Cento Universitário UNA

Sistemas de Informação

Tecnologias Emergentes

Práticas de Laboratório

Wesley Dias Maciel

2019/02

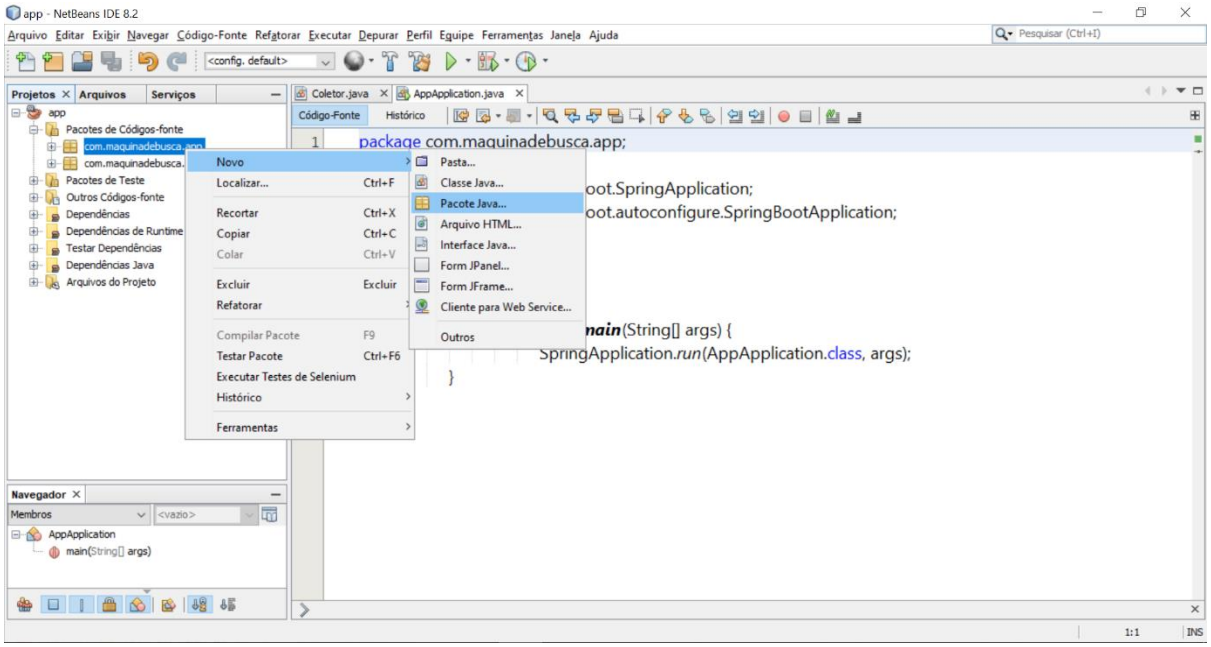
Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/02

Spring Boot, Jsoup

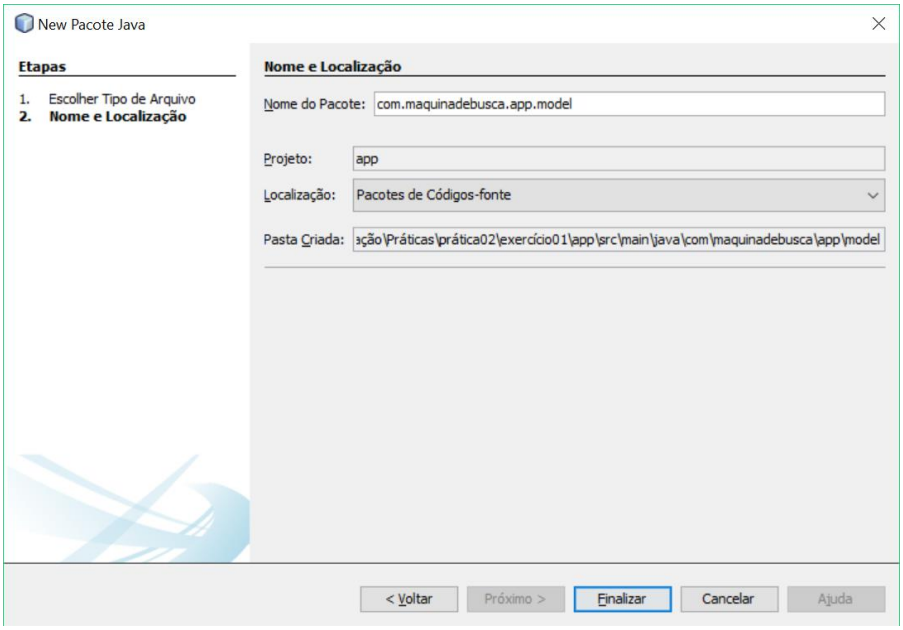
Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/02

