

Proyecto 1, Etapa 2: Analítica de textos

Federico Melo Barrero, 202021525, f.melo. Sección 1.

Shadith Pérez Rivera, 202014687, s.perezr. Sección 2.

El presente documento describe el proceso de despliegue de la solución analítica en el ambiente de producción de una organización.

Contenido

Proceso de automatización del proceso de preparación de datos	1
Construcción y persistencia del modelo	2
Acceso por medio de API	3
Desarrollo de la aplicación y justificación	4
Importancia que tiene para ese rol la existencia de esta aplicación.....	6
Mejoras gracias al grupo de estadística	7
Trabajo en equipo.....	8
Lista de Referencias.....	9

Proceso de automatización del proceso de preparación de datos

Para la preparación de los datos se recibe el texto y se le hace el debido preprocesamiento en el lado del back-end. Esto se puede ver en el archivo **src\services\classifier.py**. El procedimiento de preparación de los datos realiza lo siguiente:

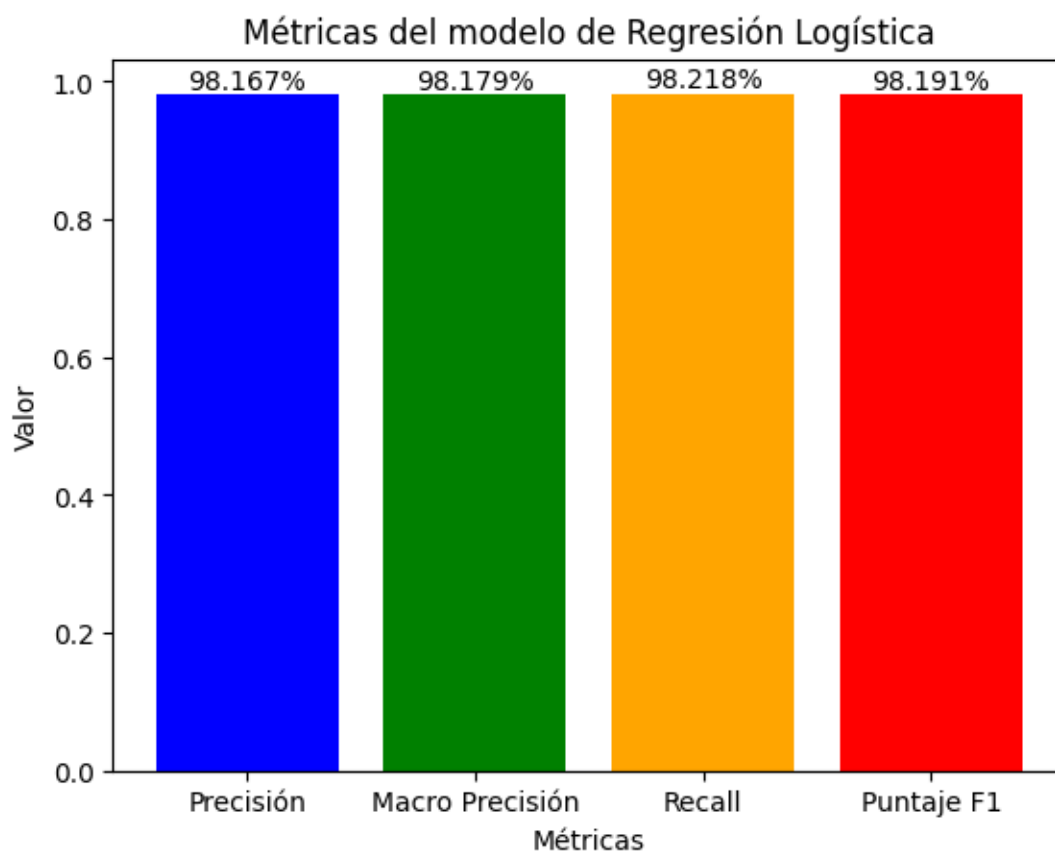
1. **Preprocesamiento (y remoción de las palabras de parada).** Remove los caracteres que no hacen parte de la codificación ASCII, remove todos los que no son alfabéticos, convertir todos los caracteres de mayúscula a minúscula, reemplazar los caracteres numéricos por su equivalente en palabras (“2” pasa a “dos”) y remove las palabras de parada (que en español suelen ser artículos, preposiciones, conjunciones, determinantes, entre otros).
2. **Tokenización.** Se realiza una tokenización con granularidad de palabras en la que distintas palabras se identifican por un mismo término, dividiendo el texto para aplicar los modelos. Se utiliza como separador el espacio.
3. **Normalización mediante stemming.** Se realiza la normalización de las palabras haciendo uso de stemming. Se utiliza SnowBall Stemming debido a que cuida no reducir al mismo término las palabras con significados

