Detector y clasificador de tripletas (DCTR)

Mayo 28, 2016

Título Detector y clasificador de tripletas en R para Text Mining (DCTR)

Versión 0.1

Fecha 28 Mayo 2016

Dependencias R (>=3.2.2)

Librerías tm, xlsx, stringr

RequisitosDelSistema FreeLing 3.1, Java

Descripción Conjunto de métodos en R para detectar y clasificar tripletas.

Licencia GNU General Public License 3.0

URL https://github.com/robertomoreu/DCTR.git

Repositorios_adicionales –

Requiere compilación Sí

Autor Roberto Moreu Rubio <rmoreu92@gmail.com>

Fecha/Publicación 5 Junio 2016

Crear_corpus	Crear un corpus a partir de un fichero

Descripción

Crea un corpus a partir de texto en un fichero de texto plano, CSV o Excel.

Utilización

Crear_corpus(fichero, indiceHoja=1, cabecera=FALSE, separador="\n", columna=1)

Argumentos

fichero	ruta completa al fichero que se pretende importar. Puede ser .txt, .csv o .xlsx.
indiceHoja	indica el índice de la hoja en que se encuentra el texto en caso de un fichero Excel. Por defecto 1.
cabecera	valor booleano que indica si el fichero contiene cabecera. Por defecto FALSE.
separador	cadena con el separador utilizado para separar los campos del CSV. Por defecto "\n".
columna	indica el índice de la columna en la que se encuentra el texto dentro del fichero. Solo necesario en Excel y CSV. Por defecto 1.

Salida

Objeto de la clase VCorpus o Corpus con una entrada por cada línea del documento.

Ejemplos

```
corpusQuijoteXLSX <- Crear_corpus(\D:/Quijote.xlsx", cabecera=TRUE,
    columna = 2, indiceHoja=1)
corpusQuijoteCSV <- Crear_corpus("D:/Quijote.csv", separador=";")
corpusQuijoteTXT <- Crear_corpus("D:/Quijote.txt")
corpusQuijoteCSV <- Crear_corpus("D:/Quijote.csv", separador=";",
    cabecera=TRUE)
```

Preparar_corpus	Aplica sustituciones sobre un corpus

Descripción

Aplica una serie de sustituciones sobre el contenido de un corpus a partir de un fichero de sustituciones.

Utilización

Preparar_corpus(corpus, ficheroSustitucion)

Argumentos

corpus	objeto de la clase Corpus o VCorpus.
ficheroSustitucion	ruta completa al fichero .txt donde se encuentran las sustituciones. Fichero formado por dos columnas separadas por tabulador, la primera con el término a sustituir, y la segunda con la sustitución. Si el segundo está vacío el término se elimina del corpus.

Salida

Objeto de la clase VCorpus o Corpus.

Ejemplos

```
corpusQuijoteTXT <- Crear_corpus("D:/Quijote.txt")
corpusPrep <- Preparar_corpus(corpusQuijoteTXT,
    "D:/sustituciones.txt")
```

Analizar_corpus	Analiza morfológicamente un corpus

Descripción

Analiza morfológicamente las palabras de cada entrada de un corpus con la ayuda de FreeLing.

Utilización

Analizar_corpus(corpus, directorio = getwd(), lenguaje = "es")

Argumentos

corpus	objeto de la clase Corpus o VCorpus.
directorio	ruta al directorio donde se encuentra la herramienta FreeLing. Por defecto getwd().
lenguaje	cadena de caracteres representativa del lenguaje del análisis morfológico permitido por FreeLing. Por defecto "es" (castellano).

Salida

Objeto de la clase VCorpus o Corpus.

Ejemplos

```
corpusQuijoteTXT <- Crear_corpus("D:/Quijote.txt")
corpusPrep <- Preparar_corpus(corpusQuijoteTXT,
    "D:/sustituciones.txt")
corpusAnalizado <- Analizar_corpus(corpusPrep,
    directorio="C:/Documents and Settings/", lenguaje="es")
```

Transformar_corpus — Separa el contenido de un corpus en frases

---

Descripción

Separa el contenido de un corpus para obtener uno nuevo en el que cada una de sus entradas represente una sola frase.

Utilización

Transformar_corpus(corpus)

Argumentos

corpus	objeto de la clase Corpus o VCorpus.

Salida

Objeto de la clase VCorpus o Corpus.

Ejemplos

```
corpusQuijoteTXT <- Crear_corpus("D:/Quijote.txt")
corpusPrep <- Preparar_corpus(corpusQuijoteTXT,
    "D:/sustituciones.txt")
corpusAnalizado <- Analizar_corpus(corpusPrep,
    directorio="C:/Documents and Settings/", lenguaje="es")
corpusTransformado <- Transformar_corpus(corpusAnalizado)
```

Detectar	Detecta las tripletas contenidas en un corpus

Descripción

Detecta las tripletas (estructuras Sujeto + Verbo + Objeto directo) contenidas en cada una de las entradas (frases) de un corpus con el análisis morfológico.

Utilización

Detectar_corpus(corpus)

Argumentos

corpus	objeto de la clase Corpus o VCorpus.

