

IMPLEMENTACIÓN DE UN CHATBOT INTELIGENTE MEDIANTE EL USO DE MACHINE LEARNING Y PYTHON

Sandoval, Pablo

Universidad Católica de Salta, Facultad de Ingeniería

Abstract:

La presente investigación se propone, haciendo uso de las tecnologías provenientes del Aprendizaje Automático (Machine Learning en inglés) y métodos propios de uno de los campos de la inteligencia artificial llamado procesamiento de lenguaje natural, diseñar una entidad conversacional artificial. La misma consiste en el desarrollo de un sitio web con un formato de sala de chat que actúe como canal de distribución de información, asistiendo a aquellos interesados en participar en el Grupo de incubación de trabajos de investigación de alumnos conocido como IDEAS de la Facultad de Ingeniería de la Universidad Católica de Salta el cual tiene como objetivo incentivar a los alumnos interesados en la investigación, el desarrollo y la innovación (I+D+i) respondiendo con considerable precisión las diversas dudas e inquietudes que se le presente en relación al mismo. Esto facilitaría la incorporación de nuevos aspirantes al grupo proveyendo una atención sin interrupción, inmediata y útil.

Palabras Clave

Chatbot; Python; Machine Learning; Procesamiento de Lenguaje Natural; Stanza

Introducción:

Actualmente, la innovación en las organizaciones es un factor imprescindible, esta asegura la participación activa en el mercado a través del tiempo de las mismas como también su triunfo frente a otras.

En base a esto, las actualizaciones tecnológicas definen un cambio considerable que determinará el beneficio de las empresas y trazará un rumbo a una economía creciente [1]. Es claro entonces que existirá una fuerte demanda en cuanto investigadores como personal técnico y de apoyo en I+D tanto del sector público, privado y externo.

Ante esta necesidad de profesionales relacionados a este ámbito, se comenzó a fomentar durante la formación universitaria el interés de convertir ideas en potenciales proyectos con diversas posibilidades,

abriendo las puertas a los actuales estudiantes a diversas oportunidades tanto laborales como de formación. En el caso de la Facultad de Ingeniería de la Universidad Católica de Salta, se propuso la creación del grupo de incubación de trabajos de investigación de alumnos IDEAS el cual propone un conjunto de líneas de acción tendientes a incentivar y motivar a los alumnos interesados en la I+D+i [2].

Este grupo desde sus inicios hasta en el corriente año, ha logrado atraer a diversos aspirantes que lograron proponer proyectos prometedores y vivir la experiencia de poder realizar y presentar un proyecto por sí mismos. También el mismo grupo dispone de un sitio web con información básica sobre objetivos y constitución del grupo, de los proyectos presentados en años anteriores y contacto de los mismos miembros del grupo.

Potenciar el sistema de información del grupo:

Si bien como se destacó, el grupo logra atraer aspirantes de distintos años y carreras, es reconocible la dificultad que existe para los miembros del grupo el poder estar en constante disposición para clarificar las diversas dudas que estos traen. Por ejemplo, en algunos casos por cuestiones de tiempo y disponibilidad, es difícil definir una reunión presencial, la cual es una de las formas más efectivas de despejar dudas previas a la reunión y las que surjan en la misma. A raíz de esto, se propone emplear una tecnología capaz de simular una conversación humana a través de una interfaz conversacional a fin de que la misma sirva para despejar algunas dudas comunes de los potenciales participantes aclarando fines y objetivos del

mismo grupo, el beneficio que trae participar en el mismo, entre otros aspectos. Realizando el trabajo de responder consultas básicas, esta herramienta podrá potenciar la captación de nuevos participantes de diversos ámbitos, facilitando considerablemente el trabajo de los integrantes.

Desarrollo

Descripción de la propuesta

Este trabajo propone que, a partir de la aplicación de las ciencias de la computación y ramas de la inteligencia artificial definido como Machine Learning y Procesamiento de Lenguaje Natural (PLN), la creación de un bot conversacional que permita a los usuarios, empleando como interfaz un sitio web, poder interactuar con el mismo. De esta forma podrá ingresar una consulta o pregunta y obtener una respuesta casi inmediata que contenga información acerca del grupo IDEAS, propuestas de como formular una idea, consejos a la hora de consolidar esta última como proyecto e incluso sobre donde se presentará dicho proyecto. Se busca que los aspirantes logren comprender más detalladamente lo que brinda el grupo empleando este servicio con atención permanente y poder avivar su interés para participar en el mismo.