diferentes y evita realizar stemming en casos complicados (recordando que en cuestión de stemming es peor hacer un stemming excesivo que hacer menos). No siempre reduce los términos a una palabra existente, pero eso no representa una dificultad en este contexto de negocio.

Al igual que en la parte 1 del proyecto, **se hace uso de stemming y no de lematización**, en tanto que fuentes como [1] indican que es preferible para idiomas como español.

Construcción y persistencia del modelo

Para la construcción y persistencia del modelo utilizamos el modelo que arrojó las mejores métricas al realizar diferentes pruebas y realizar una validación cruzada. Este modelo fue el de regresión logística, que obtuvo los siguientes indicadores:



Estos indicadores son explicados en la parte 1 del proyecto y también en la página web. Se creó un *pipeline* para este modelo y fue entonces guardado como un **pkl**, almacenando así los pesos de la regresión y los bias, en el archivo **ods_classifier.pkl**.

En las funciones **classify_multiple_texts** y **classify_single_text** del archivo `src\services\classifier.py` se puede evidenciar el uso del *pipeline* para clasificar los textos dados por un usuario.

Acceso por medio de API

Para poder acceder al modelo creamos un back-end en FastAPI utilizando Python.

Se puede acceder a la documentación del API, generada automáticamente por Swagger UI, mediante <http://localhost:8000/>, una vez se esté ejecutando el back-end de acuerdo con las instrucciones dadas en el [repositorio de github](#).

El backend almacena información en una base de datos relacional, en este caso usando SQLAlchemy para conectar FastAPI con SQLite.

Los endpoints expuestos entran principalmente en 3 categorías.

1. **Autenticación:** Esta categoría está compuesta de 2 endpoints fundamentales. Uno permite crear un nuevo usuario, la ruta para poder acceder a este endpoint es: `localhost:8080/logins/creáre` y debe recibir 2 parámetros, un login y una contraseña, para crear un usuario nuevo de la aplicación.

POST `/logins/create` Create Login

El otro endpoint importante es un POST a `localhost:8080/logins/`, que revisa si un *username* y *password* están en la base de datos y con eso permite o no iniciar sesión en el fornt.

POST `/logins/` Check Login

2. **Clasificador y CRUD para apartados de texto:** Se realizó una serie de endpoints que permiten obtener, modificar, crear y eliminar apartados de texto. Una vez se crea un apartado de texto, el API:
 - a. Revisa si ya existe en la base de datos y por ende ya se tiene una clasificación para él. Si ese es el caso, no vuelve a computar la clasificación, sino retorna la ya existente.
 - b. Si no existe en la base de datos, utiliza el modelo antes descrito para asignarle una clasificación y retornarla. Además, almacena el proceso de preprocesamiento, para que lo pueda ver el usuario final si lo desea.

Hay dos opciones principales para clasificar apartados de texto:

- Dando a la aplicación un texto, usando el endpoint:

POST /`excerpts/` `Classify Excerpt`

- Mediante un archivo de Excel que debe tener **una sola columna** con múltiples textos a clasificar, usando el endpoint:

POST /`excerpts/excel` `Classify Excerpts From Excel`

3. **Revisión de resultados y del proceso:** Es viable obtener los textos que corresponden a una categoría dada, con el endpoint:

GET /`excerpts/ods/{category}` `Get Excerpt By Category`

También se pueden revisar qué datos fueron usados para el entrenamiento y las pruebas del modelo, respectivamente, con los endpoints

GET /`excerpts/test` `Get Test Excerpts`

GET /`excerpts/train` `Get Train Excerpts`

Es viable revisar el proceso de clasificación de un texto, e inclusive obtener la forma del texto tokenizada y preprocesada, haciendo uso del endpoint:

GET /`excerpts/log` `Get Log`

Nótese que solo se mencionan los endpoints más relevantes para el negocio, pero el backend cuenta con un total de 18 endpoints que pueden realizar cualquier tipo de acción sobre los usuarios de la aplicación y sobre los textos que se tienen almacenados o textos nuevos.

