Segmentación de Tweets comparando diferentes técnicas de modelado

Nikolas Sebastián Rodriguez Alfonso

Científico de Datos

https://github.com/NikolasRodriguezA/Modelado_Topicos_Davivienda

Exploración de Datos

- Dataset con 12 variables y 1811 observaciones, destacando la variable Embedded_text, siendo esta la variable que generará el corpus para un posterior modelamiento.

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Unnamed: 0      1811 non-null   int64
 1   UserScreenName  1807 non-null   object
 2   UserName        1811 non-null   object
 3   Timestamp       1811 non-null   object
 4   Text            1811 non-null   object
 5   Embedded_text   1811 non-null   object
 6   Emojis          295 non-null    object
 7   Comments        838 non-null    float64
 8   Likes           247 non-null    object
 9   Retweets        643 non-null    object
 10  Image link      1811 non-null   object
 11  Tweet URL       1811 non-null   object
dtypes: float64(1), int64(1), object(10)
memory usage: 169.9+ KB
```

1811 rows × 12 columns

UserScreenName	UserName	Timestamp	Text	Embedded_text	Emojis	Comments	Likes	Retweets
Andrés Langebaek	@ALangebaek	2021-12-01T20:43:12.000Z	Andrés Langebaek\n@ALangebaek\n·\n1 dic.	La confianza se afectó. El indicador de confia...	NaN	1.0	7	19
Plaza Futura	@plaza_futura	2021-12-01T21:18:10.000Z	Plaza Futura\n@plaza_futura\n·\n1 dic.	Buscamos la accesibilidad y mejor atención en ...	✅✅✅✅✅	NaN	NaN	NaN
Julián Martinez	@JulianM998	2021-12-01T22:49:11.000Z	Julián Martinez\n@JulianM998\n·\n1 dic.	Señores \n@Davivienda\n no he podido ingresar ...	NaN	1.0	NaN	1
Ferchis.	@fergomezr28	2021-12-01T12:29:07.000Z	Ferchis.\n@fergomezr28\n·\n1 dic.	Llevo toda una semana sufriendo intento de hur...	NaN	2.0	1	2
MirandaL2	@MirandaSuspLo	2021-12-01T20:52:36.000Z	MirandaL2\n@MirandaSuspLo\n·\n1 dic.	Hemos retrocedido tanto en este país con este ...	NaN	3.0	NaN	8
...	...	...	...	...	...	...	...	...
Banco Davivienda	@Davivienda	2021-12-22T18:26:38.000Z	Banco Davivienda\n@Davivienda\n·\n22 dic.	En respuesta a \n@JaimeMolina\nBuenas tardes. ...	NaN	1.0	NaN	NaN
Banco Davivienda	@Davivienda	2021-12-22T20:18:40.000Z	Banco Davivienda\n@Davivienda\n·\n22 dic.	En respuesta a \n@josefe71\nHola Jose , gracia	NaN	1.0	NaN	NaN

```
data.iloc[0]['Embedded_text']
```

✓ 0.6s

```
'La confianza se afectó. El indicador de confianza Davivienda tuvo una  leve caída en noviembre, rompiendo una tendencia de cinco meses de mejoras.
especialmente en la última semana del mes, asociado al aumento en la tasa de cambio.\n1\n7\n19'
```

Exploración de Datos

- Exploración tweets más relevantes

```
data['Likes'].value_counts()
```

```
0          1564
1           143
2            58
3            12
4             6
6             5
7             4
5             4
10            2
19            1
29            1
11            1
16            1
14            1
838           1
15            1
1,5 mil       1
115           1
9             1
8             1
326           1
13            1
Name: Likes, dtype: int64
```

```
data['Retweets'].value_counts()
```

```
Output exceeds the size limit. Oper
0          1168
1           343
2           126
3            48
4            36
5            16
6            10
7             9
11            7
14            6
9             6
8             5
10            3
13            3
19            3
23            2
25            2
20            1
1 mil         1
855           1
12            1
252           1
2,5 mil       1
17            1
```