Objetivos y alcance

El objetivo de este proyecto es el desarrollo de un agregado a la web del grupo IDEAS que presente una interfaz conversacional sencilla que sirva de conducto para comunicar un bot conversacional que integra tecnologías Machine Learning y Procesamiento de Lenguaje Natural (PLN), que pueda brindar información útil al usuario en función de la consulta que este último realice (véase desde una conversación informal sencilla hasta preguntas de índole informativa en relación a los temas previamente mencionados).

Para alcanzar este objetivo se plantea:

- Construir un listado de preguntas que identifiquen las incógnitas más

importantes o de interés para las diversas personas que busquen interactuar con el bot.

- Analizar la posibilidad de construir e implementar un sistema que emplee la tecnología de ML (Machine Learning).
- Desarrollar un aplicativo web capaz de emplear PLN y ML para responder las consultas de los usuarios.

A partir de estos, y siguiendo un cronograma trazado de 4 semanas, se define el siguiente proyecto cuyo alcance será el diseño de una aplicación web básica que se agregaría a la web oficial del Grupo IDEAS y que sirva de medio para entablar la conversación entre un usuario y el chatbot empleando el lenguaje Python y algunas de sus librerías disponibles. Todo el diseño del prototipo se planteará en el editor de código fuente, Visual Studio Code.

Este prototipo será apodado “I-Chan” a partir de jugar con el término “Chatbot Inteligente” y el sufijo diminutivo honorífico japonés empleado a la hora de dirigirse a un amigo cercano, que resalta la capacidad que se busca que esta herramienta posea al momento de entablar una conversación con el usuario.

Primeramente, se diseñarán varias funcionalidades que le permitan a “I-Chan” responder ante saludos, despedidas, conversaciones triviales (“¿Cómo estás?”, “¿Todo bien?”, etc) y agradecimientos.

Posteriormente se propone diseñar un archivo .CSV que contenga un listado de las preguntas con sus respectivas respuestas en una columna adyacente las cuales fueron consultadas directamente a los miembros del grupo.

Por último se diseñará un algoritmo que evalúe la similitud que tenga la consulta en función de todos los datos que tiene a disposición “I-Chan”, y a partir de ese grado de semejanza devuelva una respuesta adecuada al contexto expresado por el usuario. Se emplearán conceptos propios del

preprocesamiento de texto como, a su vez, herramientas proporcionadas por el paquete Stanza [3] creado por el grupo de PLN de Stanford y algoritmos de aprendizaje automático pertenecientes a la librería Scikit-learn [4].

En cuanto se concluya lo planteado previamente y se haya diseñado un sitio web básico empleando el framework Flask e incorporando CSS y HTML (en el cual se basará en modelos ya existentes) [5], se podrá concretar la interfaz (Figura 1):



Figura 1. Interfaz prototipo de Chat

Características funcionales del sistema

Primeramente, “I-Chan” se presenta al usuario a través de un globo de dialogo. El usuario desde ese mismo instante tiene la posibilidad de ingresar cualquier tipo de consulta en el espacio de texto. Una vez cargado y al interactuar con el botón, se dispara el algoritmo cuyo bloque se observa en Figura 2.

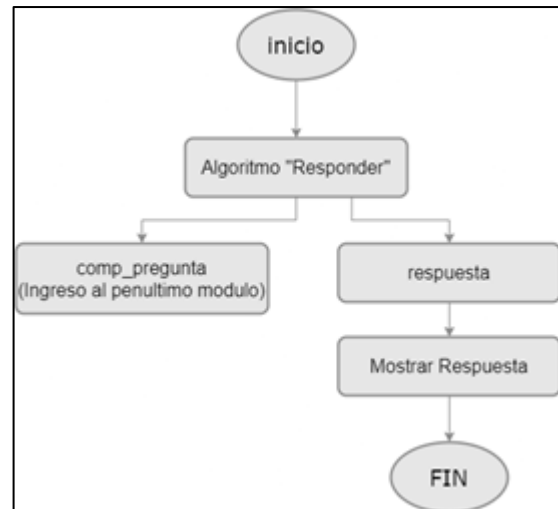


Figura 2. Algoritmo de la funcionalidad “Responder”

