datascience.pe

Topic modelling sobre comentarios negativos de restaurantes usando NLP

Data Science

Data Science

Expositor

Diego Bernales Valdivia

https://github.com/dbernalesv

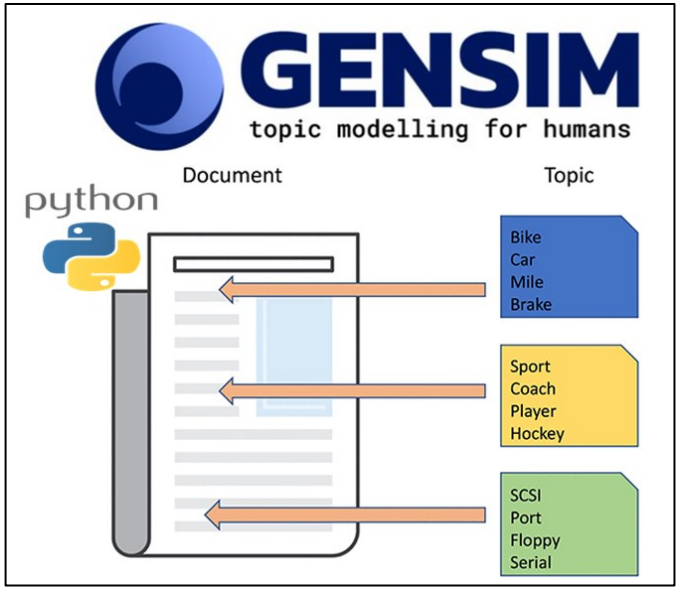
diegobernales3@gmail.com

https://www.linkedin.com/in/diego-bernales-valdivia/

RESUMEN

- **En el presente proyecto se propone presentar un modelo que nos permite encontrar los Tópicos más importantes en los reviews de quejas que reciben los restaurantes limeños.**

- **Al final del proyecto se obtendrá un dataset que contenga los Tópicos de quejas más relevantes para distintos restaurantes de Lima y sus respectivas interpretaciones. Además se realizará un pequeño análisis respecto a la cadena de restaurantes "Señor Limón".**



Data Science

Data Science

INTRODUCCIÓN

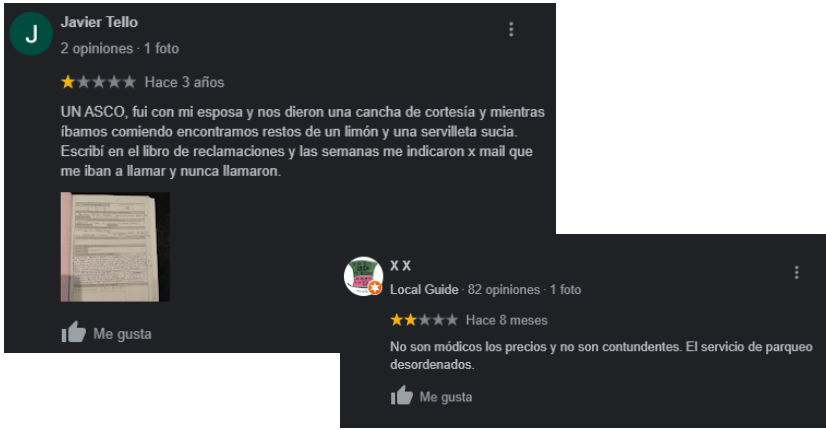
- **Problemática**

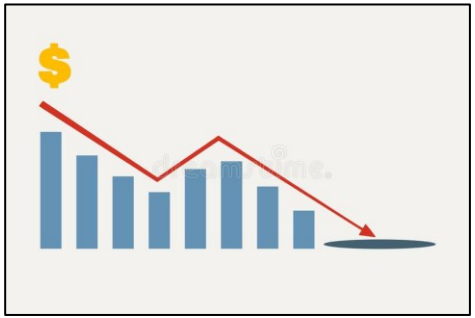
  Para los negocios del rubro de restaurantes es bastante importante la opinión de los clientes, ya que en esto se basa la popularidad del negocio, es decir, mientras se reciba una mayor cantidad de reseñas positivas más ventas y viceversa cuando se tienen muchas reseñas negativas.