Prática 02

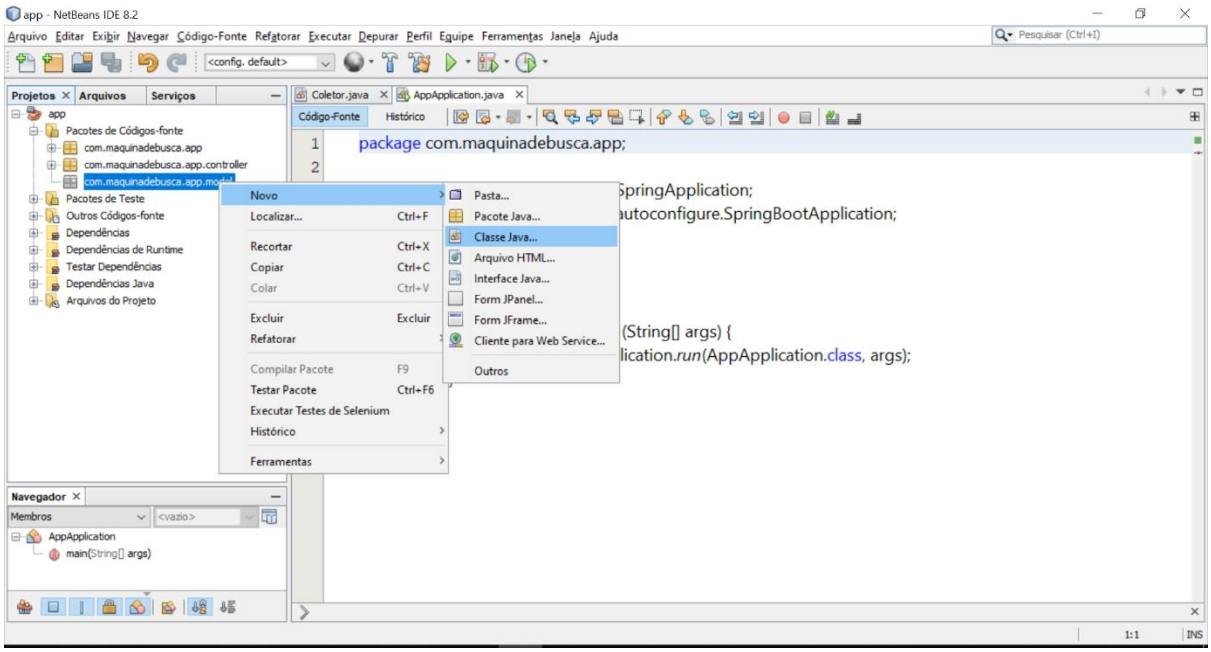
1) Altere o projeto da prática 01, exercício 02, criando um novo pacote. Clique com o botão direito do mouse sobre o pacote com.maquinadebusca.app. Em seguida, clique em Novo. Depois, clique em Pacote Java.



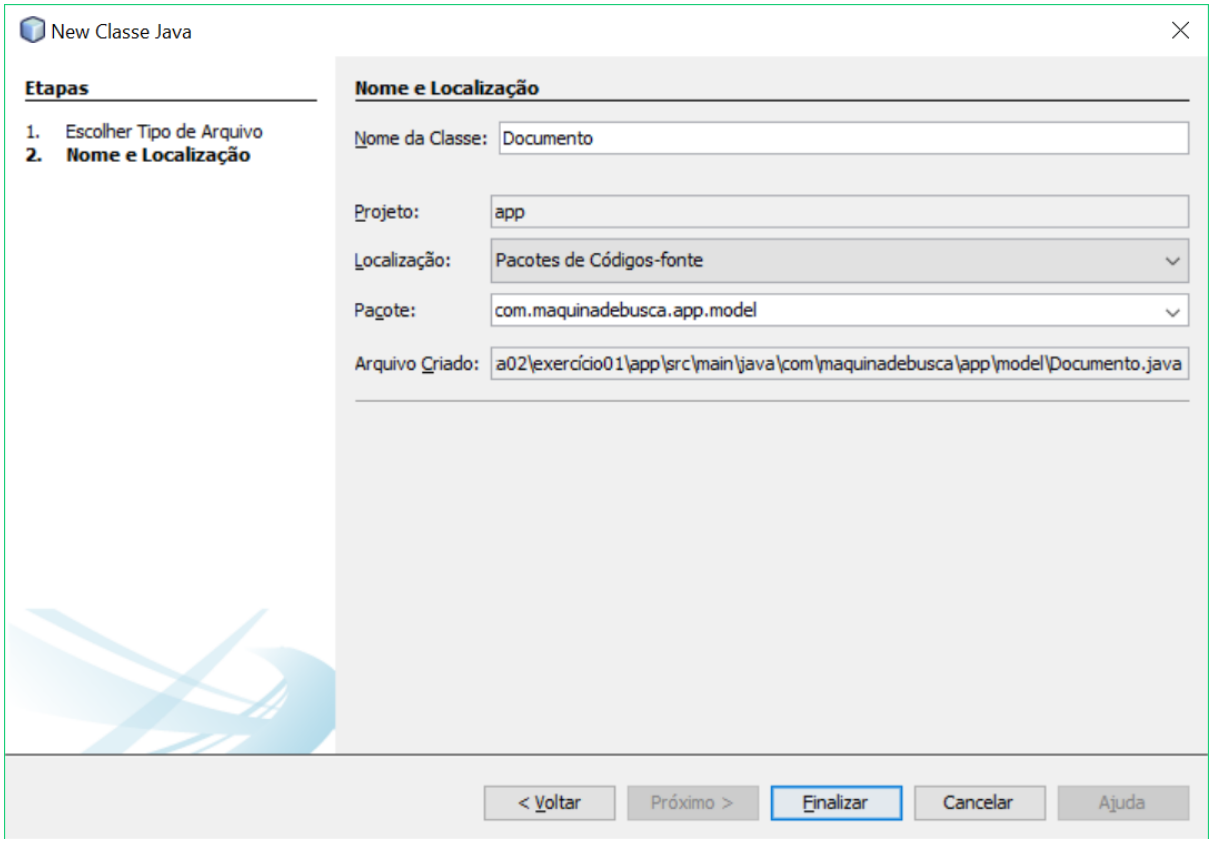
No campo Nome do Pacote da janela que se abre, informe o nome model, conforme apresentado na figura abaixo (com.maquinadebusca.app.model). Em seguida, clique em Finalizar.