```
data[data['Likes']=='1,5 mil']
```

	Unnamed: 0	UserScreenName	UserName	Timestamp	Text	Embedded_text	Emojis	Comments	Likes	Retweets	Image link	
1134	1134	Wilson Arias	@wilsonariasc	2021-12-17T10:20:00.000Z	Wilson Arias\n@wilsonariasc\n·\n17 dic.	He conocido de primera mano un caso en el que ...	0	87.0	1,5 mil	2,5 mil	[]	https://twitt

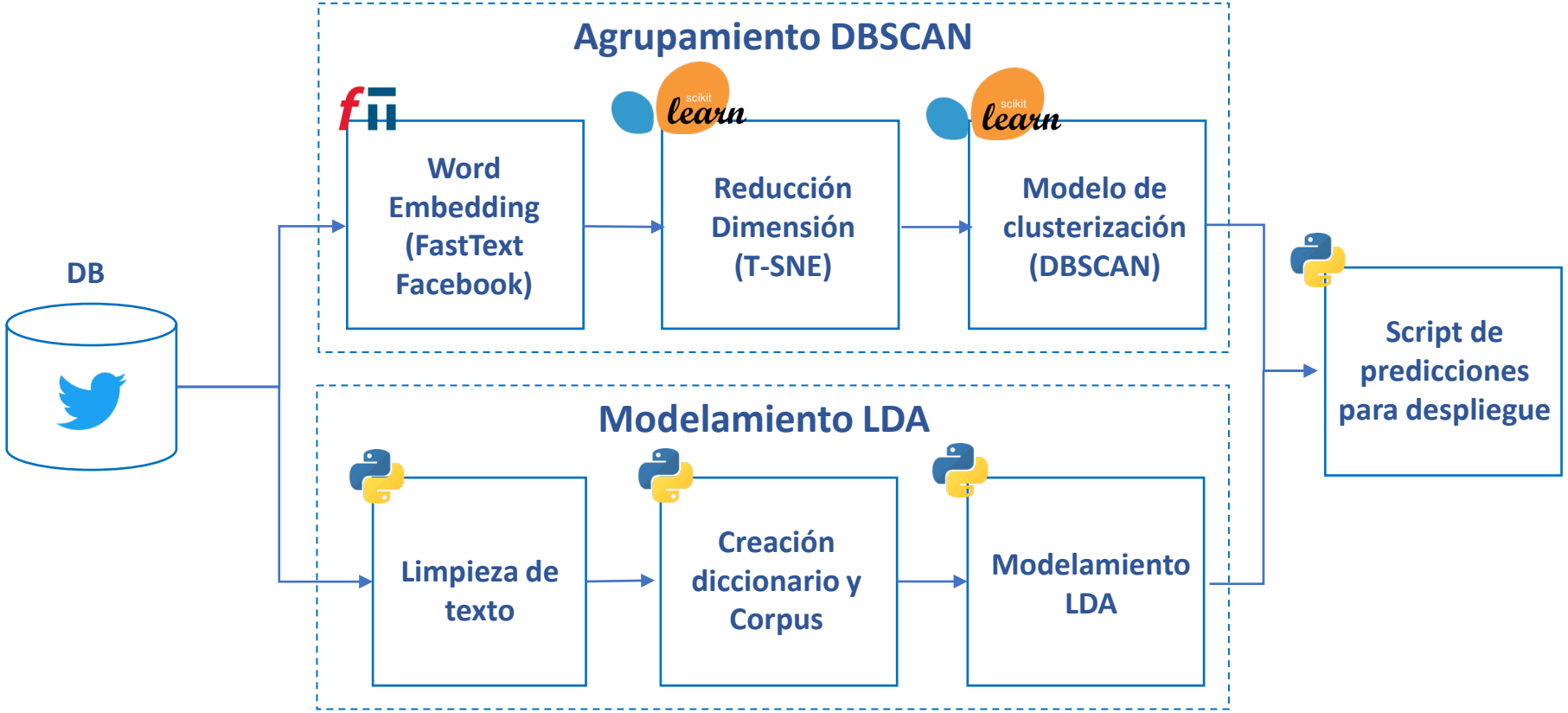
```
data[data['Likes']=='838']
```

	Unnamed: 0	UserScreenName	UserName	Timestamp	Text	Embedded_text	Emojis	Comments	Likes	Retweets	Image link
779	779	Maria Niny Echeverry	@Marianiniecheve	2021-12-15T19:01:12.000Z	Maria Niny Echeverry\n@Marianiniecheve\n·\n15 ...	Chicos: me ayudarían a denunciar a \n@SegurosB...	😡😡😡	65.0	838	1 mil	[]

Técnicas implementadas

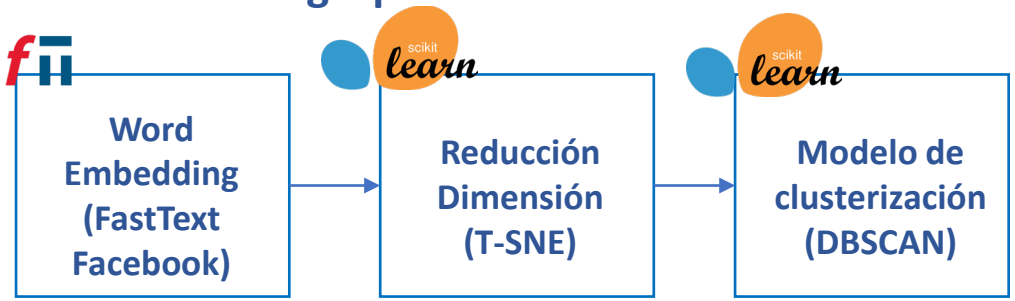
Comparación de dos técnicas:

1. Vectorización con Embeddings y clusterización
2. Modelamiento generativo con LDA



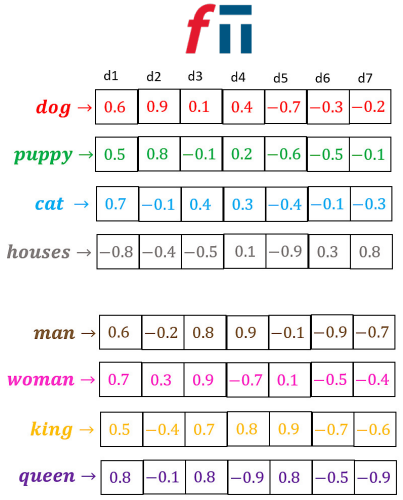
Técnicas implementadas

Agrupamiento DBSCAN



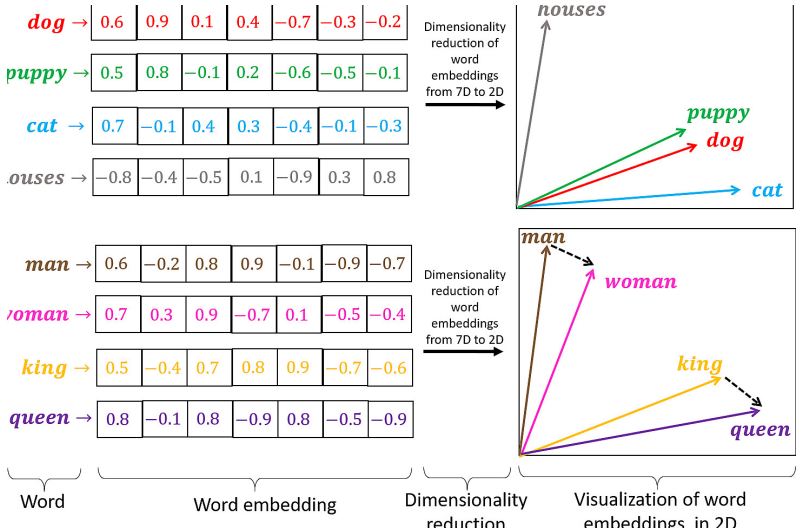
