14.08.2024

IBaroCrawler Proje Dokümantasyonu

Proje Hakkında Genel Bilgi

Proje Amacı

IBaroCrawler, İstanbul Barosu web sitesinden PDF belgelerini otomatik olarak indirip, bu belgelerden metin çıkarma işlemini gerçekleştiren bir Java uygulamasıdır. Bu proje, elde edilen metinleri yerel dosya sisteminde saklar ve istenirse https://istanbulbarosu.org.tr/yayinlar.aspx yayınlarındaki Baro Dergileri bölümünden güncel olarak yayınlanmış pdf dosyalarını ayrıştırır ve işleyerek MySQL veritabanına kaydeder.

Proje Özeti

- **Proje Adı**: IBaroCrawler
- **Yazılım Dili**: Java
- **Kullanılan Kütüphaneler:**
    - **Jsoup**: HTML işlemleri için.
    - **PDFBox**: PDF dosyalarından metin çıkarmak için.
    - **MySQL Connector**: Veritabanı işlemleri için.
    - **Jackson Databind**: JSON verisi ile çalışmak için.
    - **Log4j**: Loglama amacıyla.
    - **HttpClient**: HTTP isteklerini yönetmek için.
- **Proje Yapısı**: Maven projesi.

Proje Yapısı

1. Sınıflar ve Metodlar

App Sınıfı

App sınıfı projenin merkezidir ve tüm iş akışını yöneten sınıftır. Bu sınıf, PDF indirme, metin çıkarma, dosya kaydetme ve veritabanı etkileşimlerinden sorumludur.

- **Metodlar**:
  - **main(String[] args):** Uygulamanın başlangıç noktası. HTTP isteklerini yapar, bağlantıları işler ve dosya indirme sürecini başlatır.

1. downloadFile(String href, File pdfFile)

Bu metod, verilen href URL'sinden bir PDF dosyasını indirir ve bu dosyayı belirtilen pdfFile nesnesi olarak kaydeder. Metod, HTTP bağlantısı kurar, PDF içeriğini okur ve yerel dosya sistemine yazar.

- **Parametreler**:
  - **href**: PDF dosyasının indirileceği URL.
  - **pdfFile**: PDF dosyasının kaydedileceği yerel dosya nesnesi.
- **İstisnalar**:
  - **IOException**: Dosya indirme veya kaydetme sırasında oluşabilecek herhangi bir giriş/çıkış hatası durumunda fırlatılır.

2. extractTextFromPdf(File pdfFile, String href, String pdfFileName, MySqlConnection mySqlConnection)

Bu metod, verilen PDF dosyasından metin çıkarır. Metni çıkarılan her makale, Article sınıfına dönüştürülür ve MySQL veritabanına kaydedilir.

- **Parametreler**:
  - **pdfFile**: Metin çıkarılacak PDF dosyası.
  - **href**: PDF dosyasının kaynak URL'si.
  - **pdfFileName**: PDF dosyasının adı.
  - **mySqlConnection**: Veritabanı bağlantısını yöneten nesne.
- **Dönüş Tipi:**
  - **List<Article>:** PDF dosyasından çıkarılan makalelerin listesi.

3. saveTextToTxtFile(List<Article> articles, String pdfFileName)

Bu metod, Article nesnelerinden oluşan bir listeyi alır ve her makalenin içeriğini .txt formatında dosyalara kaydeder. Dosya isimleri, PDF dosyasının ismine dayalı olarak oluşturulur.

- **Parametreler**:
  - **articles**: İçeriği .txt formatında kaydedilecek makalelerin listesi.
  - **pdfFileName**: Kaydedilecek dosyaların isimlerinin türetileceği PDF dosya adı.

4. printArticleDetails(List<Article> articles)

Bu metod, verilen Article nesnelerinin detaylarını konsola yazdırır. Bu işlem, hata ayıklama ve gözden geçirme süreçlerinde kullanışlıdır.

- **Parametreler**:
    - **articles**: Detayları konsola yazdırılacak makalelerin listesi.