Clique com o botão direito do mouse sobre o pacote com.maquinadebusca.app.model. Em seguida, clique em Novo. Depois, clique em Classe Java.



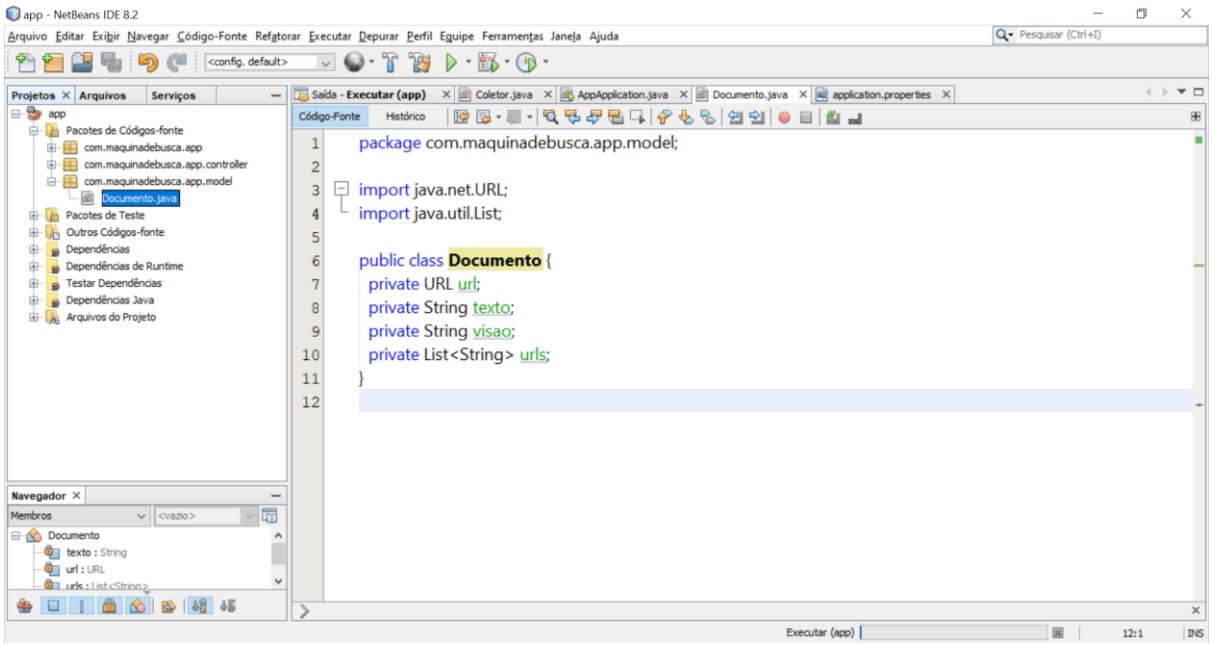
No campo Nome da Classe da janela que se abre, informe o nome Documento, conforme apresentado na figura abaixo. Em seguida, clique em Finalizar.



Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/02

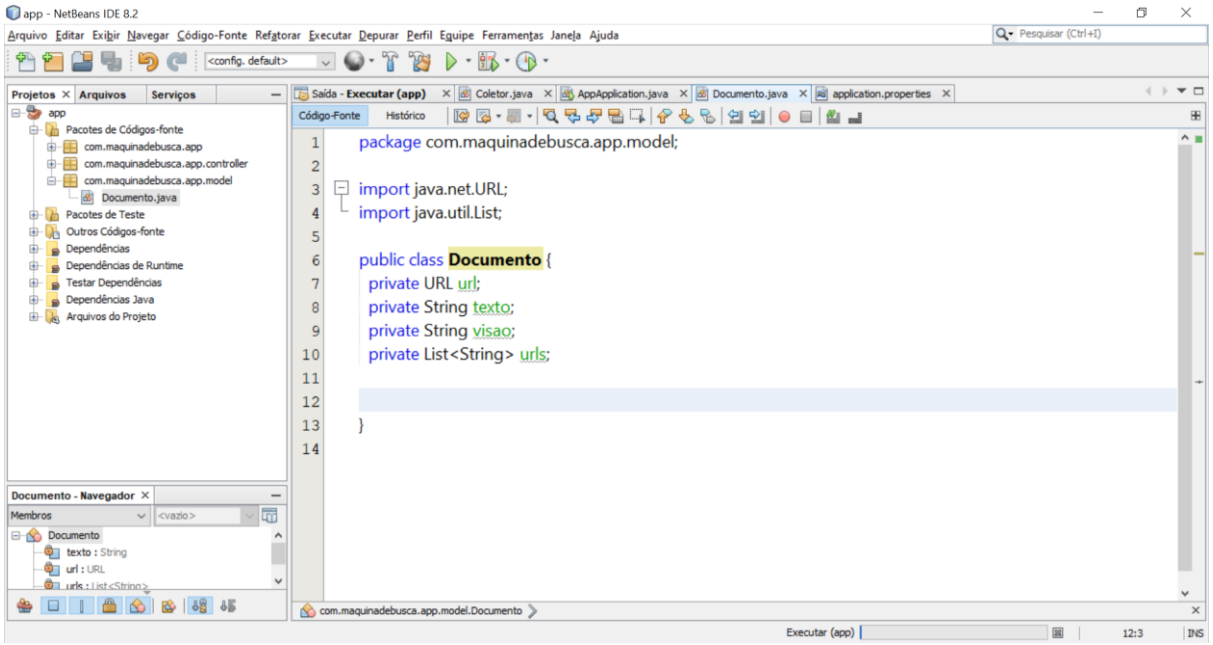
Altere o conteúdo da classe Documento, conforme indicado abaixo.

```
package com.maquinadebusca.app.model;

import java.net.URL;
import java.util.List;

public class Documento {
  private URL url;
  private String texto;
  private String visao;
  private List<String> urls;
}
```

