datistix

Documento alcance proyecto
Fecha 26/09/2019
---
info@datistix.io
datistix.io

Proyecto	S1 Gateway
Líder de Proyecto	Federico Garcia Blanco
Producto	NLP ChatBox

Autor	Federico Garcia Blanco	Fecha	26/09/2019

Introducción

El objetivo del presente informe es dar a conocer los detalles técnicos de la primera exploración de los datos entregados por el cliente S1 Gateway con el propósito de desarrollar y mejorar el comportamiento de respuesta del ChatBox para servicio al cliente. Para ello, se va a utilizar el lenguaje de programación Python a través del entorno Jupyter lab haciendo uso de las librerias pandas, numpy, matplotlib, scikit learn entre otras.

EDA

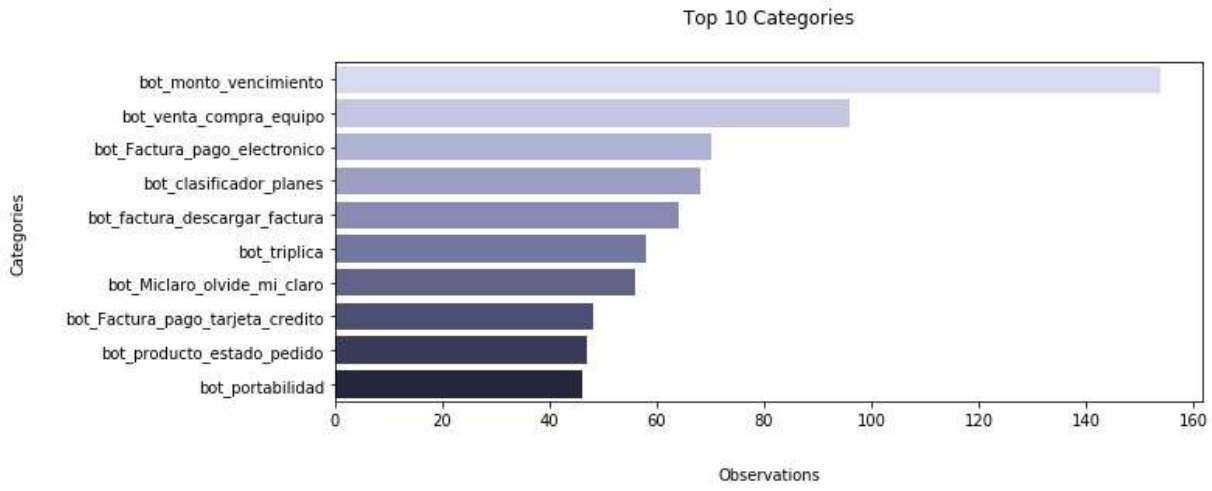
Durante la etapa de análisis exploratorio de datos, se encontró que el dataset se compone de:

1) Tres columnas: Intent, Phrase - NLP, Phrase
2) Filas: 3213
3) Categorías o clases: 291
4) Observaciones por clase: Entre 1 y 154 observaciones por clase.

info@datistix.io
datistix.io

Visualización

Durante la etapa de visualización se hizo un análisis directo, en el cual se identificaron cuales son las categorías con mayor cantidad de resultados como se muestra en el gráfico siguiente:



Por otro lado, en una nube de palabras se identifican cuales fueron los términos mas utilizadas, excluyendo las preposiciones, y caracteres especiales.

info@datistix.io
datistix.io

Feature Engineering

Durante la etapa de preprocesamiento se optó por las siguientes modificaciones en el dataset:

1) Se transformaron todas las palabras a minúscula dado que se considera que no influyen en lo que quiere expresar el cliente. Análogamente se procede con las tildes encontradas. “Lazy evaluation”.
2) Se trabajó con las columnas Intent y Phrase, porque se considera que Phrase–NLP está altamente correlacionada con Phrase, y por lo tanto no hace un aporte significativo a la predicción de la categoría.
3) Se eliminan las preposiciones y caracteres especiales para que queden solamente las palabras que apunten a la necesidad del usuario y ayuden a identificar la categoría.
4) Se redujo el dataset a categorías con 20 o más observaciones para quedarnos con la información más robusta y de esta forma sacar los outliers.
5) Se utilizó steeming para asociar palabras que se refieren al mismo concepto.
6) Se optó por el método de tokenizacion para organizar el dataset.

Con estas modificaciones se logra una tabla relacional con redundancia que permite la aplicación de un modelo de aprendizaje automatizado para lograr predecir las categorías del ChatBox.

Machine Learning

Para la etapa de Machine Learning se separaron los datos en 80% para el entrenamiento del modelo y 20% para el testeo, de esta forma se puede evaluar el modelo con datos que no fueron incluidos en su entrenamiento. Luego aplicando el modelo de Support Vector Machine, como primer testeo, y con la métrica ‘mean accuracy’ se obtuvo una precisión de 73%.

info@datistix.io
datistix.io

Conclusión

Siendo el primer acercamiento realizado a los datos entregados, se concluye que un proceso de aprendizaje automatizado puede brindar valor agregado al servicio de ChatBox de S1Gateway ya que es capaz de predecir con cierta precisión la categoría de las preguntas que elaboran los clientes.

Adicionalmente, se podría mejorar el modelo si además de contar con mas cantidad de datos para tener una muestra robusta, se programa el proceso de aprendizaje automatizado con más parámetros y mejores practicas de ciencia de datos. Generar mecanismos de validación cruzada para evitar errores de sobre ajustamiento, evaluar diferentes métricas de error, reconocer categorías poco dominantes, identificar errores de asignación de categorías entre otras, serian las practicas a tener en cuenta para profundizar aún más en el desarrollo de este proceso de lenguaje natural.