5. getFileName(String url)

Bu metod, verilen bir URL'den dosya adını alır. URL içindeki dosya ismini elde etmek için son / karakterinden sonrasını alır.

- **Parametreler**:
    - **url**: Dosya adının çıkarılacağı URL.
- **Dönüş Tipi:**
    - **String**: URL'den çıkarılan dosya adı.

6. extractArticleInfo(String text, int startPage, String url, String filename)

Bu metod, verilen metin parçasından makale bilgilerini çıkarır ve bir Article nesnesine dönüştürür. Makalenin başlık, yazar, başlangıç sayfası gibi detaylarını analiz eder.

- **Parametreler:**
    - **text**: İşlenecek metin parçası.
    - **startPage**: Makalenin başladığı sayfa numarası.
    - **url**: Makalenin bulunduğu PDF dosyasının URL'si.
    - **filename**: Makalenin kaydedileceği dosya adı.
- **Dönüş Tipi:**
    - **Article**: Çıkarılan makale bilgilerini içeren Article nesnesi.

7. extractTitle(String text)

Bu metod, verilen metin parçasından makale başlığını çıkarır. Başlık, genellikle metnin ilk satırlarından elde edilir.

- **Parametreler**:
    - **text**: Başlığın çıkarılacağı metin parçası.
- **Dönüş Tipi**:
    - **String**: Metinden çıkarılan başlık.

8. extractAuthor(String text)

Bu metod, verilen metin parçasından makale yazarını çıkarır. Yazar adı, genellikle başlıktan hemen sonra gelir.

- **Parametreler:**
  - **text**: Yazarın çıkarılacağı metin parçası.
- **Dönüş Tipi:**
  - **String**: Metinden çıkarılan yazar adı.

9. writeArticleToFile(Article article, String fileName)

Bu metod, Article nesnesinin içeriğini belirtilen dosya adına yazar. Bu işlem, makale içeriğini dosya sisteminde saklamak için kullanılır.

- **Parametreler:**
  - **article**: Dosyaya yazılacak makale.
  - **fileName**: Makalenin yazılacağı dosya adı.

Article Sınıfı

Article sınıfı, bir makale için gerekli tüm bilgileri içerir ve bu bilgilerin tutulduğu veri yapısını temsil eder.

- **Alanlar (Fields):**
  - **title**: Makale başlığı.
  - **author**: Makale yazarı.
  - **startPage**: Makalenin başladığı sayfa numarası.
  - **pageCount**: Makalenin toplam sayfa sayısı.
  - **fileName**: PDF dosya adı.
  - **url**: PDF dosyasının URL'si.
  - **text**: Çıkarılan metin.

MySqlConnection Sınıfı

MySqlConnection sınıfı, MySQL veritabanı ile etkileşimden sorumludur. Makalelerin veritabanına eklenip eklenmediğini kontrol eder ve ekleme işlemini gerçekleştirir.

- **Metodlar**:
  - **isArticleExists(String title, String author)**: Veritabanında aynı başlık ve yazara sahip bir makalenin var olup olmadığını kontrol eder.
  - **insertArticle(Article article):** Yeni bir makaleyi veritabanına ekler.

2. İş Akışı

1. **Bağlantıların Alınması ve İşlenmesi:**
    - App sınıfı, HTTP istekleri aracılığıyla PDF bağlantılarını elde eder ve bu bağlantıları links.txt dosyasına kaydeder.
2. **PDF İndirme:**
    - Her bağlantı için downloadFile metodu çağrılır ve dosya yerel olarak indirilir.
3. **Metin Çıkarma:**
    - İndirilen her PDF dosyası extractTextFromPdf metodu ile işlenir ve metin çıkarılır.
4. **Dosya Kaydetme:**
    - Çıkarılan metinler saveTextToTxtFile metodu ile .txt dosyalarına kaydedilir.
5. **Veritabanı Entegrasyonu:**
    - Eğer bir veritabanı bağlantısı yapılandırılmışsa, MySqlConnection sınıfı kullanılarak makaleler veritabanına eklenir.