Este algoritmo toma como insumo la consulta del usuario y toma las palabras claves de la misma y evalúa su similitud con las frases previamente cargadas en varias instancias. Para realizar la comparación, se deberá convertir tanto la información que dispone el chatbot como la consulta en minúscula, para que el algoritmo no trate de forma diferente, palabras que fueron expresadas de distintas formas. Posteriormente se empleará un algoritmo perteneciente a la librería Scikit-Learn que sigue el enfoque “Term Frequency-Inverse Document Frequency” (TF-IDF) [6] el cual determina la frecuencia de palabra por el total de las palabras y cuan rara es esa palabra en función de su aparición en la información que dispone el chatbot. Con esto, puede determinar la importancia o peso de una palabra. Para poder emplear esta función es necesario convertir las cadenas de texto que se dispone (incluyendo la consulta) en una lista de tokens, es decir, en una lista de palabras claves, eliminando signos de puntuación y empleando el proceso de lematización el cual consiste en reemplazar las palabras que están en su forma flexionada por su lema correspondiente a fin de que se hallen similitudes (para ello se empleara la librería de Stanza que dispone un modelo de lematización en español).

A partir de los valores que se obtienen por TF-IDF, se pueden representar como vectores de valor real en espacio vectorial, por lo que se puede emplear la similitud de cosenos (otro algoritmo empleado de la librería Scikit-Learn) que permite evaluar la similitud de la consulta con las diversas frases referentes que posee “I-Chan”; si son semejantes, entonces el ángulo será pequeño, por lo que el coseno del mismo será cercano a 1, que representará una gran similitud [7].

La función de Scikit-Learn para similitud de coseno retorna una matriz con los “grados de similitud” obtenidos. A partir de esto se vectoriza y ordena de menor a mayor de forma que el penúltimo “grado de similitud” represente a la frase que tiene más similitud con la consulta realizada. En función de esta se devuelve una respuesta aleatoria en caso de ser un saludo, agradecimiento, despedida o conversación cotidiana. Para que esto suceda se diseñaron 4 módulos que harán uso de listas que contienen expresiones que podría emplear el usuario y otra de respuestas que podría dar “I-Chan” en el respectivo contexto. A partir de la entrada realizada por el usuario, se comparará con las listas de ejemplos en cada contexto y si la similitud existente es mayor a un 50%, se devuelve una cadena de texto aleatoria proveniente de una de las respectivas listas de respuestas del bot.

Si dicho modulo no devuelve ningún valor se procede a probar en la siguiente instancia y así sucesivamente hasta que devuelva una respuesta y la muestre al usuario. Esta evaluación de retorno se realiza en un método propio de la clase denominado “Responder” que busca descartar la posibilidad de que sea una consulta respecto al grupo; que en caso de que fuese así, accederá a las ultimas 2 funciones del sistema que son “comparación de preguntas” y “Comparación de respuestas”, definiendo un total de 9 métodos contenidos en la clase(incluyo los de preprocesamiento de texto). Las funciones cuyo retorno es

evaluado en la instancia “Responder” se pueden observar en la figura 3.

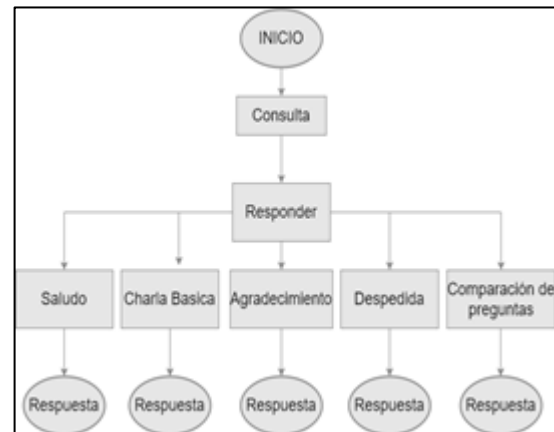


Figura 3. Función “Responder”

A partir de descartar que fuera una conversación cotidiana, existe la posibilidad entonces que este consultando acerca de información relacionada con el grupo; por ende se procede acceder a la función de “Comparación de preguntas” en el que se contrapondrán las preguntas que existen en el csv que se tomaron como referentes para “entrenar” al chatbot y la consulta del usuario.