Desarrollo de la aplicación y justificación

La aplicación fue desarrollada principalmente con 3 capas, cada una desarrollada en tecnologías diferentes.

1. Base de datos (SQLite):

La base de datos utilizada en el proyecto fue SQLite y almacena los datos relevantes a los usuarios y a peticiones ya realizadas. El backend se conecta a la base de datos mediante SQLAlchemy. También almacena los datos que fueron usados para entrenamiento y prueba del modelo. De este modo, los usuarios y contraseñas de los usuarios son almacenados para permitir la autenticación de los usuarios. Por otro lado, la base de datos también

almacena los resultados del modelo pertinentes al hacer peticiones. Esto permite que el back-end no tenga que utilizar el modelo y predecir la categoría si un texto ya registrado es solicitado por el usuario.

2. Back end (FastAPI):

El back-end fue desarrollado en FastAPI con Python, este back end permite la comunicación con la base de datos y ofrece los endpoints necesarios, expuestos arriba, para llevar la operación principal de aplicación.

3. Front (ReactJS): El front-end fue llevado a cabo utilizando ReactJS y posee una interfaz intuitiva y confiable que le permite al usuario hacer múltiples acciones:



- Realizar el debido proceso de autenticación para cerciorarse de qué la información que se encuentra en la aplicación no puede hacer accedida por usuarios no autorizados
- Examinar el modelo realizado y conocer tanto sus métricas como los datos que fueron utilizados para su entrenamiento y para su proceso de prueba.
- Acceder al historial de todos los textos que han sido ingresados a la aplicación, ver la clasificación como ODS que se les ha otorgado e incluso ver cómo fueron preprocesados y tokenizados.
- Ingresar textos nuevos a la aplicación ya sea escritos manualmente o como archivo de Excel para que el modelo pueda clasificarlos en uno de los tres objetivos de desarrollo sostenible: 6, 7 o 16.

Rol de la organización que va a utilizar la aplicación y proceso de negocio que va a apoyar

Recuérdese que el objetivo de este proyecto es diseñar un modelo de aprendizaje automático que, dado un apartado de texto arbitrario, permita al UNFPA clasificarlo como relacionado al ODS 6, 7 o 16.

El rol de la organización que hará uso de este modelo es el Área de recolección de información de la UNFPA y el Equipo de comunicaciones de la UNFPA. Esto se evidencia en el siguiente apartado del mapa de actores entregado en la fase 1 del proyecto:



Rol	Tipo de Actor	Beneficio	Riesgo
Área de recolección de información de la UNFPA	Usuario / Cliente	Utiliza la clasificación automática para no tener que clasificar a qué ODS corresponde cada	Debe comprenderse que este modelo en particular solo clasifica en 3 ODS,

 <div>Universidad de los Andes Colombia</div> <div><div>Acreditación Institución de Alta Calidad 10 años</div><div>Reconocimiento Exención ISO 9001 9 de agosto de 21</div></div>		Ingeniería de Sistemas y Computación Pregrado ISIS-3301 – Inteligencia de Negocios Primer Proyecto Semestre: 2023-20		
			registro de forma manual. Con eso, asigna recursos de manera eficiente a cada área relacionada con los ODS y brinda la información apropiada al área que la necesita sin que haya una inversión grande de tiempo y capital humano.	por lo que información que no corresponda con ninguno puede ser mal clasificada. Para el resto de ODS debe invertirse capital humano o solicitarnos un nuevo modelo enfocado a eso.
Equipo de comunicaciones de la UNFPA	Beneficiado	A partir de la clasificación tiene acceso a información sobre cada tema para así poder difundir más fácilmente el impacto de los problemáticas y proyectos relacionados con los ODS 6, 7 y 16. Eso ayuda a hacerlos públicos y a una mayor difusión, lo cual a su vez brinda más atención y colaboración a la causa de cumplir los ODS.	Nuevamente es imperativo que el equipo reconozca que el modelo no es la verdad absoluta y solo clasifica en 3 ODS, por lo que información que no corresponda con ninguno puede estar mal clasificada.	