  Esto significa que siempre se debe procurar destinar los recursos necesarios para mejorar en aquellos puntos débiles del negocio, por lo tanto resulta necesario analizar las quejas. Sin embargo, estas pueden ser muchas, largas o poco entendibles y agruparlas de manera eficiente por tema se convierte en un desafío.

- **Objetivo Principal**

  - DISEÑAR UN SISTEMA INTELIGENTE EFICIENTE PARA DETECTAR TÓPICOS EN REVIEWS NEGATIVAS



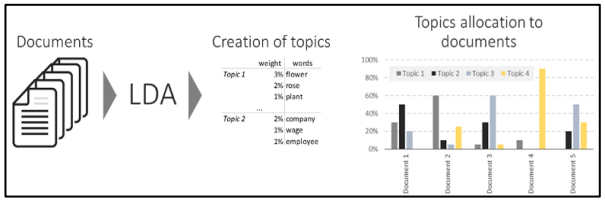


Data Science

Marco Teórico

Modelo LDA

- Es uno de los algoritmos generativos más usados y con mejor desempeño en cuanto a modelado de tópicos.
- Aprovecha de buena manera los datos que se poseen, sin necesidad de tener que recolectar grandes cantidades de información.
- Flexible



LDA Mallet

- Utiliza las bases del modelo LDA de Gensim.
- Usa una técnica de muestreo diferente que le permite ser más preciso.
- Incluye elementos aleatorios que hace que se obtengan resultados distintos en cada ejecución.

$$P(z, w|\alpha, \eta) = \int_{\theta} \int_{\beta} P(z, w, \theta, \beta|\alpha, \eta) d\beta d\theta$$

Coherencia

- Esta métrica mide si las palabras de un tema tienden a coexistir.
- Ayuda a saber qué tan bueno es el modelo entrenado.

$$\sum_{i} \sum_{j<i} \log \frac{D(w_j, w_i) + \beta}{D(w_i)}$$

Data Science

Marco Teórico

Pre-procesamiento

- Sirve para mejorar el rendimiento del modelo final.
- Es necesario para el funcionamiento adecuado de los algoritmos.
- Es un paso obligatorio en el análisis de texto.

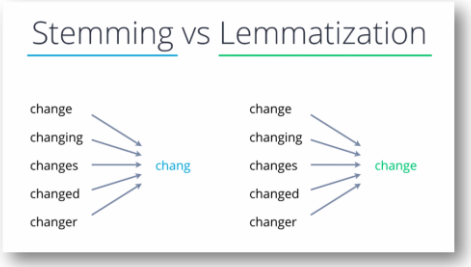
StopWords

- Son palabras que no aportan información relevante o muy poca sobre el texto que se esté analizando, estos pueden ser preposiciones, conectores, algunas acotaciones informales propias del lenguaje, etc.

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Lematización

- Derivar cada palabra a su forma canónica.
- Derivaciones tal y como se encuentra en el diccionario



Data Science

Marco Teórico

Corpus

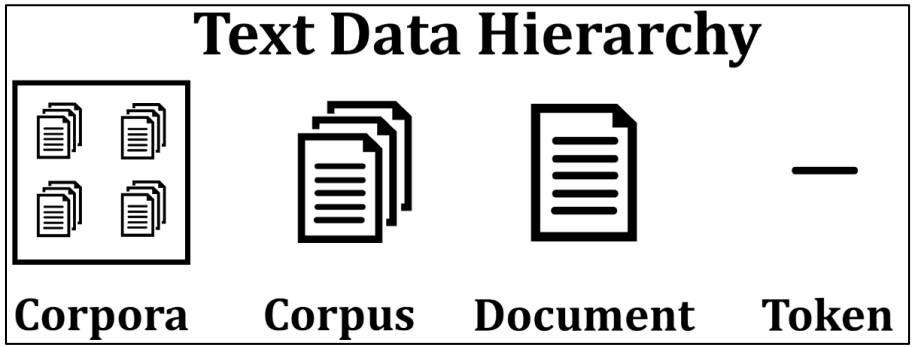
- El set de documentos, en este caso es el set de Reviews.

