Trabajo práctico
Text Mining

Equipo N° 3: Stefano Canossini, Marcela Distefano, Hernán Ifrán, Damián Joglar y José Valdés

UNIVERSIDAD
AUSTRAL

Proyecto: Clasificación de tipo de fallas de productos

Alcance:

- Desarrollar una solución que permita automatizar la clasificación de reclamos a partir de la información obtenida en formularios web.

Descripción del problema

La empresa Escorial se dedica a la producción en serie de termotanques, cocinas y calefones, siendo el foco del negocio ofrecer el mejor precio del mercado.

Al ser elevada la cantidad de productos manufacturados, se requiere de un área de postventa encargada de atender solicitudes concernientes a los productos. Se brinda la posibilidad al cliente de completar un formulario en la página web https://escorial.com.ar/postventa en caso de existir algún problema con el producto. El formulario solicita datos de contacto, del producto con inconvenientes y descripción del problema para que el área pueda resolver. A continuación, se presenta screenshot del formulario que se observa en el sitio web:

Formulario de reclamos en la web de la empresa

Servicio Postventa

Solicitud de Servicio Técnico a Domicilio.
Complete el siguiente formulario y nos contactaremos con Usted a la brevedad.

DATOS DE CONTACTO

Nombre*
Apellido*
Email*
Teléfono*
Celular*
DNI*
Dirección Completa*
Entre Calles
Localidad/Barrio*
Provincia*
Motivo de la solicitud*

DATOS DEL PRODUCTO

Producto*
Tipo de Instalación*
Modelo*
Número de Serie*

En la cocina, el número de serie lo podrá encontrar en el Manual de Instrucciones (primera hoja del manual, "número de garantía"); y en la etiqueta frontal de la cocina (ángulo derecho puerta del horno), o en la placa de marcado (visible al retirar la plancha, en uno de los laterales). En el termotanque el número de serie se encuentra en la placa de marcado en la parte frontal del artefacto.

DESCRIPCIÓN DEL PROBLEMA*

Tareas de los operadores

- En base a la observación que escriba el cliente, el operador debe clasificar el problema para que el técnico lo resuelva, realizando lo que por tabla está tabulado.
- El operador cuenta con muchas tareas operativas, para agilizar esta tarea se realizará la clasificación del problema de forma automática quedando solo el control posterior a cargo del operador.

Hoja de ruta: Pasos tentativos.

1. Obtener los datos de la base de datos donde se aloja la página web de la empresa.
2. Exportarlos como csv o documento de excel para su posterior procesamiento.
3. Limpieza de los datos dejando solamente las clasificaciones realizadas por los operadores.
4. Análisis de datos faltantes o incorrectos y su manipulación.
5. Creación del modelo a partir de las observaciones y clasificaciones.
6. Realizar el test y medir el nivel de precisión del modelo.
7. Implementar el script correspondiente para que realice la clasificación requerida y se aplique sobre la base de datos.
8. Presentación de la mejora a la gerencia.

Expectativas

- Por cada tipo de producto existe una lista de problemas específicos y estimamos que vamos a tener complicaciones en designar la categoría correcta sin mezclar entre tipo de producto.

- Al procesar las observaciones muy probablemente tengamos problemas al clasificarlos en caso de que el cliente escriba de una manera incorrecta.

- Un punto importante para considerar es la inexperiencia del equipo en el procesamiento de texto lo cual la va a provocar que el tiempo del proyecto se extienda más allá del tiempo normal de cualquier proyecto de esta índole hasta lograr obtener el modelo correcto.

Problemas en el camino

- La clasificación se realiza en base a categorías de desperfectos. Del análisis de los formularios web resulta:
- Categorías redundantes.
- Categorías que fueron utilizadas muy pocas veces
- Categorías que, si bien están contempladas, el cliente no usa e indica OTROS
- En una primera instancia nos confiamos con un conjunto de datos que resulto insuficiente.
- Los clientes escriben de manera coloquial y con mala ortografía