Si existe una similitud superior a la del 50%, entonces tiene un considerable grado de probabilidad de que sean preguntas de un mismo tema y por ende, procederá a devolver dicha respuesta. En caso contrario, se almacena el índice y el grado de precisión obtenido y se pasa al último modulo. En este se compara las diversas respuestas cargadas en el archivo CSV para evaluar la similitud que tenga con la consulta hecha por el usuario en función de las palabras claves de la misma. A partir de esto, se evalúa si ambos grados de precisión son nulos; en caso de que no lo sean, se pregunta si la precisión del módulo anterior es mayor a la obtenida en esta instancia. En caso de que no lo sea, y que la precisión de esta instancia sea superior a un 50%, se devolverá dicha respuesta y adjuntará la pregunta hecha por el usuario junto a dicha respuesta en el archivo CSV. Para ello, se diseñará e

implementará una clase a parte la cual se instanciará en esta función a fin de poder abrir y modificar el archivo. Cabe destacar que esta también contendrá algunas funciones para el tratamiento del texto.

En caso contrario, o bien el grado de precisión del módulo anterior es mayor, por lo que se devolverá la respuesta empleando el índice que se almaceno previamente y guardando dicha pregunta en el archivo; o de otro modo, la precisión es nula o por debajo del porcentaje antes mencionado, por lo que se supone que no existe similitud entre la consulta y la información que tiene a disposición el bot por lo que devolverá una respuesta afirmando ser incapaz de responder a su consulta.

Cabe destacar finalmente que cuando se emplea como condición que el grado de similitud entre la consulta y la información almacenada sea mayor que un 50%, se basa en que al estar por encima de la media, se presupone una considerable semejanza. Sin embargo, la fiabilidad y precisión de la respuesta que “I-Chan” pueda proveer en función de su consulta quedara determinada por cuan extenso sea el conjunto de datos que disponga.

Durante la etapa de prueba del prototipo, emplearemos la consulta “que es ideas?” la cual, como se puede observar, carece de signos de acentuación e interrogación como también inicia su consulta con minúscula. A partir de que se envía la consulta, se buscara la semejanza con las frases que posee a su disposición el bot en diversas instancias hasta que allá un cierto grado similitud con la pregunta “¿Qué es el grupo ideas?”; por lo que devolverá su respectiva respuesta (Figura 4).

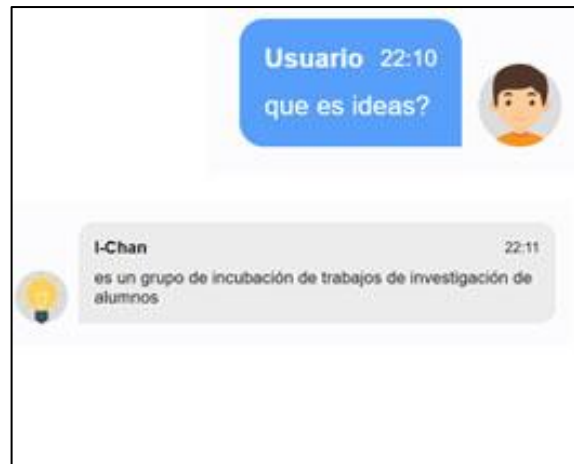


Figura 4. Ejemplo de Conversación

Trabajos Relacionados:

Desde sus inicios, el termino chatbot ha tenido un crecimiento casi exponencial y paulatino, destacando que en 2016 tomó una particular relevancia dentro del comercio digital cuando diversas plataformas de interfaz conversacional (como Facebook Messenger, Telegram, Skype, etc) habilitaron herramientas de desarrollo para la creación de chatbots, consiguiendo que el time-to-market de este tipo de soluciones se tornara altamente atractivo.

A su vez, en el ámbito académico, se puede encontrar beneficioso un sistema de asistencia perpetua con respuesta instantánea, por lo que se pudo encontrar una serie de trabajos orientados al contexto universitario semejantes al presentado en este documento.

A nivel internacional, se puede destacar el proyecto “Smart College Chatbot using ML and Python” [8] que presenta un sistema de chatbot universitario basado en una aplicación web que responde consultas de usuario.