"Los modelos de ML & IA entienden números no palabras"

1. Vectorización con Embeddings



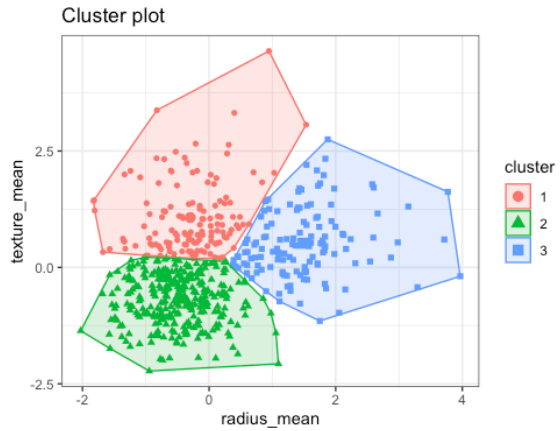
Modelo pre-entrenado al español por el equipo de investigación de Facebook. La salida es una matriz que relaciona palabras con palabras contexto obteniendo un espacio vectorial n-dimensional.

2. Reducción de la dimensionalidad



Utilizando técnicas de reducción de dimensionalidad se reduce el espacio vectorial obtenido del proceso de embedding para obtener resultados explicables y con sentido.

3. Clusterización o agrupamiento



Una vez obtenido un espacio entendible se realiza la clusterización de este espacio vectorial para obtener los clusters, temas, tópicos o dolores de los cuales se hablan en twitter, entrenando un modelo DBSCAN.

Técnicas implementadas

Modelamiento LDA

“Los modelos de ML & IA entienden números no palabras”

1. Limpieza de texto

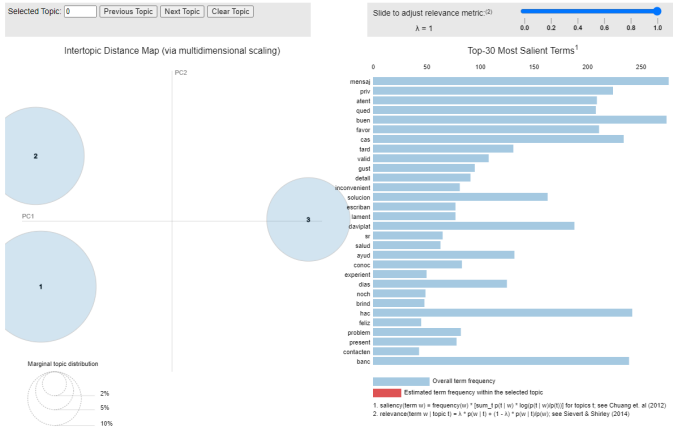
```python
def limpiado_de_texto_1(texto, remove_stop_words=True, stemming_words=True):