Modelo

- Se procede a probar el aplicativo con información generada por la empresa desde el año 2021 hasta finales del mes de agosto 2023. De los datos adquiridos se resumen los principales enunciados para el modelo en la imagen siguiente:

	problema_id	descripcion	alias_8_nombre3	problema_n	obsitem	Concaobsitem
0	70e03aaf-6cfb-42e8-92ae-fcb46ff479dd	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.)	GN - Gas Natural	B1 - Pérdidas de gas con piezas dañadas	HAY UNA PERILLA DE LAS HORNALLA QUE NO SE PUED...	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.) GN...
1	e5be47a0-c506-4a79-87e9-2a45f06d496c	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.)	GN - Gas Natural	G - Perilla no gira / gira mal	HAY UNA PERILLA DE LAS HORNALLA QUE NO SE PUED...	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.) GN...
2	2b2af341-20c5-4607-baa0-01972617647e	TERMO ELECTRICO EL-55	ELE - Termo eléctrico	D - No enciende	DESPUÉS DE QUE UNA PERSONA SE DUCHO NO VOLVIÓ ...	TERMO ELECTRICO EL-55 ELE - Termo eléctrico " ...
3	ab0ed03a-8aab-4317-9aad-fcc06dd41520	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.)	GN - Gas Natural	D2 - Hornalla Mal funcionamiento	NaN	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.) GN...
4	ab0ed03a-8aab-4317-9aad-fcc06dd41520	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.)	GN - Gas Natural	D2 - Hornalla Mal funcionamiento	LAS HORNALLAS TARDAN MUCHÍSIMO TIEMPO EN PREND...	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.) GN...

- La variable problema_n se convierte en el objeto del modelo para predecir la categorización que en la operación es realizada por el operador en turno, esto en función de lo consignado por el usuario.

Modelo

- Limpieza de datos: La base trabajada en los análisis iniciales se observa que presenta campos vacíos, esto se asume que son de los clientes que en ocasiones no diligencian la información completa del formulario, estos campos son eliminados para desarrollar un análisis más robusto.
- Preprocesamiento: Se implementa stemmer para el desarrollo de un modelo de text mining, a continuación, se presenta el bloque de código utilizado:

```
# Preprocesamiento de texto
stemmer = SnowballStemmer('spanish')
stop_words = set(stopwords.words('spanish'))
```

Por otro lado, en este mismo bloque se definen las palabras que no agregan valor en la revisión.

- Definición de función para preprocesar el texto a trabajar:

```
def preprocesar_texto(texto):
    palabras = nltk.word_tokenize(texto.lower())
    palabras = [stemmer.stem(palabra) for palabra in palabras if palabra.isalpha() and palabra not in stop_words]
    return ' '.join(palabras)
```

Modelo

- Se define un procedimiento para la clasificación de texto implementando TF-IDF y un kernels SVC lineal:

```
# Creación del pipeline para el clasificador basado en texto
pipeline = Pipeline([
    ('tfidf', TfidfVectorizer(max_features=5000)),  # Experimenta con diferentes valores para max_features
    ('clf', SVC(kernel='linear'))  # Experimenta con diferentes kernels
])
```

- En el proceso de validación se divide de una parte de los datos en entrenamiento y prueba, en los resultados obtenidos se alcanza una validación cruzada Accuracy de 74,68%.

```
Accuracy en validación cruzada: 0.7468176381841768
```

Modelo

- Clases identificadas en el conjunto de pruebas:

	precision	recall	f1-score	support
B - Descargas de electricidad	0.33	0.15	0.21	20
B - La llama no enciende al abrir el grifo de agua	0.78	0.97	0.86	33
B - Pérdidas de gas – sin/con piezas quemadas	0.40	0.11	0.17	19
B1 - Pérdidas de gas con piezas dañadas	0.33	0.12	0.18	67
B2 - Pérdidas de gas sin piezas dañadas	0.67	0.71	0.69	195
C - La llama se apaga luego de algunos minutos de estar funcionando	0.93	0.76	0.84	17
C - Pérdidas de agua	0.91	0.95	0.93	347
C2 - Horno Mal funcionamiento	0.81	0.67	0.74	301
D - El agua sale con temperatura baja	1.00	1.00	1.00	2
D - No enciende	0.56	0.51	0.53	215
D - No enciende piloto	0.43	0.28	0.34	57
D2 - Hornalla Mal funcionamiento	0.71	0.85	0.77	469
E - El agua sale con temperatura baja / alta	0.70	0.82	0.75	382
E - Piloto se apaga	0.76	0.88	0.81	204
E - Puerta de horno - No cierra / Se cae	0.89	0.96	0.92	487
F - No calienta agua Mal funcionamientoQuemador no enciende Quemador se apagaQuemador no se apaga	0.79	0.55	0.65	67
F - Puerta Parrilla - No cierra	0.57	0.19	0.29	21
F - Pérdida de agua	1.00	0.50	0.67	6
.50 0.57 38				
M - Vidrio estallado	0.91	0.83	0.87	36
micro avg	0.76	0.76	0.76	3127
macro avg	0.73	0.57	0.60	3127
weighted avg	0.75	0.76	0.74	3127