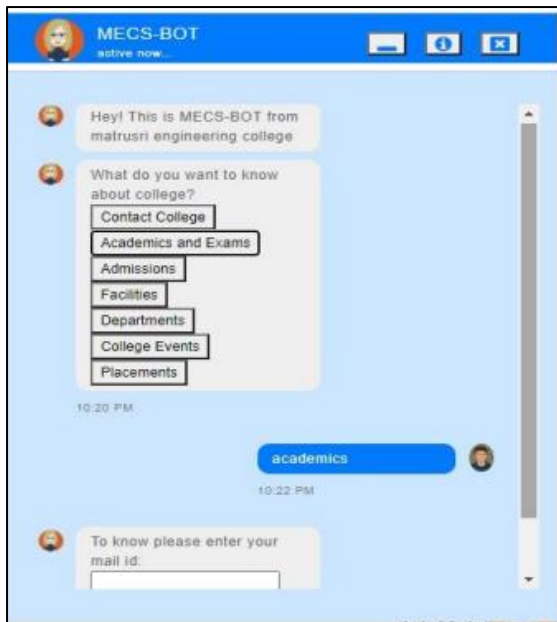


Figura 5. Imagen del menú del chatbot

Este proyecto fue presentado en diciembre de 2020 por varios alumnos de la facultad de ingeniería de la universidad Matrusri en Hyderabad ,India.

Este presenta la posibilidad de definir entre diversos tópicos para poder hacer consultas, requiriendo a su vez que este se registre y así poder hacer diversas consultas acerca del tema seleccionado, un ejemplo se muestra en la Figura 5.

Otro trabajo realizado es “Development of a chatbot for informing students of the Schedule” [9] que fue desarrollado en la Universidad Pedagógica Estatal Bohdan Khmelnytsky Melitopol en Ucrania. Este artículo se basa en el desarrollo de un chatbot para informar a los estudiantes sobre sus horarios diseñado para Telegram de forma que puedan minimizar el tiempo perdido y optimizar sus actividades estudiantiles. La conversación con el mismo puede visualizarse en la figura 6.

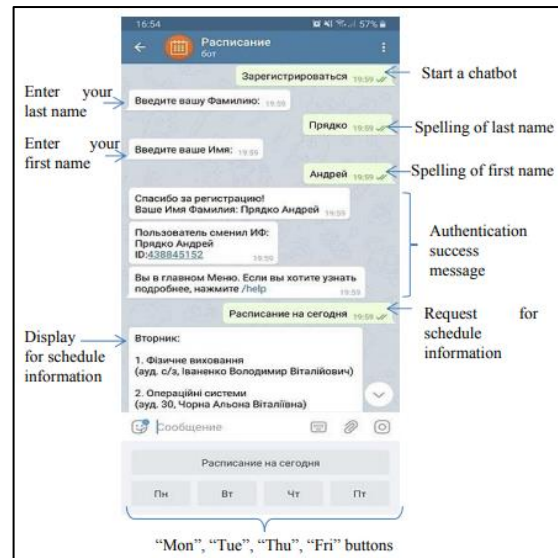


Figura 6. Imagen del chatbot en Telegram

Otro ejemplo es el artículo “Realtime Processing of IoT Events using a Software as a Service (SaaS) Architecture with Graph Database” [10] redactado en 2017 desde la universidad “St.john de Ingeniería y Tecnología” en India el cual propone un sistema para entablar una conversación y ayudar a un usuario utilizando LUIS (Language Understanding) e implementado en la nube publica AWS.

Este Chatbot analiza si los mensajes del usuario tienen solución o no en función de datos almacenados previamente. Si existe, el chatbot iniciara su interacción con el usuario.

Finalmente, en julio del corriente año, se presentó a “Clementina”, chatbot diseñado por la Facultad de Ciencias Exactas y Natural de la Universidad de Buenos Aires [11], la cual puede responder a todo tipo de consultas relacionadas a la facultades y colegios de la universidad, así como también de trámites y otros tópicos relacionados.

Es claro destacar que las propuestas mencionadas siguen la estructura de un bot informativos e inteligentes. A partir de esto, es resaltable que la diferencia entre estos o la propuesta realizada en el actual proyecto está directamente ligada al ámbito de implementación.

Esta tecnología no solo puede proveer una ventaja en cuanto a optimización de tareas y captación de nuevos participantes para el grupo, sino también se la puede llegar a considerar como el primer paso en la integración de esta herramienta innovadora y versátil en diversos entornos de la Universidad Católica de Salta.

Conclusión y Trabajos Futuros:

Actualmente el proyecto se encuentra en fase de prototipo, limitado en cuanto a la información que dispone y el sistema que emplea para almacenarlas siendo un máximo de 20.000 líneas. También cabe destacar que permanece en la etapa de “entrenamiento”, en la que se evalúa la precisión de las respuestas en función de la entrada de datos que tiene.