Con eso en mente, el proceso de negocio que va a ser apoyado por la aplicación es el proceso de distribución adecuada de información. Este proceso refiere al hecho de que una vez realizada la recolección de información y apartados de texto relacionados con los ODS, la UNFPA debe organizar esta información, clasificarla cada una en un ODS, y brindársela a las organizaciones que la requieran y la pueden aprovechar. Por ejemplo, la información relacionada con el ODS 6 puede ser dirigida a organizaciones que pueden colaborar con el acceso a agua potable y este tipo de organizaciones no deben recibir información relacionada con ningún otro ODS, pues será distractora e innecesaria.

Importancia que tiene para ese rol la existencia de esta aplicación

Así pues, la importancia que tiene la aplicación para los roles mencionados es muy grande, pues el hecho de que la aplicación clasifique automáticamente los textos en uno de los tres ODS ahorra muchísimo trabajo manual, disminuye la inversión de

 <p>Universidad de los Andes Colombia</p> <p>Acreditación institucional de alta calidad 10 años MinEducación Resolución 1812 9 de enero de 21</p>	<p>Ingeniería de Sistemas y Computación Pregrado</p> <p>ISIS-3301 – Inteligencia de Negocios Primer Proyecto Semestre: 2023-20</p>	 <p>Engineering Accreditation Commission</p>
---	--	---

capital humano, aumenta la eficiencia en el proceso de clasificación de los textos, y por ende aumenta también la eficiencia en el proceso de distribuir la información a las organizaciones a las que les corresponde, dándoles únicamente información que les es pertinente. La aplicación representa entonces un valor agregado muy grande para la UNFPA y cualquier otra organización afiliada a las Naciones Unidas que propenda por el cumplimiento de los ODS, pues permite ahorrar tiempo y dinero y acelerar el proceso de impulsar el cumplimiento de los ODS.

Mejoras gracias al grupo de estadística

La contribución del estadístico Andrés David Castañeda en nuestro proyecto ha sido fundamental para enriquecer nuestras capacidades analíticas y ofrecer una visión más detallada y precisa de los datos, lo que resultó en una comprensión más profunda de los patrones subyacentes en los textos y su relación con los Objetivos de Desarrollo Sostenible.

A continuación, se destacan las mejoras y sugerencias proporcionadas por el estadístico, que han fortalecido significativamente nuestro enfoque analítico:



1. Mejora de Gráficas: La visualización de datos desempeña un papel crucial en la comprensión de patrones y tendencias. En particular, nos recomendó incluir la gráfica de barras y la matriz de confusión del modelo en la aplicación web, lo que facilita al usuario entender la calidad del modelo.

2. Revisión y Mejora del Proceso de Preparación de Datos: El preprocesamiento de datos es una etapa crítica en la construcción de modelos precisos. Andrés nos ayudó a revisar y mejorar el proceso de preparación de datos para garantizar que los textos se transformen adecuadamente.

Si bien estas mejoras han enriquecido nuestro proyecto, es importante tener en cuenta que algunos aspectos pueden resultar desafiantes debido a las limitaciones existentes. No obstante, la colaboración con Andrés no solo ha mejorado la calidad y precisión de nuestra aplicación, sino que también ha enriquecido la visualización de los datos y los procesos analíticos. Sus aportes han sido invaluable en la búsqueda de una aplicación más sólida y efectiva para predecir los Objetivos de Desarrollo Sostenible.