Modelo

- Validación con nuevos datos:

	Descripcion
0	TERMO 45 L GN (U.) GAS - Termo a gas " EL EQUI...
1	TERMO ELECTRICO EL-55 ELE - Termo eléctrico " ...
2	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.) GN...
3	TERMO 80 L GN (U.) GAS - Termo a gas " NO SE M...
4	TERMO ELECTRICO EL-90 ELE - Termo eléctrico " ...
5	COCINA CANDOR S2 (GN) (U.) GN - Gas Natural " ...
6	COCINA MASTER S2 BL CLASSIC. (GN.) (U.) GN - G...
7	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.) GN...
8	COCINA MASTER S2 NEG.CLASSIC (GN) (U.) GN - Ga...
9	COCINA MASTER S2 NEG.CLASSIC (GN) (U.) GN - Ga...
10	TERMO ELECTRICO EL-55 ELE - Termo eléctrico " ...
11	TERMO 45 L GN (U.) GAS - Termo a gas "

- Clasificación del algoritmo:

	Problema	Etiqueta de Reclamo
0	TERMO 45 L GN (U.) GAS - Termo a gas " EL EQUI...	C - Pérdidas de agua
1	TERMO ELECTRICO EL-55 ELE - Termo eléctrico " ...	C - Pérdidas de agua
2	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.) GN...	E - Puerta de horno - No cierra / Se cae
3	TERMO 80 L GN (U.) GAS - Termo a gas " NO SE M...	E - Piloto se apaga
4	TERMO ELECTRICO EL-90 ELE - Termo eléctrico " ...	C - Pérdidas de agua
5	COCINA CANDOR S2 (GN) (U.) GN - Gas Natural " ...	E - Puerta de horno - No cierra / Se cae
6	COCINA MASTER S2 BL CLASSIC. (GN.) (U.) GN - G...	D2 - Hornalla Mal funcionamiento
7	COCINA PALACE CRISTAL BLACK LX S2 (GN) (U.) GN...	D2 - Hornalla Mal funcionamiento
8	COCINA MASTER S2 NEG.CLASSIC (GN) (U.) GN - Ga...	D2 - Hornalla Mal funcionamiento
9	COCINA MASTER S2 NEG.CLASSIC (GN) (U.) GN - Ga...	B2 - Pérdidas de gas sin piezas dañadas
10	TERMO ELECTRICO EL-55 ELE - Termo eléctrico " ...	D - No enciende
11	TERMO 45 L GN (U.) GAS - Termo a gas "	D - No enciende piloto
12	COCINA MASTER STYLE INOX. MULTIGAS (U.) MG - M...	D2 - Hornalla Mal funcionamiento
13	TERMO ELECTRICO EL-90 ELE - Termo eléctrico " ...	D - No enciende
14	COCINA CANDOR S2 BLACK GL (U.) GE - Gas Envasa...	E - Puerta de horno - No cierra / Se cae

Modelo

- Validación con nuevos datos, 1 registro:

	Descripcion
0	TERMO 45 L GN (U.) GAS - Termo a gas " EL EQUIPO PIERDE AGUA POR DENTRO, INCLUSO LO FALLA ES DE ANTES DE ENCENDERLO POR PRIMERA VEZ.

- Clasificación del algoritmo, 1 registro:

	Problema	Etiqueta de Reclamo
0	TERMO 45 L GN (U.) GAS - Termo a gas " EL EQUIPO PIERDE AGUA POR DENTRO, INCLUSO LO FALLA ES DE ANTES DE ENCENDERLO POR PRIMERA VEZ.	C - Pérdidas de agua

- Se observa que el registro se clasifica acorde a los históricos de la falla, en las pruebas realizadas se observa que la clasificación fue correcta al 100%, 16 registros de 16 registros clasificados.

Resultado

- ACCURACY = 74.68

- OBSERVACIONES: Si bien consideramos que el modelo clasifica con un accuracy aceptable un problema, en una segunda etapa se implementarán mejoras al mismo con el objeto de clasificar múltiples problemas.

GRACIAS

- UN EXPERIMENTO NO SE LE NIEGA A NADIE