    # Eliminamos los caracteres especiales
    texto = re.sub(r'\W', ' ', str(texto))
    # Eliminado las palabras que tengo un solo caracter
    texto = re.sub(r'\s+[a-zA-Z]\s+', ' ', texto)
    # Sustituir los espacios en blanco en uno solo
    texto = re.sub(r'\s+', ' ', texto, flags=re.I)
    # remover numeros
    texto = re.sub(r'\b\d+(?:\.\d+)?\s+', '', texto)
    texto = re.sub('respuesta', '', texto)  # Se elimina la palabra respuest
    texto = re.sub('Respuesta', '', texto)
    texto = re.sub('1', '', texto)
    texto = re.sub('2', '', texto)
    # Convertimos textos a minusculas
    texto = texto.lower()

    # Tokenizado
    tokenizer = ToktokTokenizer()
    tokens = tokenizer.tokenize(texto)

    # Eliminacion de stopwords
    stop_words =  stopwords.words('spanish')
    if remove_stop_words:
        tokens = [w for w in tokens if not w in stop_words]

    # Stemming
    stemmer = SnowballStemmer("spanish")
    if stemming_words:
        tokens = [stemmer.stem(token) for token in tokens]

    text = " ".join(tokens)
```

Se desarrolla una función para realizar una limpieza del texto, eliminando caracteres especiales, números, espacios, tokenizando y eliminando stopwords y conectores del idioma español.

2. Creación de corpus y diccionario

- Diccionario: Codificación de cada palabra con un identificador.
- Corpus: A partir del diccionario creado, se crean “documentos” por cada tweet el cual contendrá una bolsa de palabras del diccionario anterior.

