Workshop de Investigadores de Ciencias de la Computación

USO DEL ANÁLISIS DE TEXTO ESTRUCTURADO Y NO ESTRUCTURADO PARA LA GENERACIÓN DE MATERIAL DE ENTRENAMIENTO PARA MODELOS DE IA

Nicolás Hernández, Julio Castillo, Marina Cardenas, Sandra Olariaga, Matias Negrelli, Sofia Britos, Tito Santino

Laboratorio de Investigación de Software/Dpto. Ingeniería en Sistemas de Información/Facultad Regional Córdoba/Universidad Tecnológica Nacional

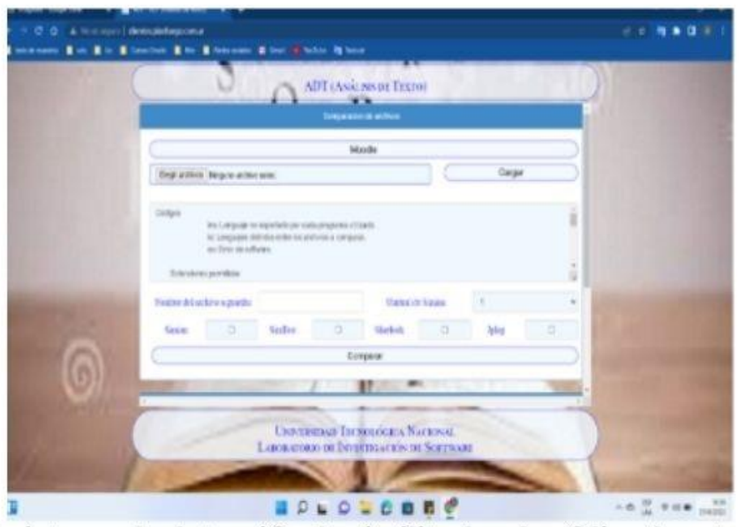
{damiannicolas05, dr.jotacastillo, ing.marinacardenas, sandramolariaga, matutecba19, soffibritos01, zsantinoz}@gmail.com

Objetivos

Crear herramientas que pemitan analizar grandes volumenes de datos que se encuentran en forma de archivos de textos estructurados o no estructurados, adecuando la información para su utilización en el entrenamiento de sistemas de análisis de texto.

Contexto

- Proyecto denominado "Modelo para el procesamiento de textos estructurados Fase 2" (cód. SIECACO0008518), que es un proyecto homologado por la SCyT de la UTN.
- Actualmente, el proyecto se encuentra dentro del grupo de investigación denominado Grupo de Aprendizaje Automático, Lenguajes y Autómatas (GA2LA).
- Físicamente, los integrantes del proyecto desarrollan sus actividades en el Laboratorio de Investigación de Software LIS del Dpto. de Ingeniería en Sistemas de Información.



Subsistema de detección de similitudes de código fuente.

Formación de Recursos Humanos

- Una doctoranda en ingeniería con mención en sistemas de información en la UTN-FRC, que está trabajando específicamente en el subsistema de detección de similitudes.Ademas realiza la dirección de becarios.
- Un doctor en ciencias de la computación quien desarrolló su tesis en el área de investigación, que realiza la dirección de becarios.
- Un maestrando en Ingeniería en Sistemas de Información de la UTN-FRC, que está desarrollando su tesis.
- Docentes Investigadores, alumnos y becarios.

Lineas de Investigación y Desarrollo

- Uso de redes neuronales, en aprendizaje supervisado, semi supervisado y machine learning.
- Creación y utilización de corpus, que es el estudio empírico de la lengua.
- Reconocimineto y comprención del lenguaje humano mediante la creación de modelos computacionales.

Resultados Obtenidos/Esperados

Como resultado se ha obtenido un conjunto de herramientas que han sido desarrolladas para el estudio e investigación sobre análisis de texto estructurado y el área de minería de datos. A continuación se enumeran las herramientas:

- Programa de Mapeo de Datos (PMD).
- Banco de Prueba de Algoritmos de Semejanza (BPAS).
- Subsistema de detección de similitudes en archivos de código fuente (SDS).

Pantalla principal del Programa de Mapeo de Datos

Agradecimientos

-Secretaría de Ciencia y Tecnología (SCyT) de la UTN.

-Dpto. de Ingenieria en Sistemas de Información.