Por otro lado, a futuro, se propone el desarrollo de una interfaz más interactiva y dinámica que pueda presentarse adjunta a la página oficial del grupo IDEAS, a fin de que en el primer instante que el usuario acceda al sitio, sea recibido por “I-Chan”. De esta forma se maximiza la interacción y el acceso a la información por parte del usuario, evitando la necesidad de acceder a otra pestaña para poder realizar las consultas pertinentes.

También será necesario proponerse a largo plazo, implementar un mejor medio para preservar los datos que posee el usuario a fin de que el acceso a esto sea más eficiente y rápido.

Al mismo tiempo se deberá plantear con los miembros del grupo la incorporación de esta herramienta al sitio web reconociendo si es necesario la adquisición de infraestructura para su implementación o si la adhesión de este no demandará ningún cambio a excepción del diseño de una estructura de datos acorde al modelo que actualmente para preservar la información.

Por último, sería idóneo añadir funcionalidades extras al sistema actual como:

- En caso de que el sistema no pueda contestar alguna incógnita del usuario, pueda proponer que dicha consulta sea realizada personalmente (sea presencial o virtual) presentando un modelo de solicitud de reunión, a fin que una vez que el usuario la complete, pueda ser enviada a un respectivo miembro del grupo.
- Poder recibir solicitudes de nuevos aspirantes y comunicarlas a un miembro del grupo por vía mail.

Finalmente, es claro resaltar las diversas posibilidades que provee el uso de chatbots como medio informativo e interactivo, pudiendo proveer atención constante y útil para los usuarios en cualquier tipo de ámbito (tanto académico, comercial o en salud), automatizando cualquier tipo de tarea.

También cabe destacar que el avance en el Procesamiento del Lenguaje Natural y en la Inteligencia Artificial en general han servido para la aparición de chatbots más inteligentes, capaces de comprender diversas expresiones del usuario y poder brindar una respuesta adecuada. Ante estas tecnologías, es necesario hacer un buen uso para poder incentivar a los estudiantes a que sigan el camino o mínimamente vivan la experiencia que trae la participación en trabajos de investigación y desarrollo.

Agradecimientos:

Agradezco a la Universidad Católica de Salta, en especial, al grupo IDEAS por el apoyo y acompañamiento desde el inicio de este proyecto.

Referencias:

- [1] N. . A. Manzaba Campaña y E. . D. I. L. Perez, «Importancia del I+D+i en las Empresas y el Desarrollo Nacional,» 03 Julio 2019.
- [2] «Grupo IDEAS,» [En línea]. Available: <https://ideas.ucasal.edu.ar/>.
- [3] S. N. Group, «Stanza – A Python NLP Package for Many Human Languages,» 2020.

- [En línea]. Available:
<https://stanfordnlp.github.io/stanza/>.
- [4] F. Pedregosa, G. Varoquaux, A. Gramfort y V. Michel, «Scikit-learn: Machine Learning in Python,» 2007. [En línea]. Available:
<https://scikit-learn.org/stable/>.
- [5] H. Sayyed, «Create Web Based ChatBot in Python, Django, Flask,» *studygyaan*, 30 marzo 2020.
- [6] P. Pandey, «Construyendo un Chatbot simple desde cero en Python (usando NLTK),» *Planeta CHATBOT*, 12 junio 2019.
- [7] J. Vembunaryanan, «Tf-Idf and Cosine similarity,» *Seeking Wisdom*, 27 Octubre 2013.
- [8] K. Ashok Kumar , «Smart College Chatbot using ML and Python,» 2020.
- [9] K. Osadcha, A. . O. Priadko y B. Khmelnytsky, «Development of a chatbot for informing students of the schedule,» 2020.
- [10] G. D'silva, S. Thakare y V. . A. Bharadi, «Real-time processing of IoT events using a Software as a Service (SaaS) architecture with graph database,» 2016.
- [11] L. U. N. d. B. Aires, «Clementina,» 2021. [En línea]. Available:
<https://www.uba.ar/#/clementina>.

Datos de Contacto:

*Pablo Sandoval. Universidad Católica de Salta –
Facultad de Ingeniería. Campus Castañares, A4400
Salta. pablosandoval2191@gmail.com*