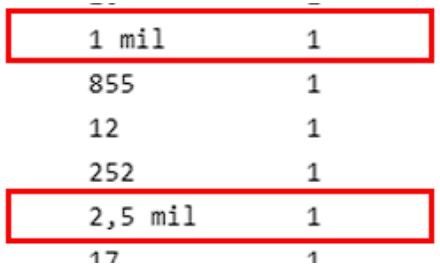
3. Modelamiento LDA



El algoritmo LDA es no supervizado y asigna a cada documento un valor dependiendo de los diferentes tópicos que logra discriminar. Al final se obtiene un vector de probabilidades por cada tópico.

Dificultades encontradas

- Para la ejecución del modelo pre-entrenado de FastText, es necesario tener un ambiente basado en arquitectura UNIX (Linux/MacOs) o en su defecto si se desea ejecutar en Windows, será necesario instalar el compilador de C++ debido a que el core de este modelo está desarrollado en C++.
- Para la extracción de Insights la utilización de variables como likes o retweets resulta relevante, sin embargo el valor numérico de algunas observaciones no era el deseado. Esto se puede evidenciar en cifras superiores a los mil, como por ejemplo:

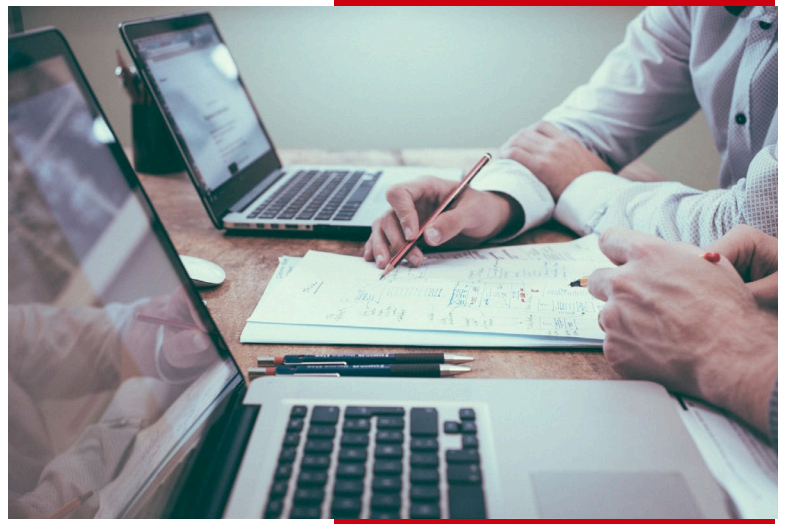


1 mil	1
855	1
12	1
252	1
2,5 mil	1
17	1

- Sin embargo, se logra hacer una limpieza de estos valores para poder ser convertidos a numéricos y obtener los insights requeridos.

Aspectos importantes

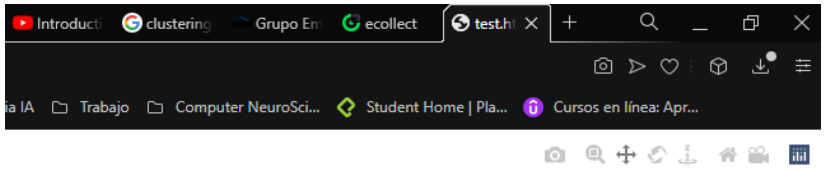
- Para la ejecución del modelo de FastText en español se debe realizar una limpieza de texto diferente al modelo LDA, razón por la cual se cuentan con dos funciones de limpieza.
- Ambos resultados tienen agrupamientos o segmentaciones de tweets significativos, sin embargo las técnicas difieren en sus resultados, seleccionando como última técnica el modelamiento LDA por los agrupamientos que genera y la distribución de los mismos.
