

Dossier Next Report

Resumen. Dossier Next Report es una aplicación que permite generar resúmenes automáticos y análisis estadístico a partir de la base de datos de una aplicación Dossier Next. Simplemente se ha de indicar un intervalo de fechas y se generarán automáticamente varios resúmenes: uno general, teniendo en cuenta todos los documentos de ese período, y otro para cada entidad detectada dentro de los documentos de ese período. De esta forma, es posible obtener la información más importante de una colección de textos y de cada una de las entidades mencionadas en esa colección.

Summary. Dossier Next Report is an application that allows you to generate automatic summaries and statistical analysis from the database of a Dossier Next application. All you have to do is indicate a date range and several summaries will be automatically generated: a general one, taking into account all the documents of that period, and another one for each entity detected within the documents of that period. In this way, it is possible to obtain the most important information from a collection of texts and from each one of the entities mentioned in that collection.

Especificaciones técnicas / Technical Specifications

Lenguaje de programación <i>Development Language</i>	Java, Javascript, HTML, CSS
Entorno Operativo <i>Operating Environment</i>	Linux, Windows, Mac
Versión <i>Version</i>	1.0

Requerimientos / Requirements

- **Java 8** (o superior)
- **Maven 3** (o superior)
- 1GB de RAM o superior
- 100 MB de espacio en disco aproximadamente, pero depende del número de documentos que se utilizarán para generar los resúmenes automáticos

Instalación / Installation

- Descargar el fichero con la aplicación y descomprimirlo.
- Descargar el modelo de entidades nombradas (NER) de OpenNLP:
<http://opennlp.sourceforge.net/models-1.5/es-ner-person.bin>
- Crear un fichero de texto plano con el nombre de las entidades que no se desea procesar (llamadas *entidades spam*), una por línea.
- Crear un fichero de texto plano con las frases que no se desea incluir en los resúmen (llamadas *frases spam*), una por línea.
- Ejecutar el siguiente comando:
mvn clean package -Pshade

Ejecución / Run

- Dentro de la carpeta descargada, ejecutar el siguiente comando:

```
java -jar core/target/oportunity-report-core-0.1-shaded.jar \  
  --db-url URL_BASE_DE_DATOS \  
  --db-username USUARIO_BASE_DE_DATOS \  
  --db-password CONTRASEÑA_BASE_DE_DATOS \  
  --openlp-model RUTA_MODELO_NER_OPENLP \  
  --date DD/MM/YYYY \  
  --period [day/week/month] \  
  --spam-content-file RUTA_FICHERO_FRASES_SPAM \  
  --spam-entity-file RUTA_FICHERO_ENTIDADES_SPAM \  
  --o CARPETA_DESTINO
```

Dependencias / Dependencies

El listado completo de dependencias se encuentra en el archivos **core/pom.xml**. Aquí se incluye un listado de las dependencias que se descargarán de varios repositorios:

Dependencias Java / Java Dependencies · pom.xml

```
io.airlift:airline:jar:0.8
org.mariadb.jdbc:mariadb-java-client:jar:2.4.4
es.ua.oportunity:summarization-core:jar:0.1
org.apache.commons:commons-dbcp2:jar:2.7.0
opennlp:tools:jar:1.5.0
```

Estructura de ficheros / Files' structure

```
./core/pom.xml
./core/src/main/resources/template-topic.html
./core/src/main/resources/template-keyword.html
./core/src/main/resources/template-source.html
./core/src/main/resources/template-sentence.html
./core/src/main/resources/template-entity.html
./core/src/main/resources/template.html
./core/src/main/java/es/ua/oportunity/report/Main.java
./pom.xml
./README.md
```