Trabajo en equipo

Integrante	Participación y roles adoptados
Federico Melo Barrero	Roles: Líder de proyecto, Ingeniero de software responsable del diseño de la aplicación y resultados, Ingeniero de software responsable de desarrollar la aplicación final. Acumulado de número de horas dedicadas al proyecto: aproximadamente 20 horas de trabajo neto. Retos enfrentados en el proyecto y formas planteadas para resolverlos: Gestionar eficazmente el tiempo y la carga de trabajo, implementé una estructura de gestión del tiempo más efectiva y me rodeé de un entorno de trabajo inspirador. Se realizaron todas las reuniones propuestas en el enunciado y el proyecto final resultó ser de muy buena calidad. Puntos repartidos: 50/100 . Cosas por mejorar: Realizar una planeación más detallada de la arquitectura de la aplicación para no tener que realizar cambios sobre la marcha.
Shadith Pérez Rivera	Roles: Ingeniero de datos, Ingeniero de software responsable de desarrollar la aplicación final. Acumulado de número de horas dedicadas al proyecto: aproximadamente 20 horas de trabajo neto. Retos enfrentados en el proyecto y formas planteadas para resolverlos: La gestión del tiempo. Para abordar este reto, implementé una planificación detallada, asignando tareas específicas a intervalos regulares y estableciendo plazos realistas. Se realizaron todas las reuniones propuestas en el enunciado y el proyecto final resultó ser de muy buena calidad. Puntos repartidos: 50/100 . Cosas por mejorar: Realizar una planeación más detallada de la arquitectura de la aplicación para no tener que realizar cambios sobre la marcha.

 <p>Universidad de los Andes Colombia</p> <p>Acreditación institucional de alta calidad 10 años Ministerio de Educación Resolución 1823 9 de enero de 2018</p>	<p>Ingeniería de Sistemas y Computación Pregrado</p> <p>ISIS-3301 – Inteligencia de Negocios Primer Proyecto Semestre: 2023-20</p>	 <p>Engineering Accreditation Commission</p>
---	--	---

Lista de Referencias

- [1] ONU. (2023). *Objetivos de Desarrollo Sostenible* [En línea]. Disponible en: <https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/#>
- [2] ONU. (2023). *Objetivo 6: Garantizar la disponibilidad de agua y su gestión sostenible y el saneamiento para todos* [En línea]. Disponible en: <https://www.un.org/sustainabledevelopment/es/water-and-sanitation/>
- [3] ONU. (2023). *Objetivo 7: Garantizar el acceso a una energía asequible, segura, sostenible y moderna* [En línea]. Disponible en: <https://www.un.org/sustainabledevelopment/es/energy/>
<https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/#>
- [4] ONU. (2023). *Objetivo 16: Promover sociedades justas, pacíficas e inclusivas* [En línea]. Disponible en: <https://www.un.org/sustainabledevelopment/es/peace-justice/>
- [5] La República. (2023). *En Colombia, 3,2 millones de personas no tienen acceso al servicio de agua potable*. [En línea]. Disponible en: <https://www.larepublica.co/economia/en-el-colombia-3-2-millones-de-personas-no-tienen-acceso-al-servicio-de-agua-potable-3576736>
- [6] Statista. (2023). *Número de homicidios cometidos por cada 100.000 habitantes en Colombia de 2014 a 2022*. [En línea]. Disponible en: <https://es.statista.com/estadisticas/1289833/tasa-de-homicidios-colombia/#:~:text=En%202022%2C%20hubo%20aproximadamente%2026,por%20debajo%20de%20los%2025.>
- [7] E. Perdomo, J. Díaz, A. Ojeda y N. Amador. *Análisis de los procesos de lematización y estemizado en lingüística computacional*. Instituto de Literatura y Lingüística. Cuba. 2017. Disponible en: https://www.researchgate.net/profile/Josval-Diaz-Blanco/publication/322364515_ANALISIS_DE_LOS_PROCESOS_DE_LEMATIZACION_Y_ESTEMIZADO_EN_LINGUISTICA_COMPUTACIONAL/links/5a560e1445851547b1be8080/ANALISIS-DE-LOS-PROCESOS-DE-LEMATIZACION-Y-ESTEMIZADO-EN-LINGUEISTICA-COMPUTACIONAL.pdf