- Puede considerarse este desarrollo como una PoC en donde se pueden encontrar bastas oportunidades de mejora sobre todo en el ámbito de modelamiento, implementando modelos más avanzados de aprendizaje profundo.



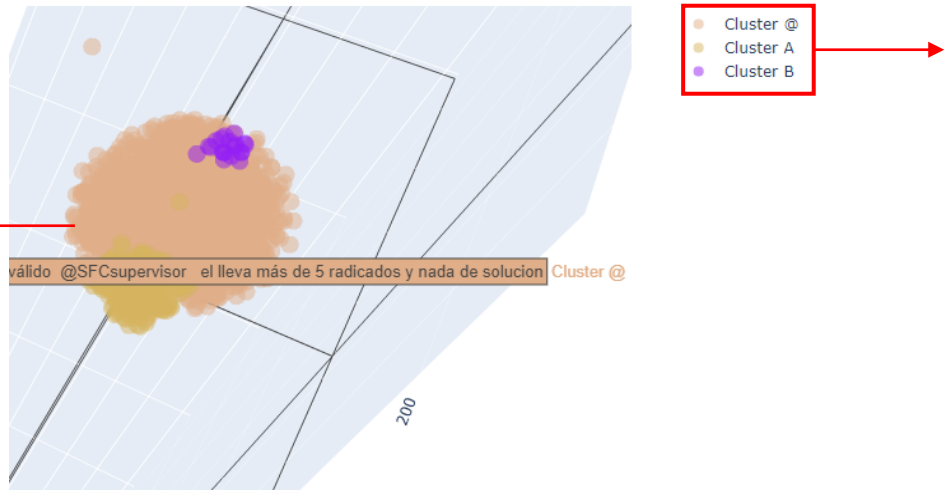
Resultados FastText - DBSCAN

- Como resultados de esta primera técnica se tiene una pequeña interfaz interactiva (‘test.html’) que se puede ejecutar en los navegadores más utilizados. En esta entrega se pueden observar la siguiente segmentación:

    1. Cluster A: Respuestas dadas por Davivienda
    2. Cluster B: Burlas o comentarios irónicos de los clientes con antiguo slogan de Davivienda
    3. Cluster @: Quejas y/o reclamos por parte de los clientes.



- Pasando el cursor del mouse por encima de algún punto se podrá leer el comentario perteneciente al cluster indicado por el color.



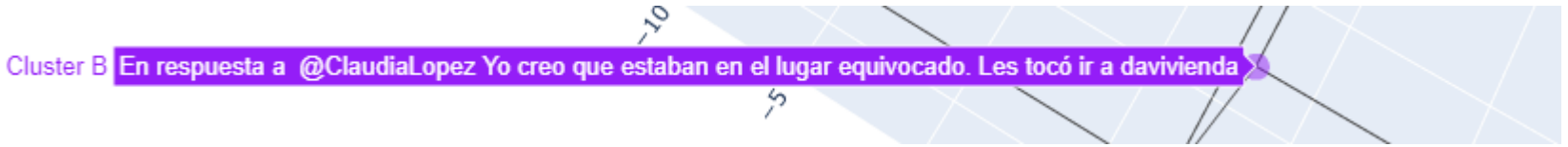
- Haciendo Click en cada cluster se puede activar o desactivar la visualización del mismo

Resultados FastText - DBSCAN

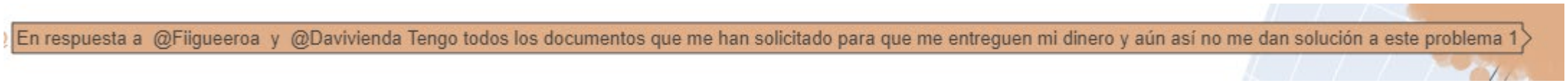
- Visualizando ejemplos de cada cluster:

1. Cluster A: Respuestas por Davivienda

2. Cluster B: Respuestas irónicas con el slogan “el lugar equivocado”



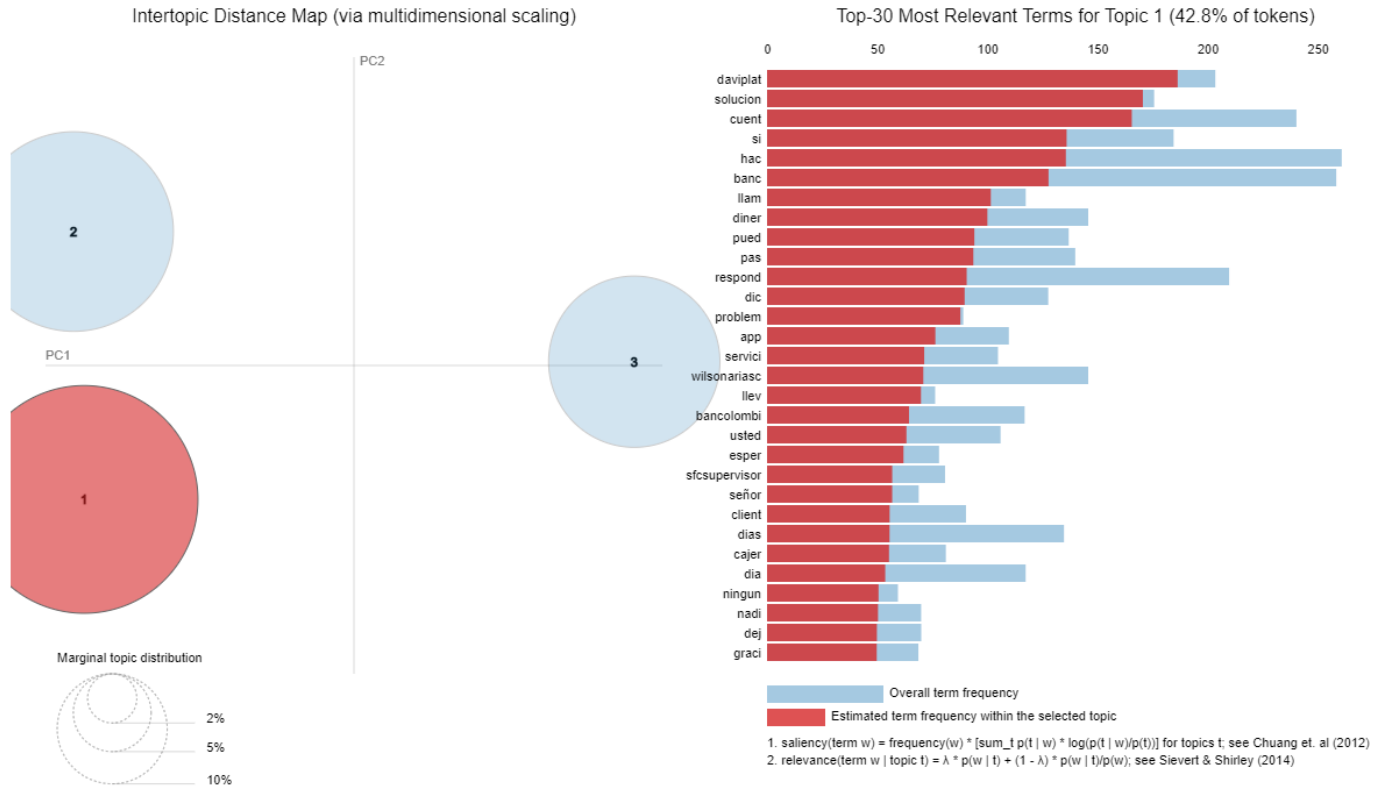
3. Cluster @: Quejas y/o reclamos por parte de los clientes



Resultados Modelado LDA