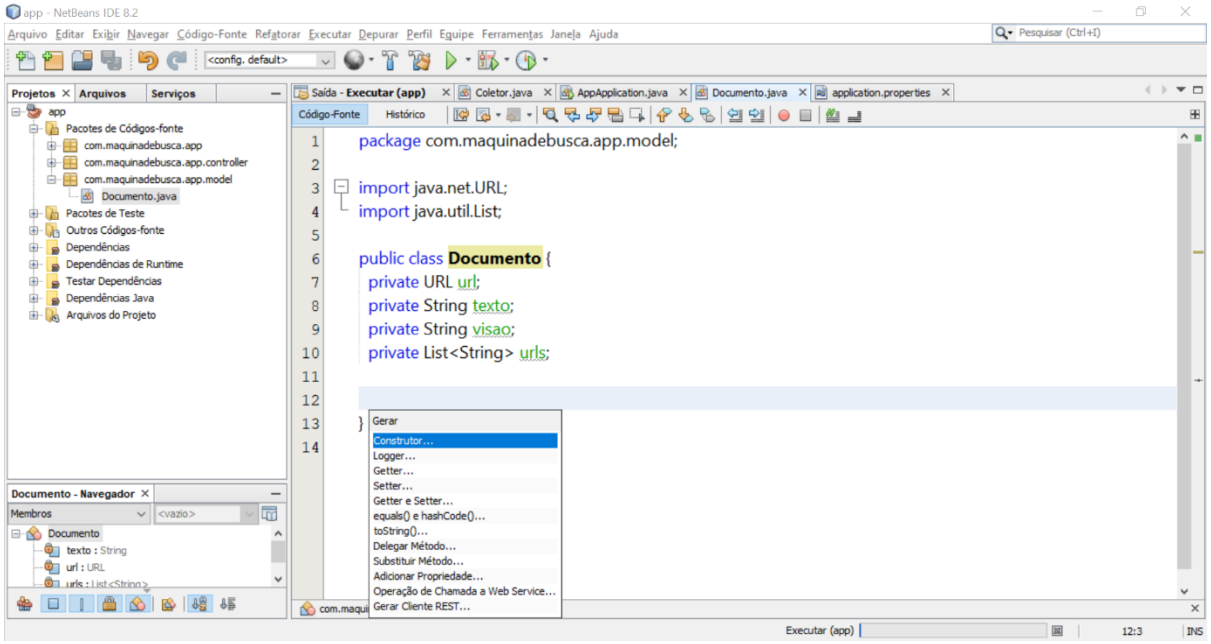


Crie duas linhas em branco após o último atributo da classe Documento, private List<URL> urls. Deixe o cursor nesse ponto da classe.

Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/02

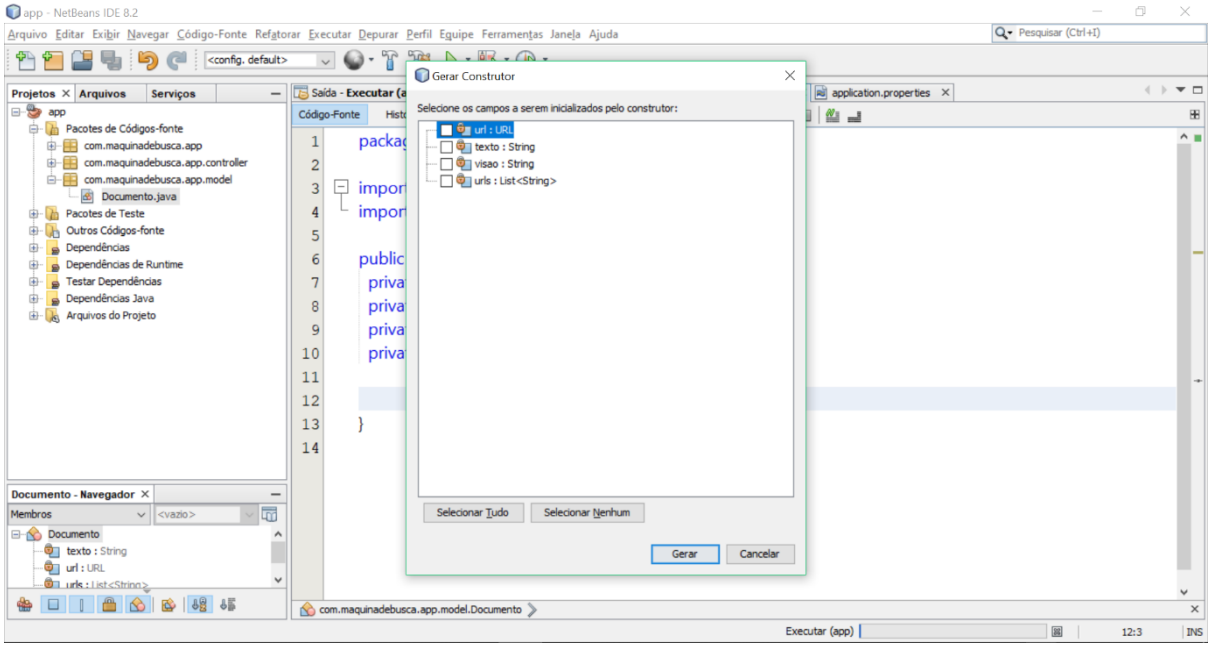


Em seguida, tecle “ALT + Insert”. Na janela que se abre, clique sobre a opção Construtor.

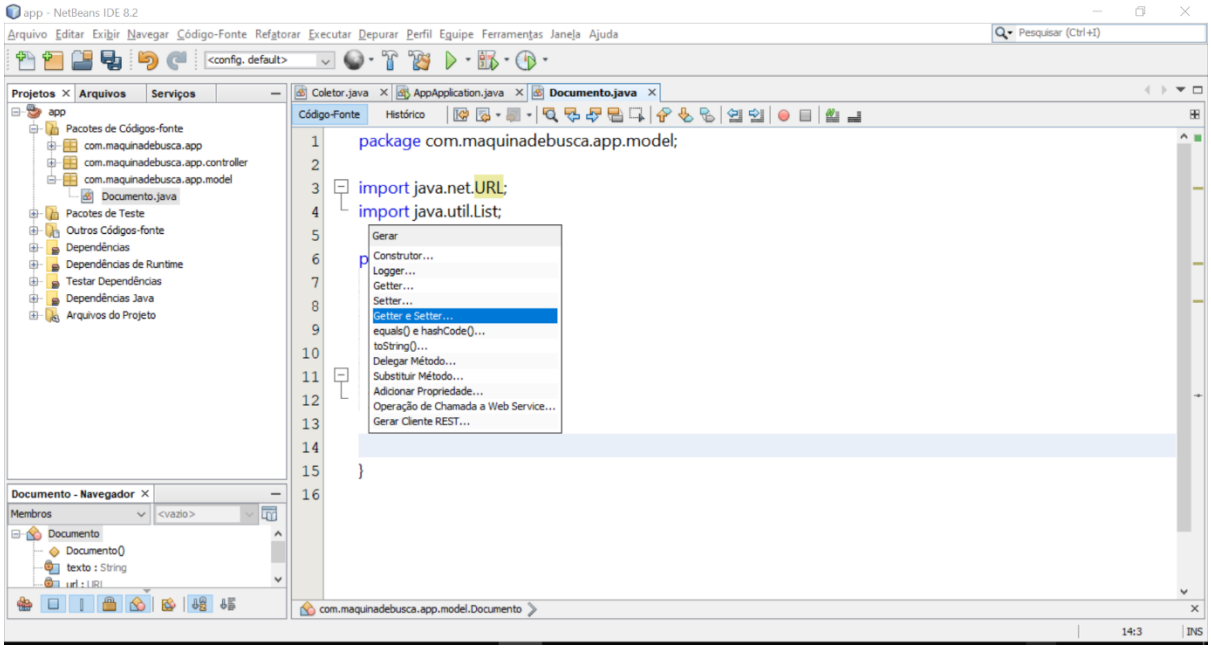


Na janela que se abre, simplesmente clique em Gerar.

Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/02

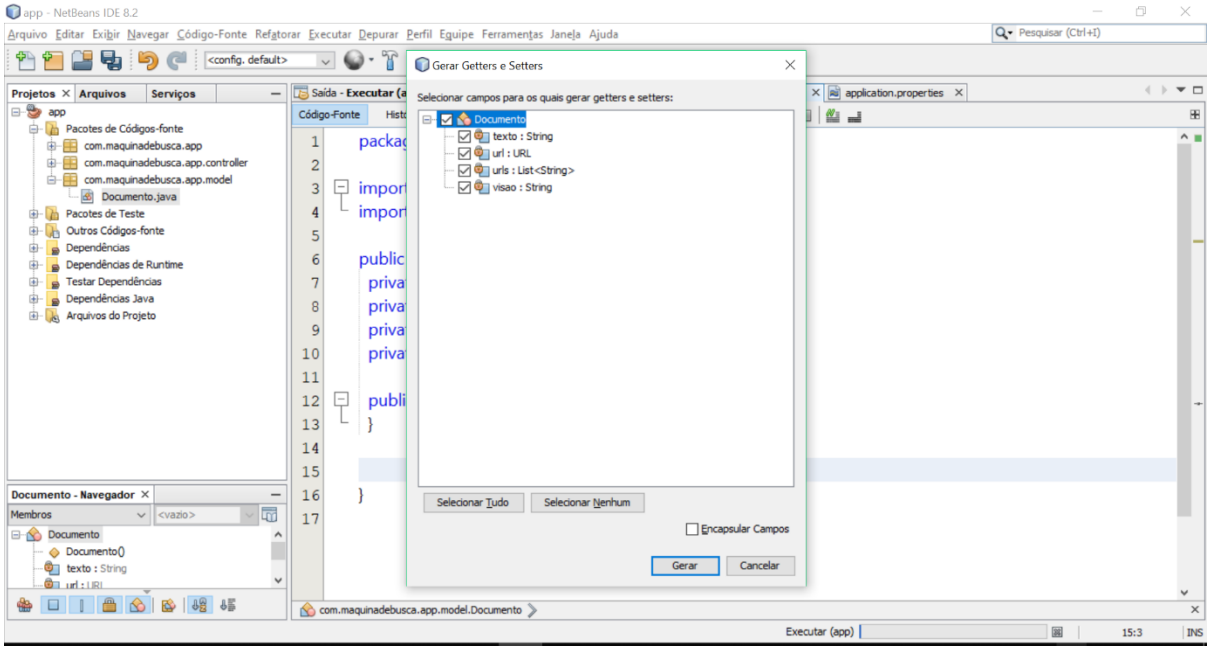


Novamente, tecle “ALT + Insert”. Na janela que se abre, clique sobre Getter e Setter.

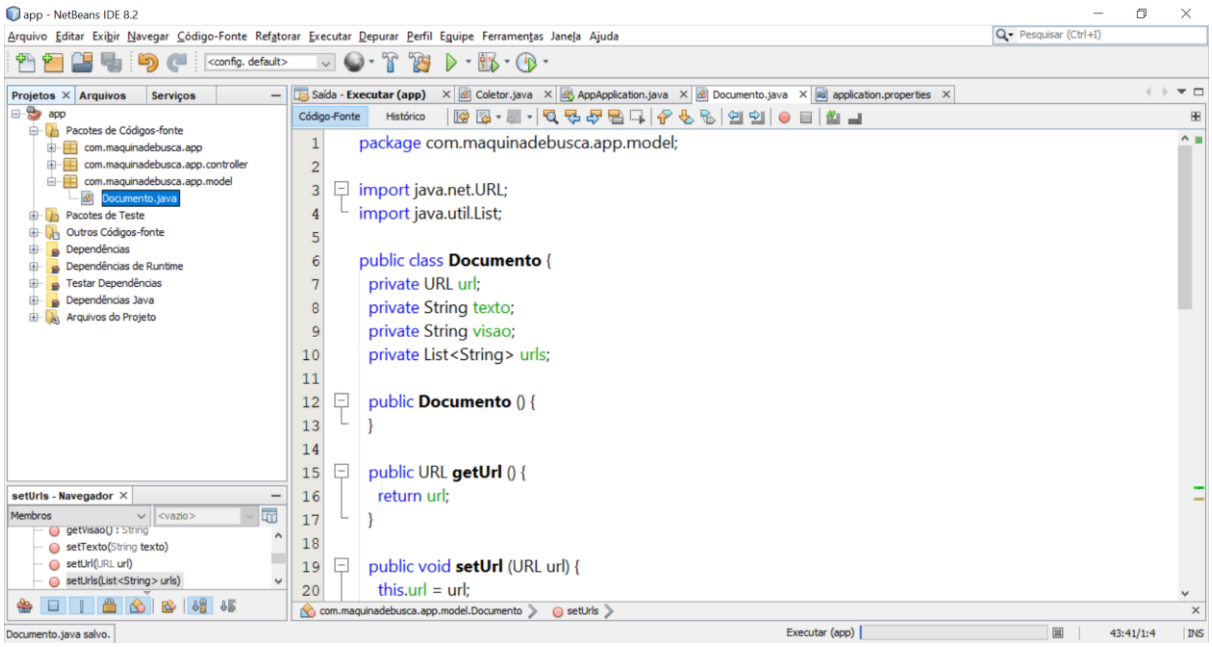


Na janela que se abre, selecione todos os atributos da classe Documento, conforme ilustrado abaixo. Em seguida, clique em Gerar.

Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/02



Observe a definição gerada para a classe Documento.



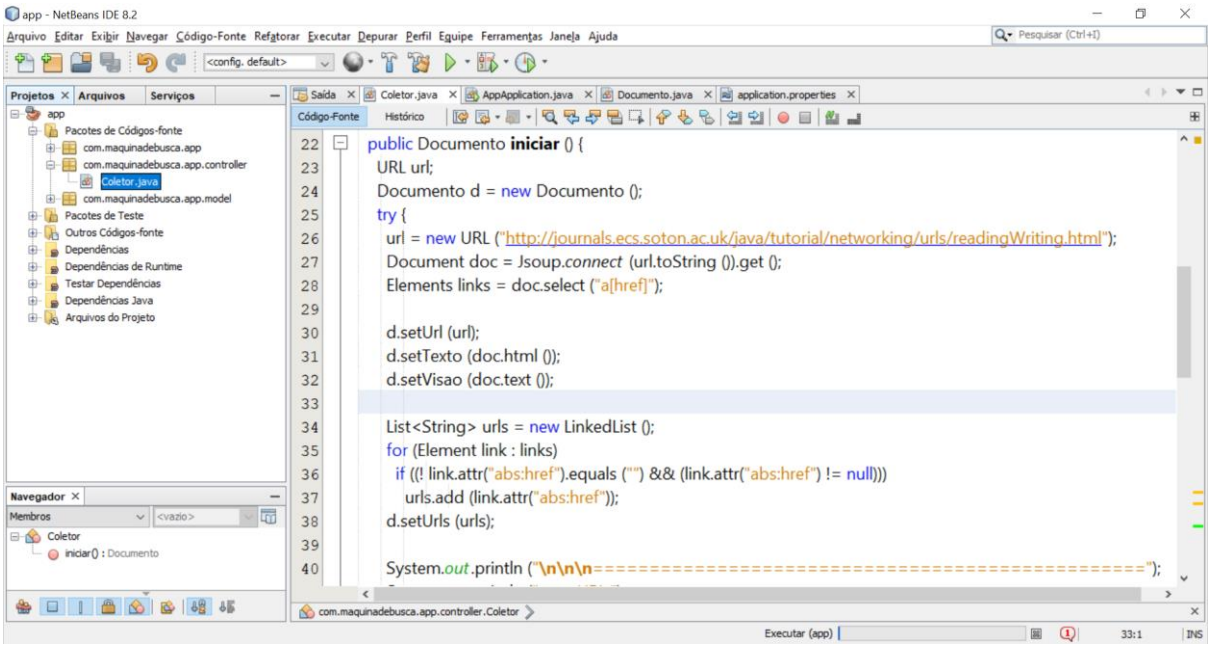
Abra a classe Coletor. Altere o método iniciar () da classe Coletor que foi criado na prática 01, exercício 02, conforme indicado abaixo.

```
package com.maquinadebusca.app.controller;

import java.net.URL;
import java.util.List;
import java.util.LinkedList;
```

Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/02

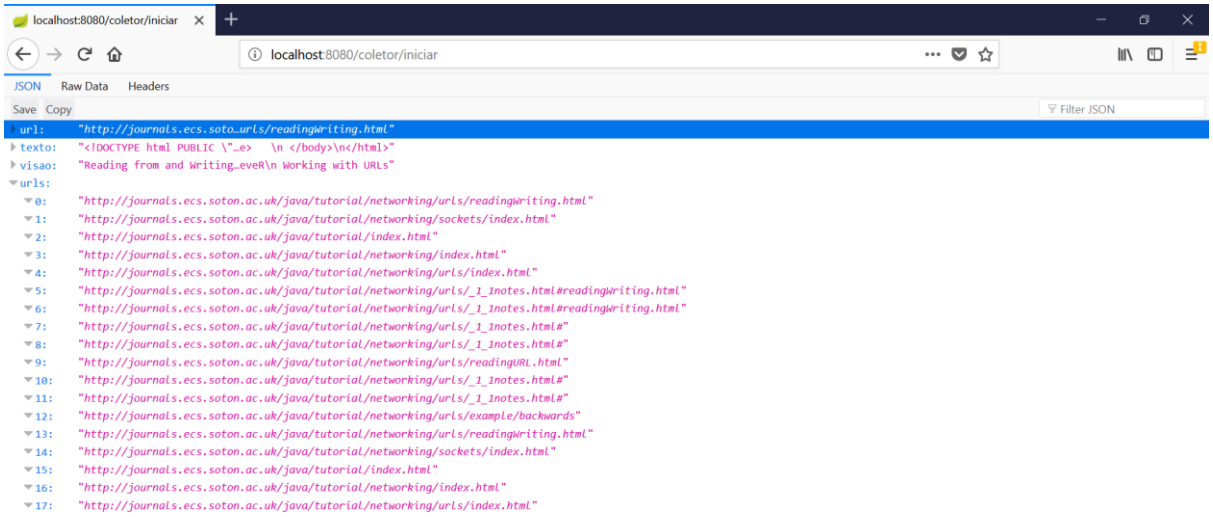
```
import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;
import org.springframework.http.MediaType;
import org.springframework.web.bind.annotation.GetMapping;
import org.springframework.web.bind.annotation.RestController;
import org.springframework.web.bind.annotation.RequestMapping;
import com.maquinadebusca.app.model.Documento;

@RestController
@RequestMapping ("/coletor") // URL: http://localhost:8080/coletor
public class Coletor {

  // URL: http://localhost:8080/coletor/iniciar
  @GetMapping (value = "/iniciar", produces =
MediaType.APPLICATION_JSON_UTF8_VALUE)
  public Documento iniciar () {
    URL url;
    Documento d = new Documento ();
    try {
      url = new URL
      ("http://journals.ecs.soton.ac.uk/java/tutorial/networking/urls/readingWriting.html");
      Document doc = Jsoup.connect (url.toString ()).get ();
      Elements links = doc.select ("a[href]");

      d.setUrl (url);
      d.setTexto (doc.html ());
      d.setVisao (doc.text ());

      List<String> urls = new LinkedList ();
      for (Element link : links)
        if ((! link.attr("abs:href").equals ("") && (link.attr("abs:href") != null)))
          urls.add (link.attr("abs:href"));
      d.setUrls (urls);

      System.out.println
("\n\n\n================================================");
      System.out.println (">>> URL:");
      System.out.println ("================================================");
      System.out.println (d.getUrl ());
```

