Aplicación de Técnicas Descriptivas de Minería de Textos sobre Contenido Digital Realizando Análisis Inteligente

Sánchez Rivero, Víctor David; Farfán, José Humberto; Rodríguez, Mariela Ester; Vargas, Luis Alejandro; Vega, Ariel Alejandro; Garcete, Christian Brian; Llampa, Álvaro Facundo; Ramos, Pablo Nicolás; Contreras, Facundo; Churquina, Cintia Noelia; Águila, Viviana Emilia y Iogna Prat Genzel, Nicolás

Área de Ingeniería Informática - Facultad De Ingeniería – Universidad Nacional De Jujuy

ivansrivero@gmail.com, jhfarfan@hotmail.com, maru972@gmail.com, alevar98@yahoo.com, arielalejandrovega@gmail.com, chrisbriancerrudo@gmail.com, a.facundollampa@gmail.com, pablonicolasr777@gmail.com, facucontreras21@gmail.com, cintia7828@gmail.com, vivi.agui.4@gmail.com, iognapratnicolas@gmail.com

RESUMEN

El presente proyecto pretende implementar técnicas de Minería de Textos o Text Mining en conjunción con técnicas de Minería Web o Web Mining (metodología para la recuperación y extracción de información desde páginas web) para poder realizar un estudio de los Patrones de Escritura empleados para la confección de documentos digitales científicos. Tanto Text Mining como Web Mining se encuadran dentro de las técnicas de Minería de Datos y son técnicas que permiten descubrir patrones usados en grandes volúmenes de texto. El proyecto también incluirá una investigación sobre la aplicación de técnicas o algoritmos orientados al Procesamiento del Lenguaje Natural usados en el análisis de textos o documentos obtenidos de Redes Sociales, por ejemplo, y se persigue, a través de su empleo, la obtención de prototipos de sistema que faciliten el análisis en cuestión.

Los textos o documentos digitales, sobre los cuales se trabajará en este proyecto, se obtendrán principalmente desde la Web, considerando que en la Sociedad del Conocimiento actual, la gestión de la información y conocimiento es un componente estratégico para el análisis inteligente de la información digital, para la clasificación de contenidos y la extracción de conceptos, entre algunos de los principales tópicos que estudia Text Mining.

Palabras clave: Minería de Textos, Minería de Datos, KDD, Análisis Inteligente.

CONTEXTO

El proyecto se encuentra inserto dentro de las siguientes Líneas Prioritarias de Investigación de la Facultad de Ingeniería (LIPIFI) - UNJu:

- Ingeniería del Software.
- Ingeniería de Procesos.

Es un proyecto aprobado de categoría B (Código D/B036) denominado:

"Aplicación de técnicas descriptivas de Minería de textos sobre contenido digital realizando un análisis inteligente".

Financiamiento: Secretaría de Ciencia y Técnica y Estudios Regionales (SeCTER) de la UNJu.

Vigencia del Proyecto: 01/01/2020 al 31/12/2021

1.INTRODUCCION

El actual proyecto pretende implementar técnicas de Minería de Textos o Text Mining (el cual es el proceso de derivar información nueva a partir de textos digitales) en conjunción con técnicas de Minería Web o Web Mining (metodología para la recuperación y extracción de información desde páginas web) para poder realizar un estudio de los Patrones de Escritura empleados para la confección de documentos digitales científicos. Tanto Text Mining como Web Mining se encuadran dentro de las técnicas de Minería de Datos y son técnicas que permiten descubrir patrones usados en grandes volúmenes de texto. El proyecto también incluirá una investigación sobre la aplicación de técnicas o algoritmos orientados al Procesamiento del Lenguaje Natural usados en el análisis de textos o documentos obtenidos de Redes Sociales, por ejemplo y se persigue, a través de su empleo, la obtención de prototipos de sistema que faciliten el análisis en cuestión.

Los textos o documentos digitales, sobre los cuales se trabajará en este proyecto, se obtendrán principalmente desde la Web, considerando que, en la Sociedad del Conocimiento actual, la gestión de la información y conocimiento es un componente estratégico para el análisis inteligente de la información digital, para la clasificación de contenidos y la extracción de conceptos, entre algunos de los principales tópicos que estudia Text Mining.

La Minería de Datos es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de datos [1]. El Grupo de Investigación se inició en el año 2016 con un proyecto de categoría B denominado “Data Mining aplicado a análisis telefónico” (Código SeCTER D/B026). Luego, en 2018, se trabajó en otro proyecto (categoría B Código SeCTER D/B030) denominado “Implementación de técnicas específicas de Minería de datos en aplicaciones web con motores de Base de Datos Relacionales”. Algunos integrantes del Grupo de Investigación también participan de la beca EVC-CIN 2018 como Estímulo a la Vocación Científica, con los trabajos denominados “Detección de Plagio Implícito en Tesis de Grados” (Código 9411) y “Análisis Inteligente de Técnicas de Minería de Texto e Implementación para la Detección de Ciberacoso en Redes Sociales” (Código 12100).

Las temáticas Minería de Datos, Minería de Textos y Minería Web se desarrollan en la materia Aplicaciones de Base de Datos II; asignatura en la que algunos integrantes son docentes de la misma; de tal manera que los resultados de los proyectos de investigación y becas, antes mencionados, se vuelcan en dicha asignatura.

La minería de textos es el proceso de derivar información de alta calidad del texto que podría obtenerse de páginas web, por ejemplo. La información de alta calidad se alcanza, generalmente, a través de la elaboración de patrones y tendencias de medios tales como el

aprendizaje estadístico de patrones. La minería de texto generalmente implica el proceso de estructurar el texto de entrada (típicamente el análisis sintáctico, junto con la adición de algunas características lingüísticas derivadas y la eliminación de otras, y la posterior inserción en una base de datos), derivando patrones dentro de los datos estructurados y, finalmente, la evaluación e interpretación del resultado [2].

