Informe preliminar parte 1. Proyecto.

SISTEMAS DE APOYO A LA DECISIÓN

JAVIER GIL, GUZMÁN LÓPEZ, JAVIER SUÁREZ, OLIVER YE

Tabla de contenidos

1. Introducción a la minería de textos

La minería de datos la podemos definir como el análisis matemático para deducir patrones y tendencias que existen en los datos, patrones que no pueden detectarse mediante una exploración tradicional de los datos porque las relaciones son demasiado complejas o por el volumen de datos que se maneja.

Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos. Se puede decir que *Data Mining* se refiere al conjunto de métodos estadísticos que proporcionan información (correlaciones o patrones) cuando se dispone de muchos datos.[1]

En el campo de la inteligencia artificial, la minería de textos se aplica en el Procesado del Lenguaje Natural. Pero esta contiene multitud de subdivisiones: Comprensión del Lenguaje Natural, Generación de Lenguaje Natural, construcción de Conocimiento Base, Sistemas de Gestión del Diálogo, Procesamiento del Habla…etc.

Estas disciplinas permiten, por ejemplo, que un sistema no humano aprenda un lenguaje de forma automatizada, utilizarlo de forma relativamente coherente o adquiera ciertos conocimientos derivados de estos textos.[2]

Aplicaciones[2]:

-*Social Media Analytics* (SMA) consiste en la recogida de datos de las redes sociales (Twitter, Facebook, blogs, RSS, páginas web) distribuidos en diversas fuentes, tras realizar un análisis automatizado de la información, estos son mostrados en un entorno gráfico para ayudar a los empresarios a tomar decisiones empresariales.

-IBM *Watson* es una plataforma tecnológica que utiliza el Procesado de Lenguaje Natural para obtener patrones y resúmenes de grandes cantidades de información no estructuradas. *Watson* es capaz de extraer información de todo tipo de documentos de texto enriquecido de un repositorio para entonces construir patrones sobre las relaciones de los distintos conceptos de los documentos.

Retos[3]:

Uno de los retos que debe afrontar la minería de textos frente a la minería de datos tiene que ver con el análisis semántico de textos, ya que este análisis resulta computacionalmente muy costoso, pues los sistemas únicamente son capaces de operar en el orden de unas pocas palabras por segundo.

Otro retro es el análisis de textos multilenguaje. Los sistemas actuales están, en su mayoría, centrados en el inglés. Sería interesante ver cómo podrían tratar estos sistemas otros idiomas o incluso textos compuestos por varios idiomas.

2. Pre-procesamiento

Características más relevantes de los distintos tipos de datos:

-Raw. Estos datos no están en ningún formato concreto, es la representación de los datos tal cual son obtenidos en su fuente.

-BoW (*Bag of Words*). Este formato es una representación del texto que describe la ocurrencia de las palabras en un documento o directorio. Utiliza un vocabulario de palabras conocidas y una medición de la presencia de las palabras conocidas.

-TF-IDF (*Term Frequency-Inverse Document Frequency*). Es una estadística numérica que refleja la importancia de unas ciertas palabras en el contexto de un documento o directorio.

//TODO ejemplos de transformación

En cuanto a la fase práctica de esta tarea, hemos utilizado la plataforma GitHub en conjunción con Eclipse para el desarrollo del software. Hemos consultado toda la información relevante sobre los métodos y clases de Weka y Java en la Wiki oficial de Weka (https://weka.wikispaces.com/Use+WEKA+in+your+Java+code).

La repartición de las tareas está distribuida de la siguiente manera:

Tarea	**Autor**
Parte 1: Representación raw. Implementación y pruebas.	Guzmán López
Parte 2: Transformación (BoW, TF-IDF). Implementación y pruebas.	Javier Gil
Parte 3: Hacer compatible. Implementación y pruebas.	Oliver Ye
Informe	Javier Suárez

3. Bibliografía

[1] Revista Digital INESEM

¿Qué es y cuáles son las aplicaciones del Text Mining?

https://revistadigital.inesem.es/informatica-y-tics/text-mining/ (consultado el 27/03/18)

[2] UniRioja

Text Analytics: the convergence of Big Data and Artificial Intelligence (Documento PDF)

https://dialnet.unirioja.es/descarga/articulo/5573981.pdf (consultado el 27/03/18)

[3] IEEE

Text Mining: Challenges and Future Directions (Documento PDF)

http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7124872 (consultado el 27/03/18)