- Como resultados de esta segunda técnica se tiene una pequeña interfaz interactiva (‘lda.html’) que se puede ejecutar en los navegadores más utilizados. En esta entrega se pueden observar la siguiente segmentación:

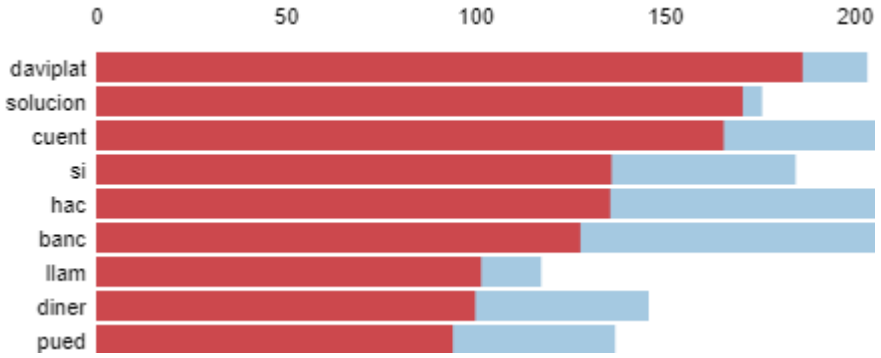
1. Cluster 1: Problemas relacionados con Daviplata
2. Cluster 2: Problemas y/ quejas relacionadas con tweets de alta influencia.
3. Cluster 3: Mensaje contestado por Davivienda.



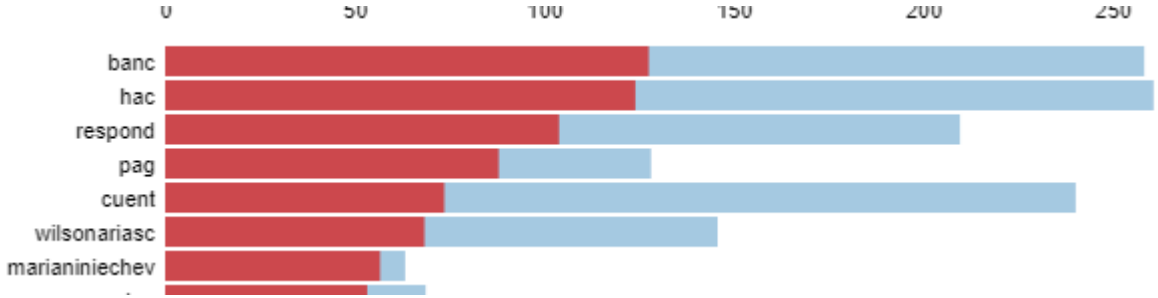
Resultados Modelado LDA

- Visualizando ejemplos de cada cluster y observando las palabras más relevantes:

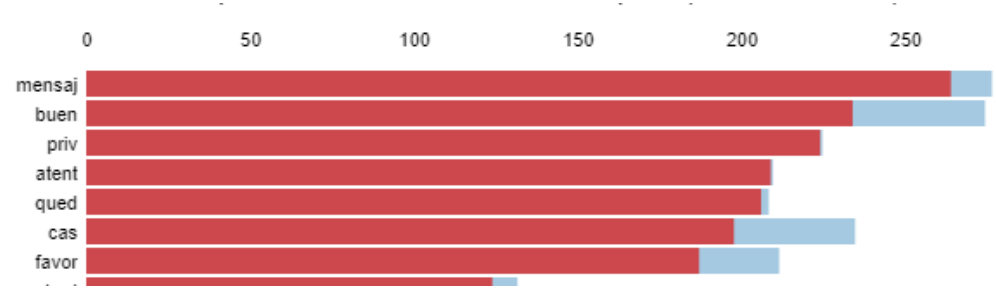
1. Cluster 1: Problemas relacionados con Daviplata



2. Cluster 2: Problemas y/ quejas relacionadas con tweets de alta influencia.



3. Cluster 3: Mensaje contestado por Davivienda.



Costos ejecución

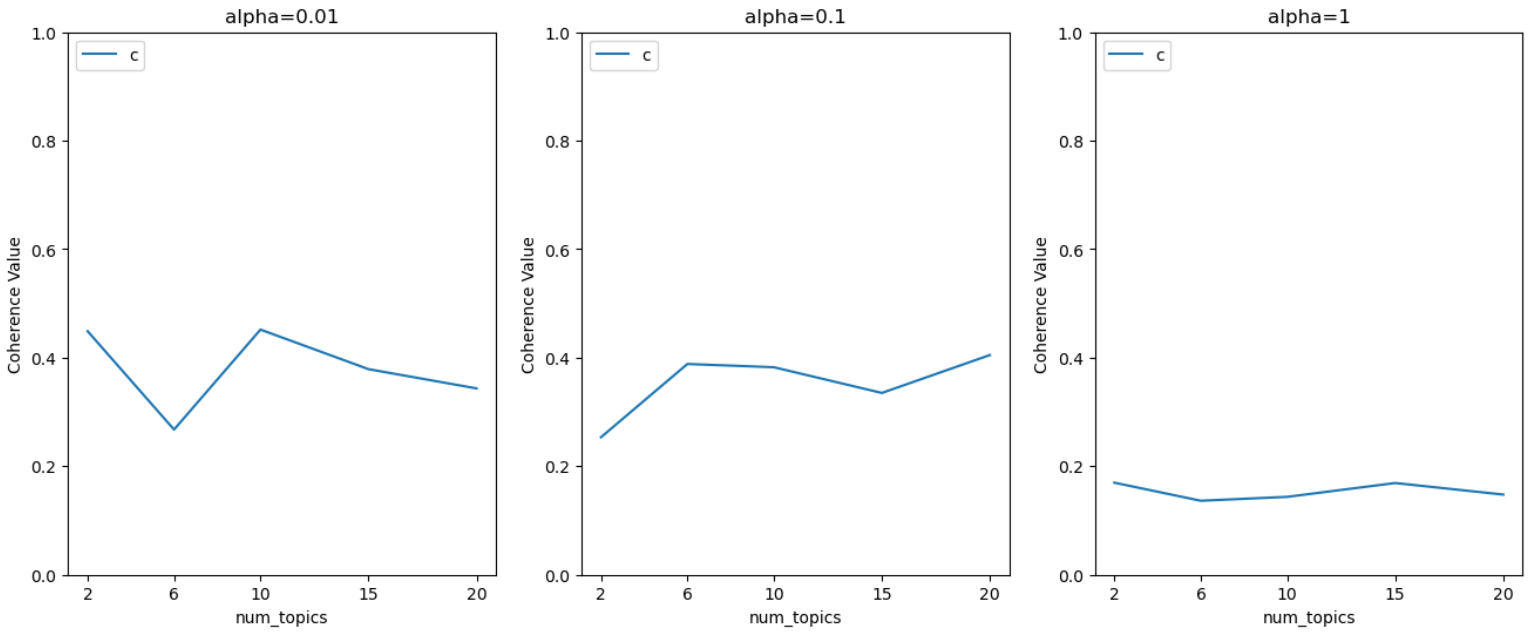
- Para la ejecución de ambos modelos se utiliza un equipo con 16 GB de RAM, procesador de 8 núcleos a 3,59 GHz tomando un tiempo de entrenamiento por cada modelo aproximado de medio minuto cada uno.

- No se utiliza ninguna API paga ni hardware en nube.

- Se utiliza FastText, modelo que corre nativamente en ambientes UNIX y pesa alrededor de 1,3 GB.

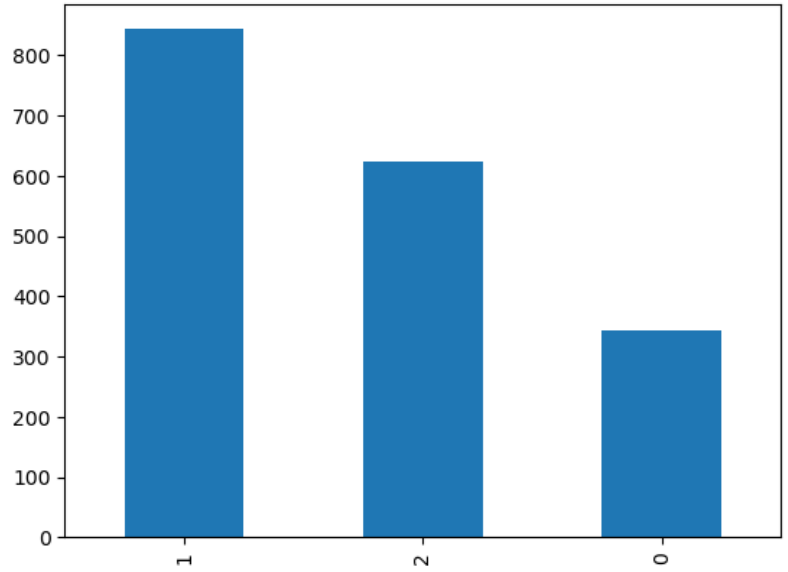
Detalles adicionales y conclusiones

- Se puede observar que ambos modelos discriminan bastante bien, sin embargo para el primer modelo se puede observar que la nube de puntos resultante parece ser poco distinguible entre clusters, razón por la cual se puede hacer más énfasis en la limpieza de texto para obtener mejores resultados.