Salida

Objeto de la clase VCorpus o Corpus.

Ejemplos

```
corpusQuijoteTXT <- Crear_corpus("D:/Quijote.txt")
corpusPrep <- Preparar_corpus(corpusQuijoteTXT,
    "D:/sustituciones.txt")
corpusAnalizado <- Analizar_corpus(corpusPrep,
    directorio="C:/Documents and Settings/", lenguaje="es")
corpusTransformado <- Transformar_corpus(corpusAnalizado)
```

```
corpusDetectado <- Detectar(corpusTransformado)
```

Crear_diccionario	Crea un diccionario de términos a partir de un fichero

Descripción

Crea un diccionario de términos sin repetir a partir de un fichero que contiene textos sobre cierto tema.

Utilización

Crear_diccionario(tema, ruta_fichero, ruta_sustituciones, directorio_freeling)

Argumentos

tema	string con el nombre del tema.
ruta_fichero	ruta al directorio donde se encuentra el fichero con los textos. Puede ser .csv, .txt o .xlsx.
ruta_sustituciones	ruta completa al fichero .txt donde se encuentran las sustituciones. Fichero formado por dos columnas separadas por tabulador, la primera con el término a sustituir, y la segunda con la sustitución. Si el segundo está vacío el término se elimina del corpus.
directorio_freeling	ruta al directorio donde se encuentra la herramienta FreeLing.

Salida

Fichero Excel con el nombre del tema introducido. Contiene un diccionario de términos con pesos de relevancia (1 por defecto).

Ejemplos

```
Crear_diccionario(tema ="conflicto",
    ruta_fichero="D:/Código/data/conflicto.txt",
    ruta_sustituciones="D:/Código/data/sustituciones.txt",
    directorio_freeling= directorio="C:/Documents and Settings/")
```

Clasificar	Clasifica un corpus según un tema

Descripción

Clasifica las entradas de un corpus con tripletas detectadas para un tema dado a partir de un diccionario de términos. Puntúa cada entrada del corpus según su relevancia en dicho ámbito.

Utilización

Clasificar(corpus, tema)

Argumentos

corpus	objeto de la clase Corpus o VCorpus.
tema	string con el nombre del tema.

Salida

Objeto de la clase VCorpus o Corpus.

Ejemplos

```
corpusQuijoteTXT <- Crear_corpus("D:/Quijote.txt")
corpusPrep <- Preparar_corpus(corpusQuijoteTXT,
    "D:/sustituciones.txt")
corpusAnalizado <- Analizar_corpus(corpusPrep,
    directorio="C:/Documents and Settings/", lenguaje="es")
corpusTransformado <- Transformar_corpus(corpusAnalizado)
corpusDetectado <- Detectar(corpusTransformado)
Crear_diccionario(tema ="conflicto",
    ruta_fichero="D:/Código/data/conflicto.txt",
    ruta_sustituciones="D:/Código/data/sustituciones.txt",
    directorio_freeling=directorio="C:/Documents and Settings/")
corpusClasificado <- Clasificar(corpusDetectado,
    tema = " conflicto ")
Crear_diccionario(tema ="conflicto",
    ruta_fichero="D:/Código/data/conflicto.txt",
    ruta_sustituciones="D:/Código/data/sustituciones.txt",
    directorio_freeling=directorio="C:/Documents and Settings/")
```

```
corpusClasificado <- Clasificar(corpusDetectado,
    tema="conflicto")
```

Exportar — Exporta el contenido de un corpus a un Excel

Descripción

Exporta el contenido de un corpus detectado y clasificado a un fichero Excel.

Utilización

Exportar(corpus, directorio)

Argumentos

corpus	objeto de la clase Corpus o VCorpus.
directorio	ruta del directorio donde se escribirá el fichero de salida.

Salida

Fichero 'salida.xlsx' cuyo contenido son las frases originales, las tripletas y los pesos de clasificación.

Ejemplos

```
corpusQuijoteTXT <- Crear_corpus("D:/Quijote.txt")
corpusPrep <- Preparar_corpus(corpusQuijoteTXT,
    "D:/sustituciones.txt")
corpusAnalizado <- Analizar_corpus(corpusPrep,
    directorio="C:/Documents and Settings/", lenguaje="es")
corpusTransformado <- Transformar_corpus(corpusAnalizado)
corpusDetectado <- Detectar(corpusTransformado)
Crear_diccionario(tema="deportes",
    ruta_fichero="D:/Código/data/deportes.txt",
    ruta_sustituciones="D:/Código/data/sustituciones.txt",
    directorio_freeling= directorio="C:/Documents and Settings/")
corpusClasificado <- Clasificar(corpusDetectado, tema="deportes")
Crear_diccionario(tema ="conflicto",
```

```
        ruta_fichero="D:/Código/data/conflicto.txt",
        ruta_sustituciones="D:/Código/data/sustituciones.txt",
        directorio_freeling= directorio="C:/Documents and Settings/")
corpusClasificado <- Clasificar(corpusDetectado, tema="conflicto")
Exportar(corpusClasificado, "D:/Datos/")
```