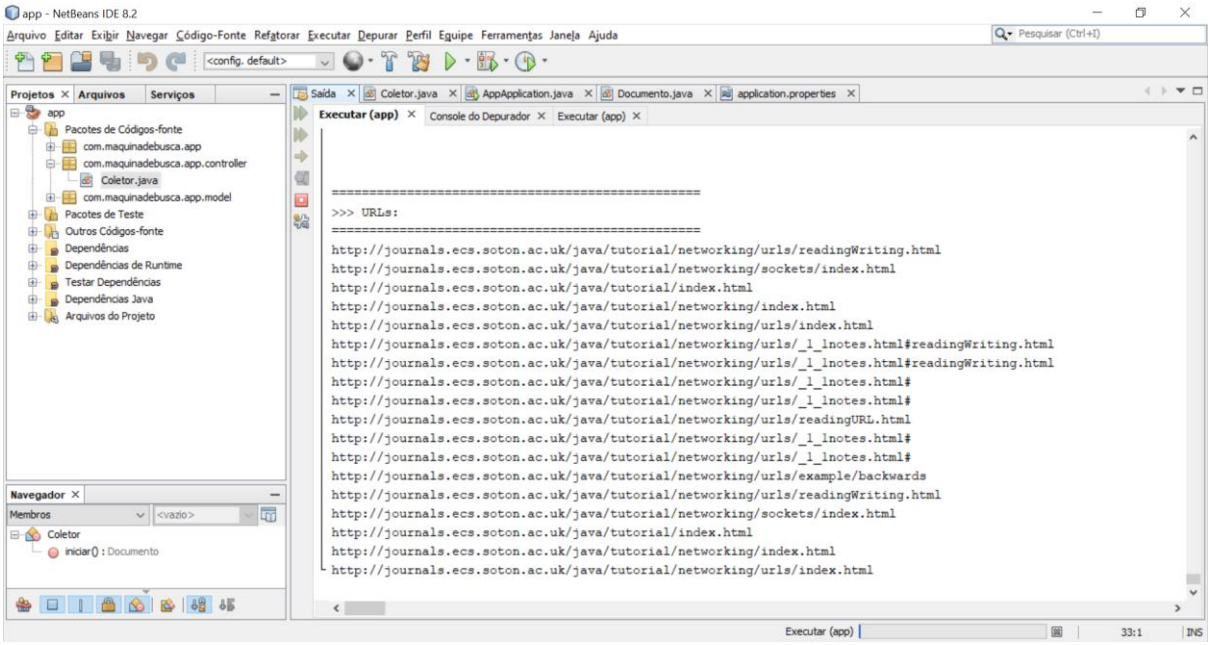
```
    System.out.println
("\n\n\n================================================");
    System.out.println (">>> Página:");
    System.out.println ("================================================");
    System.out.println (d.getTexto ());

    System.out.println
("\n\n\n================================================");
    System.out.println (">>> Visão:");
    System.out.println ("================================================");
    System.out.println (d.getVisao ());

    System.out.println
("\n\n\n================================================");
    System.out.println (">>> URLs:");
    System.out.println ("================================================");
    urls = d.getUrls ();
    for (String u: urls)
      System.out.println (u);
   } catch (Exception e) {
    System.out.println ("Erro ao coletar a página.");
    e.printStackTrace ();
   }
   return d;
  }
}
```



Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/02

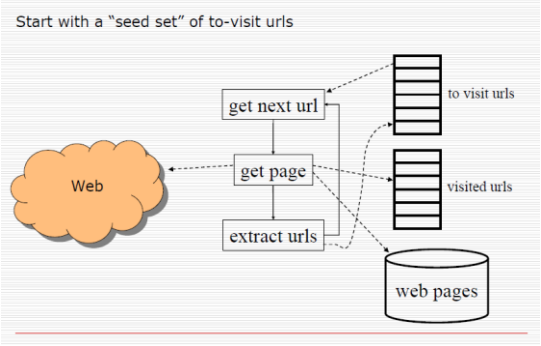
Execute o projeto e observe a saída no navegador e no console do servidor. Use o navegador **Firefox**, para que você consiga uma melhor visualização da resposta no formato JSON.





2) Altere o projeto, criando o pacote com.maquinadebusca.app.model.service.
3) No pacote com.maquinadebusca.app.model.service, crie a classe ColetorService.
4) Refatore a classe Coletor, colocando o conteúdo do método iniciar () em um método da classe ColetorService.

5) Crie uma chamada para o novo método da classe ColetorService no escopo do método iniciar () da classe Coletor, para que o projeto continue executando normalmente.
6) Altere o projeto, para que a tarefa de coleta da máquina de busca inicie com um conjunto de URLs sementes. A página de cada URL semente deve ser coletada.
7) Altere o projeto, para que as novas URLs encontradas e extraídas das páginas anteriormente coletadas também sejam coletadas, conforme ilustrado na figura abaixo.



8) Garanta que o protocolo de exclusão de robôs na Web seja respeitado, para minimizar a carga nos servidores Web:
    a. Respeitar um intervalo de tempo entre duas coletas consecutivas.
    b. Exemplo: não realizar mais de 1 requisição em um mesmo servidor a cada 10 segundos.
    c. Protocolo para exclusão de robôs:
        i. Concede acesso limitado ao robô.
        ii. Há páginas que não podem ser coletadas.
        iii. http://www.robotstxt.org/orig.html
9) Na classe com.maquinadebusca.app.model.service.ServicoColetor, crie um método para identificação e exclusão de URLs repetidas, iguais.
10) Na classe com.maquinadebusca.app.model.service.ServicoColetor, crie um método para filtrar stopwords em cada página coletada.
11) O atributo visao da classe com.maquinadebusca.app.model.Documento deve receber o conteúdo da página sem as tags HTML, sem os sinais de pontuação e sem as stopwords. Além disso, os caracteres de todos os termos devem ser convertidos para minúsculo. Alguns links para listas de stopwords em português:
    a. https://gist.github.com/alopes/5358189
    b. http://snowball.tartarus.org/algorithms/portuguese/stop.txt
12) Forme uma coleção de documentos com no mínimo 50.000 documentos HTML.