- Para la selección de hiperparámetros del modelo LDA se realiza una verificación de la métrica conocida como “coherencia” obteniendo diferentes métricas variando sus hiperparámetros obteniendo los siguientes resultados, observando una mejoría con un Alpha de 0,01 y la cantidad de tópicos entre 2 y 3,



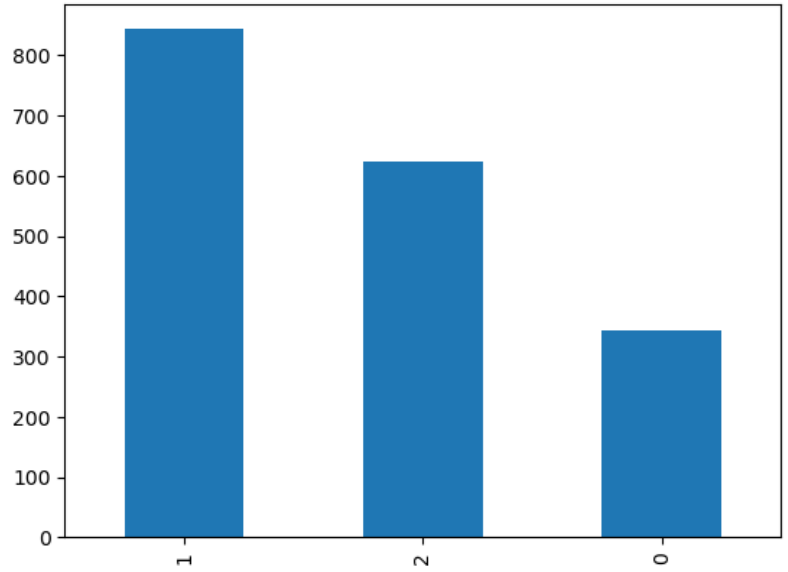
Detalles adicionales y conclusiones

- Al final del código se tienen dos funcionalidades, la primera consiste en obtener la predicción de cualquier dataset especificando la columna texto y el segundo corresponde a la generación de insights, que consiste en ver cuales son los tópicos que más se repiten y aquellos tópicos que generan interacciones (likes + retweets) superiores a 200.
- También se puede observar la distribución de tópicos para el dataset dado

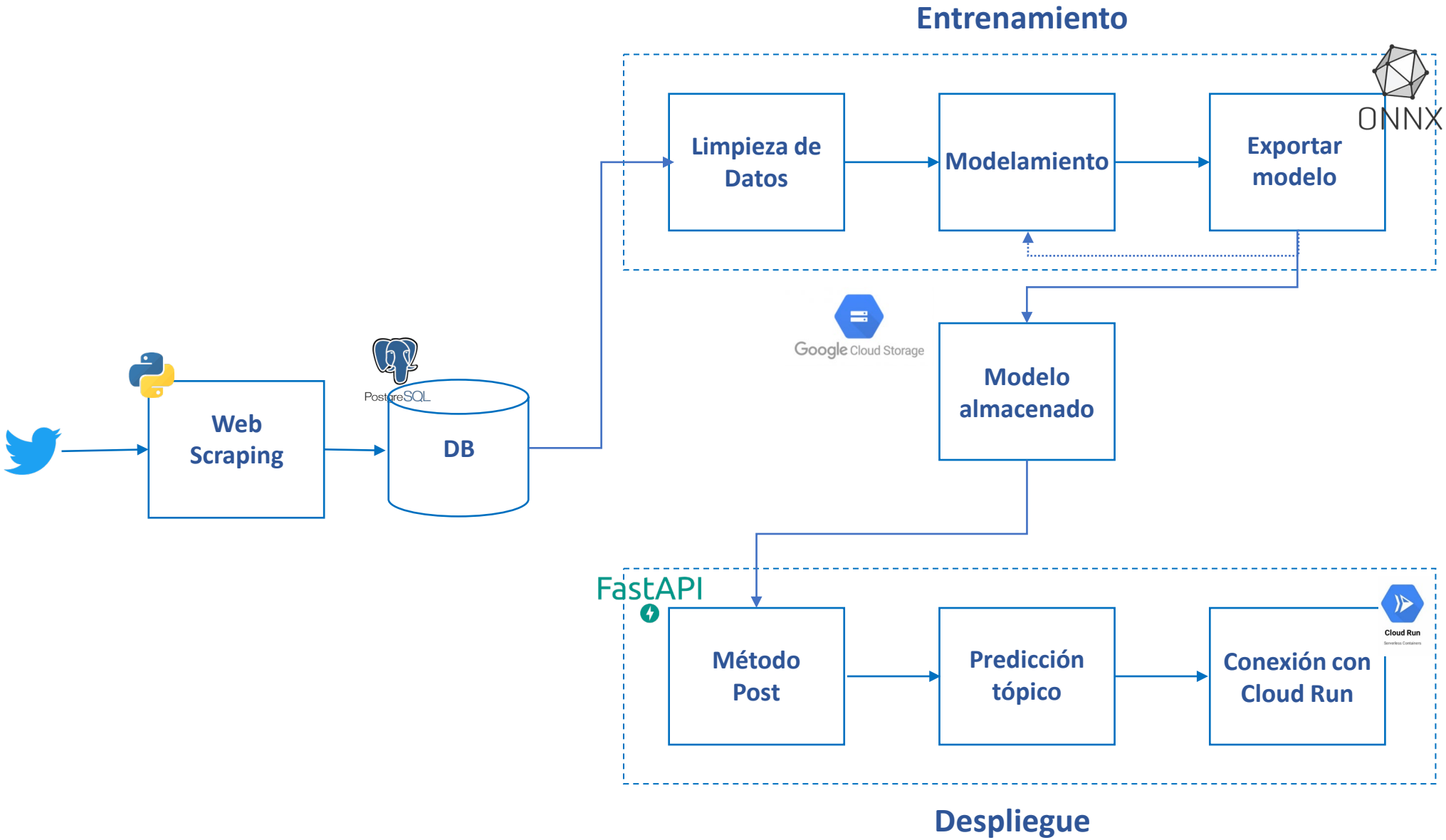


Detalles adicionales y conclusiones

- Al final del código se tienen dos funcionalidades, la primera consiste en obtener la predicción de cualquier dataset especificando la columna texto y el segundo corresponde a la generación de insights, que consiste en ver cuales son los tópicos que más se repiten y aquellos tópicos que generan interacciones (likes + retweets) superiores a 200.
- También se puede observar la distribución de tópicos para el dataset dado



Predecir tópico del tweet



Contestar tweets dependiendo el contexto