Título

Preprocesamiento: “Big Data - Investiga”

Expositor

Lic. Pedro Asis.

Abstract

Justificación:

Vistas las necesidades de 1º) Automatizar la incorporación de los datos provistos por las compañías telefónicas en distintos formatos. 2º) Normalizar los datos de manera automática 3ª) Mantener la “trazabilidad” del proceso para dejar claro el origen de los resultados 4º) Poder proveer a “Investiga” un prefiltro para lograr que investigaciones con gran cantidad de datos incorporen sólo la información relevante, de este modo optimizar el uso de las capacidades gráficas del producto. 5º) Expandir las capacidades de “Investiga” permitiendo de esta manera poder ser utilizado para analizar otro tipo de delito que necesitan mucha información.

Se desarrolló “Big Data Investiga”, utilizando las tecnologías existentes más actualizadas al momento del desarrollo.

Descripción del Proceso:

El sistema importa los archivos enviados por la compañía telefónica para procesar automáticamente los datos.

El usuario debe subir los archivos desde su equipo a través de una interfaz. Cuando el archivo se encuentra en el servidor se comienza con el proceso de análisis del mismo para recuperar los datos contenidos..

El proceso de análisis del resumen telefónico consta de varios pasos:

- Obtención del texto del documento.
- Análisis y reconocimiento de la información.
- Extracción de textos para convertirlos en datos.
- Validación del dato obtenido.
- Normalización del dato.
- Guardar la información obtenida y analizada en la base de datos.

Una vez que el proceso de importación fué realizado, el sistema cuenta con todo lo necesario para comenzar con el análisis de la información.

La base de datos cuenta con la información obtenida del documento subido por el usuario, pero es necesario aclarar que esta información está asociada al documento, por lo cual, desde un registro de llamada específico, se puede acceder al documento que dió origen a esa llamada y descargarlo.

También es importante aclarar que se guarda el dato tal cual se leyó desde el documento, como también el mismo dato normalizado.

Todo el procesamiento se realiza sobre los datos normalizados, para evitar repeticiones y agrupar de una manera consistente la información.

Cabe aclarar que la normalización de datos consiste en dar un formato unificado a los valores obtenidos, así, por ejemplo, la fecha *1/8/18* y la fecha *1 de Agosto de 2018* de manera normalizada quedaría como *01/08/2018*

Esto también es aplicable a los números telefónicos, ya que los mismos pueden escribirse de diferentes maneras.

El análisis datos, de la información contenida en la base de datos se realiza con tecnologías de Big Data. Esto se traduce en el uso de tecnologías especialmente diseñadas para el procesamiento de gran cantidad de información de manera muy sencilla.

La información en cuestión se basa en el cruce de una gran cantidad de llamadas telefónicas, las cuales pueden ser de diferentes líneas, y por un espacio de tiempo prolongado.

Estas tecnologías permiten asociaciones y agrupamientos de manera casi automática, sin importar la cantidad de información sobre la que se esté trabajando. Esto mejora la experiencia del usuario ya que no debe esperar para que los filtros o agrupaciones se realicen.

Otra de las ventajas es que se obtiene la cantidad de ocurrencias de un valor en un conjunto de datos, de esta manera es muy fácil realizar filtros que incluyan los valores más representativos o un filtro que excluya valores únicos.

Esto también es aplicable a rangos temporales, o sea, los valores que contienen fechas y horas.

Resultados:

Podemos decir que el producto obtenido cumple con lo planteado en la “Justificación”, de todos modos hemos observado la posibilidad de poder mejorar brindando una interfaz más intuitiva a los usuarios finales de el producto como así lograr una mejor y más transparente integración con “Investiga” que es el objetivo final de todo este desarrollo.

Lic. Pedro S. Asís.
Director de Departamento de
Ingeniería en Sistemas de Información.
UTN - Delta