos os tipos de fonte, a execução do crawling deve consistir de imprimir no console (System.out.println) o título e corpo do que será considerado um documento.

   Arquivos:
   1) 1 argumento para esse tipo de fonte, a pasta onde o crawling será executado.
   2) Cada arquivo encontrado na pasta será considerado como um documento, sendo o path completo do arquivo o seu título e seu conteúdo textual o seu corpo.
   3) Os arquivos com conteúdo textual são *.txt, *.pdf, *.doc, *.docx, *.ppt, *.pptx. Os arquivos de outro tipo não precisam ter seu conteúdo impresso, apenas o path completo. Para os arquivos que não são *.txt, apenas o “texto útil” deve ser impresso sem nenhum tipo de código ou algo parecido.
   4) Durante a iteração dos arquivos, caso uma pasta seja encontrada, o programa deve entrar nela e iterar por todos os seus arquivos e sub-pastas. A profundidade do crawler deve ser “ilimitada”, ou seja, enquanto o programa encontrar uma sub-pasta ele deve entrar na mesma e seguir com o crawler.

   Web:
   1) 2 argumentos devem ser passados, a url inicial e a profundidade do crawler.
   2) Cada url será considerada como um documento, sendo a url em si o seu título e seu conteúdo textual o seu corpo.
   3) O conteúdo textual de uma url é apenas o seu “texto útil“ que aparece na página, deve ser ignorado códigos html, css, javascript, etc…
   4) Caso a profundida passada seja 0, quer dizer que apenas a url inicial deve ser lida; Todos os links dentro da url inicial são links de profundidade 1, ou seja, caso a profundidade passada seja 1, a url inicial e todas as urls desses links devem ser lidos e assim por diante.

   Gmail:

1) 2 argumentos devem ser passados, username e password de uma conta gmail.
2) Cada mensagem da Inbox será considerada um documento, sendo o subject o seu título e a mensagem o seu corpo.
3) Qualquer anexo deve ser ignorado.

O que eu preciso enviar?

Você deve nos enviar um arquivo compactado com todo o código fonte e qualquer outro arquivo utilizado na implementação do programa.