WICC 2021 Chilecito La Rioja
UNDeC
UNJu Universidad Nacional de Jujuy
VRAIn
Visualización y Recuperación Avanzada de Información

Web Mining y Text Mining: enfoques avanzados para analizar el contenido de grandes cantidades de información

José Federico Medrano; Valeria Barriento

jfmedrano@fi.unju.edu.ar ; barriento.valeria@gmail.com

Contexto

La línea de investigación aquí presentada se encuadra dentro del Proyecto Bianual 2020-2021 D/B035 denominado "**Agentes Inteligentes para Recuperación de Información y Analítica Visual en Big Data**", aprobado y financiado por la Secretaría de Ciencia y Técnica y Estudios Regionales de la Universidad Nacional de Jujuy (SeCTER – UNJu).
Este proyecto es llevado a cabo por el grupo de investigación Visualización y Recuperación Avanzada de Información (VRAIn) de la Facultad de Ingeniería de la UNJu.

Introducción

El Text Mining o Minería de Texto, un tipo particular de minería de datos, tiene como objetivo extraer conocimientos útiles como relaciones, patrones y tendencias de datos no estructurados o semiestructurados. El proceso principal en la minería de textos es transformar el texto en datos numéricos utilizando métodos estadísticos. La minería web es en realidad un área de minería de datos relacionada con la información disponible en internet. Es un concepto de extracción de datos informativos disponibles en páginas web.
Uno de los objetivos de este trabajo es la descripción avanzada o mejorada del contenido de grandes sitios web. Por ejemplo, para portales de noticias como La Nación , resultados de búsquedas como "violencia de género", "Alberto Fernández", "dólar" o "vacuna covid", por citar algunos ejemplos, ofrecen miles, decenas de miles y en algunos casos centenas de miles de resultados. Para poder procesar esta enorme cantidad de información y caracterizar de manera objetiva el contenido textual de las mismas, es necesario recurrir a las técnicas avanzadas que se mencionaron.
Así mismo, este trabajo se plantea caracterizar la demanda de empleo y la oferta de inmuebles a partir de la recolección de avisos clasificados. De este modo se podrán comparar o establecer el precio a un inmueble de acuerdo a las características propias y del conjunto de viviendas circundantes, o se podrá conocer, identificar y monitorear los requerimientos para un perfil de empleo determinado

Líneas de Investigación y Desarrollo

La presente investigación se enfocará en dos aspectos claves:
• Extracción automática de grandes cantidades de información de sitios web.
• Empleo de técnicas de PLN para encontrar información relevante.
Debido al dinamismo de la web, el contenido cambia, se actualiza, se agrega o elimina constantemente, por ello para analizar el contenido de un sitio web sería necesario contar con una instantánea completa que permita tener un vistazo de un momento determinado. Es aquí donde cobra importancia el web scraping, una técnica de Recuperación de Información empleada para extraer información de sitios web. De este modo, una vez extraídos los datos relevantes, se conforma un dataset para ser procesado y analizado para hallar patrones o tendencias que permitan relacionar distintos conjuntos temáticos.

Formación de Recursos Humanos

El Equipo de Trabajo está conformado por docentes investigadores y estudiantes de la Universidad Nacional de Jujuy. Los mismos llevan adelante esta línea de investigación desde hace años. Cada año se incorporan al proyecto alumnos avanzados de distintas carreras, quienes trabajan en temas relacionados con las temáticas planteadas. Del mismo modo, los integrantes del equipo participan en el dictado de asignaturas/cursos de grado y postgrado de la UNLP, UNJu y UCSEDASS.

Resultados esperados

Se espera en una primera instancia construir un *crawler* específico para un portal para poder extraer y recolectar toda la información necesaria. Puesto que los sitios web objeto de estudio no disponen de mecanismos que permitan recolectar la información por medio de una API, esta tarea se llevará a cabo mediante *web scraping*.
Por otro lado, se espera caracterizar la información recolectada, empleando un modelado de temas y visualizando los datos con librerías especializadas en la materia.