3. Proje Bağımlılıkları ve Kurulum

Bağımlılıklar

- **Jsoup**: org.jsoup:jsoup:1.13.1
- **PDFBox**: org.apache.pdfbox:pdfbox:2.0.24
- **MySQL** Connector: mysql:mysql-connector-java:8.0.25
- **Jackson Databind**: com.fasterxml.jackson.core:jackson-databind:2.12.3
- **Log4j**: log4j:log4j:1.2.17
- **HttpClient**: org.apache.httpcomponents:httpclient:4.5.13

Kurulum Adımları

1. **Projenin İndirilmesi:**
    - Proje kaynak kodunu bir git deposundan veya doğrudan ZIP dosyası olarak indirip çıkartın.
2. **Maven Bağımlılıklarını Yükleme**:
    - pom.xml dosyasını kullanarak Maven bağımlılıklarını yükleyin.
    - mvn clean install komutu ile projeyi derleyin ve gerekli bağımlılıkları indirin.
3. **Veritabanı Yapılandırması:**
    - MySQL üzerinde mavendb adında bir veritabanı oluşturun.
    - Article tablosunu şu SQL komutuyla oluşturun:

```
CREATE TABLE Article (
  id INT AUTO_INCREMENT PRIMARY KEY,
  title VARCHAR(255) NOT NULL,
  author VARCHAR(255),
  startPage INT,
  pageCount INT,
  fileName VARCHAR(255),
  url VARCHAR(255),
  text TEXT
);
```

4. **Yapılandırma Dosyası:**
    - Veritabanı bağlantı bilgilerini config.properties dosyasına ekleyin:

    db.url=jdbc:mysql://localhost:3306/mavendb

    db.user=root

    db.password=[Şifreniz]

5. **Projeyi Çalıştırma**:
    - App sınıfını çalıştırarak proje işlemlerini başlatın.

6. Projenin Durumu ve Eksiklikler

Proje, veri toplama ve işleme konularında güçlü bir temel oluşturmakla birlikte, bazı eksiklikler bulunmaktadır. Gönüllü stajımı maalesef beklenenden daha erken sonlandırmak zorunda kaldığım için proje tam anlamıyla tamamlanamadı. Bu durum, özellikle verilerin işlenmesi ve tam anlamıyla doğrulanması aşamalarında bazı eksikliklere yol açtı.

Projeyi daha ileri taşıma ve tam anlamıyla tamamlayabilme fırsatım olsaydı, veri işleme süreçlerinde daha fazla test ve hata ayıklama yaparak, sonuçların daha güvenilir olmasını sağlayabilirdim. Yine de, mevcut durumda proje, temel işlevsellikleri yerine getirebilmekte ve veri toplama ile ilgili süreçleri büyük ölçüde yönetebilmektedir.

Bu eksikliklerin farkında olarak, projeyi devralacak olan geliştiriciye elimden gelen en iyi dokümantasyonu sağlamak için çaba gösterdim. Umarım bu süreçte yaşanan durumlar anlaşılır ve projenin devamında kolaylık sağlar.

7. İletişim ve Destek

İletişim Bilgileri:
Projeyle ilgili herhangi bir sorunuz veya desteğe ihtiyaç duyduğunuzda benimle aşağıdaki bilgiler üzerinden iletişime geçebilirsiniz:

- İsim: Yiğit Alp Kaynak
- Telefon: +90 537 473 5458
- E-posta: yigitalpkaynak1@gmail.com
- GitHub: https://github.com/yigitalpkaynak/IBaroCrawler

Ek Belgeler:
Projeyle ilgili ekstra dokümanlar, referanslar ve gerekli linkler yukarıdaki GitHub bağlantısında mevcuttur.

Proje ile ilgili herhangi bir sorunuz olursa, yukarıdaki iletişim bilgileri üzerinden benimle iletişime geçebilirsiniz. Projeye dair her türlü güncelleme ve bilgiye GitHub bağlantısından ulaşabilirsiniz.

Yardımcı olabileceğim herhangi bir konu olursa, çekinmeden bana ulaşabilirsiniz.

Saygılarımla,

Yiğit Alp Kaynak