Web Mining

Trabajo Práctico Final

Maestría en Explotación de Datos y Gestión del Conocimiento

Layla Scheli, Franco Lianza, Lucio Scalzo, Ignacio Mujica, Alexis Walker

30 marzo 2022

Índice

Introducción

Clutch es un sitio B2B[1] de calificaciones y reseñas de empresas tecnológicas. Aquí, éstas son evaluadas en función de una serie de factores cuantitativos y cualitativos entre las que se encuentran reseñas certificadas que realizan sus cliente. Permite constituir una reputación a través de revisiones verificadas de terceros.

No solo es relevante como punto de consulta de información valiosa para la ponderación de proveedores; sino también, como se pretende en el presente, hacer un análisis de la presencia de la competencia en el mercado para establecer puntos de comparación.

Objetivo

Con este desarrollo se aspira a extraer referencias allí expuesta para constituir una radiografía del mercado. La intensión es simplificar la información en un cuadrante de posicionamiento que permita una rápida comparativa de las empresas que operan en un mismo área (delimitada por rubro y geografía).

La principales técnicas a utilizar son: Web Scraping para recopilar información de forma automática expuesta en Internet, y Text Mining para preprocesar los datos no estructurados obtenidos del portal citado.

Alcance

El entregable que resulta de este trabajo supone:

- Una pieza de código automatizable que extraiga el texto tal como se expone en el sitio web.
- Preprocesamiento de los registros utilizando técnicas de Text Mining para obtener un set de datos estructurados.
- Reconocer las entidades relevantes de cada lectura para disponer el resultado en forma adecuada para la construcción el resumen pretendido.

Consideraciones (a completar)

describir los aspectos a tener en cuenta cuando se hace scraping: CAPTCHA, contenido dinámico, etc.

Resto de tareas (a completar)

Cron en hosting Monitoreo: agente externo para ver si anda -> si anda, me autochequeo Resolver la caida de internet -> por ejemplo, reintentando Resolver errores html (40, 50) Chequear que la estructura cambio o no Describir cómo construir el dashboard sin construir nada ...

Futuras lineas de trabajo (a completar)

Extender el origen de información a otros sitios que brinde un servicio de idénticas características.

[1] Negocio a negocio (del inglés business-to-business o B2B)