Al ser una temática que involucra técnicas de gran auge, desarrollo y relevancia en los últimos años [3], permite, a los alumnos de carreras informáticas de la Facultad de Ingeniería de la UNJu, plantear y desarrollar análisis inteligente sobre textos o documentación obtenida de la web, que se traducirán en propuestas de Proyectos Finales de Carrera.

Se destaca que los seres humanos son hábiles a la hora de producir e interpretar el lenguaje cotidiano porque son capaces de expresar, percibir e interpretar significados complejos en fracción de segundos; sin embargo, no son capaces de describir y comprender eficientemente las reglas que gobiernan el lenguaje natural. Este es el principal motivo por el cual entender y producir el lenguaje por medio de una computadora es un problema difícil de resolver. Este problema, pertenece al campo de estudio de Inteligencia Artificial llamado Procesamiento del Lenguaje Natural [4].

Por las características mencionadas, en el párrafo anterior, resulta razonable experimentar la aplicación del Deep Learning dentro de las posibles técnicas que se desarrollarán en el Text Mining y analizar los resultados obtenidos.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Las líneas de investigación y desarrollo abarcan el estudio de las siguientes temáticas:

- Técnicas descriptivas de Minería de Textos.
- IDE´s o herramientas informáticas de Minería de Datos que empleen las técnicas mencionadas anteriormente.
- Obtención de contenido digital que permitan aplicar dichas técnicas, ya sea con técnicas de Minería Web u otros procesos similares.
- Análisis Inteligente sobre los resultados obtenidos.

3. RESULTADOS OBTENIDOS/ESPERADOS

Se espera a través de los resultados alcanzados y los objetivos en curso que los mismos se vuelquen a:

- Tesis de grado de alumnos de las carreras Ingeniería en Informática y Licenciatura en Sistemas de la Facultad de ingeniería de la UNJu, cuyas temáticas se enfoquen a la Minería de Textos, Minería de Datos y Análisis Inteligente.
- Alumnos que desarrollen estas temáticas aplicándolas en becas científicas, como las becas EVC-CIN de Estímulo a la Vocación Científica.
- Favorezcan la aplicación en las prácticas de los alumnos que cursen la cátedra Aplicaciones de Base de Datos II, perteneciente a la carrera Licenciatura en Sistemas.

Los conocimientos adquiridos se pondrán a disposición de cualquier

persona, institución u organización que desee implementar técnicas de herramientas de Minería de Textos para realizar Análisis Inteligente sobre datos existentes o de su propiedad mediante la generación de repositorios de información específicos.

Se destaca la vinculación existente entre la Facultad de Ingeniería y el Gobierno Provincial en relación a proyectos anteriores del Grupo de Investigación, por tal motivo es intención mantener y afianzar esta relación de trabajo e implementar técnicas que faciliten la toma de decisiones, en sus diferentes áreas.

4. FORMACIÓN DE RECURSOS HUMANOS

Cantidad de tesinas de grado en curso en el 2.019: 2 (cuatro integrantes).

Cantidad de tesinas de grado aprobadas en 2.019: 1 (dos integrantes).

El equipo que desarrollará la propuesta posee vasta experiencia en el trabajo docente con cursos de gran diversidad, en el área de tecnologías básicas y avanzadas, con promoción con distintas modalidades. Son sus miembros:

Director del Proyecto: Esp. Ing. Sánchez Rivero, Víctor David.
Investigadores Docentes:

- Esp. Ing. Inf. Farfán, José Humberto: director de beca EVC-CIN 2018 en curso.
- Ing. Rodríguez, Mariela Ester: CoDirectora de beca EVC-CIN 2018 en curso.
- Mg. Ing. Vega, Ariel Alejandro: director de beca EVC-CIN 2018 en curso.
- Ing. Vargas, Luis Alejandro.

Estudiantes Investigadores:

- Ramos, Pablo Nicolás: tesina de grado en curso, beca EVC-CIN 2018 de Estímulo a la Vocación Científica en curso.
- Garcete, Christian Brian.
- Churquina, Cintia Noelia.
- Águila, Viviana Emilia.
- Iogna Prat Genzel, Nicolas.

Egresados

- Llampa, Alvaro Facundo: tesina de grado aprobada, beca EVC-CIN 2018 de Estímulo a la Vocación Científica en curso.
- Contreras, Facundo: tesina de grado aprobada, beca EVC-CIN 2018 de Estímulo a la Vocación Científica en curso.

5. BIBLIOGRAFIA

[1] Maimon, O., & Rokac, L. (2010). ¨Data Mining and Knowledge Discovery Handbook", "O. Maimon, & L. Rokac, Data Mining and Knowledge Discovery Handbook", Nueva York, Springer, 2010, págs. 1-18.

[2] Ortiz A. (2018). "¿Qué es el análisis de texto, extracción de textos o minería de textos?". Obtenido de https://pcweb.info/que-es-analisis-de-texto-extraccion-mineria-de-textos en Junio de 2019.

[3] Ricardo & Barbosa. (2019). "Importancia de la minería de datos en el mundo empresarial actual". Obtenido de https://www.ricardo-barbosa.com/es/importancia-de-datos-mineria-en-hoy-negocios-mundo/ en Junio de 2019.

[4] López Briega R. (2019). IAAR, "Comunidad Argentina de Inteligencia

Artificial". Obtenido en https://iaarbook.github.io/procesamiento-del-lenguaje-natural/ en Junio 2019.