Diccionario

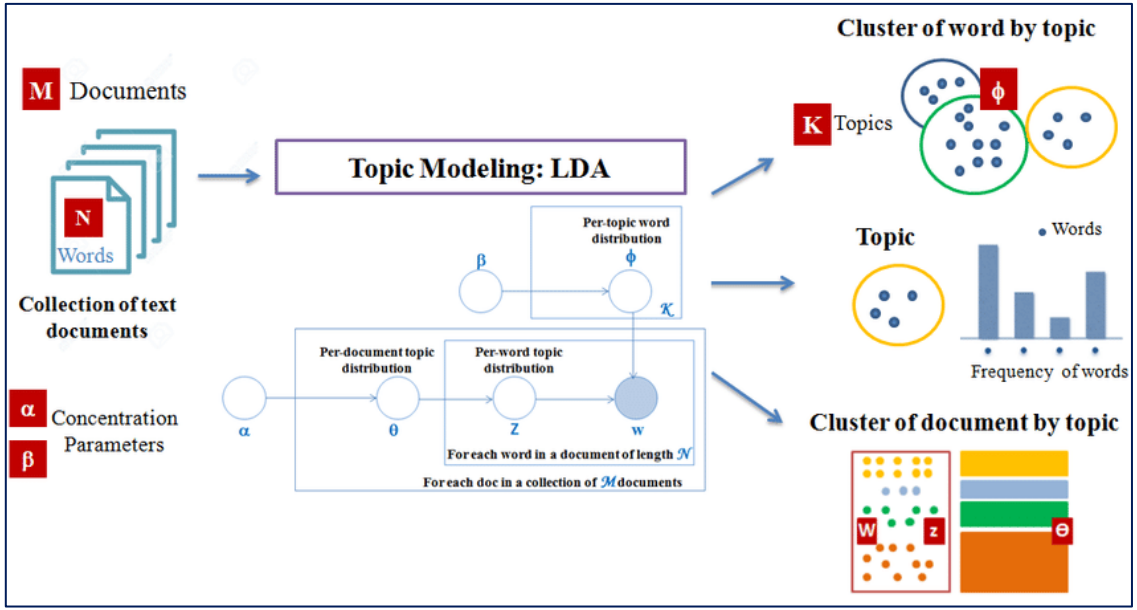
- lista de todas las palabras únicas que aparecen en el corpus.

Tokenización

- Es la manera de separar un fragmento de texto en unidades más pequeñas conocidas como tokens, las cuales son palabras o caracteres.



Entendiendo LDA



The LDA model has two parameters that control the distributions:

1. Alpha (α) controls per-document topic distribution, and
2. Beta (β) controls per topic word distribution

(c) Two outputs of LDA

(c-1) Per-document topic proportions (θ_d)

	Topic 1	Topic 2	Topic 3	...	Topic K
Doc 1	0.20	0.50	0.10	...	0.10
Doc 2	0.50	0.02	0.01	...	0.40
Doc 3	0.05	0.12	0.48	...	0.15
...	...	...	...	...	...
Doc N	0.14	0.25	0.33	...	0.14

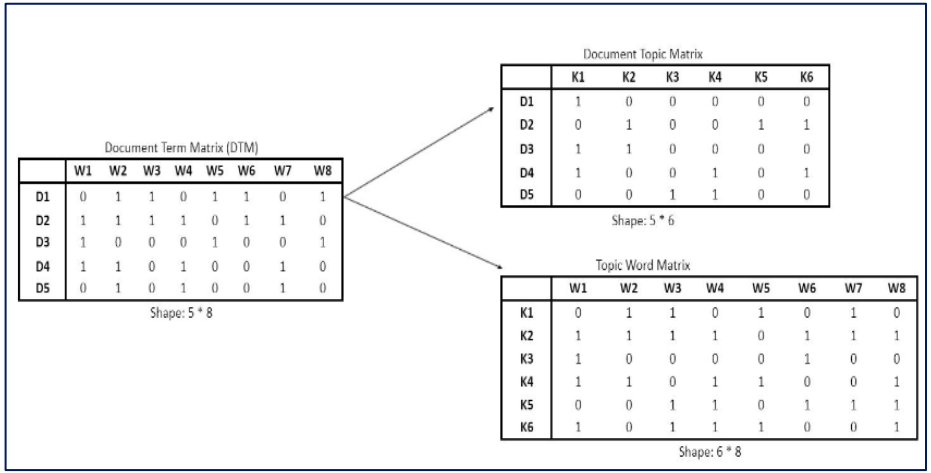
(c-2) Per-topic word distributions (ϕ_k)

	Topic 1	Topic 2	Topic 3	...	Topic K
word 1	0.01	0.05	0.05	...	0.10
word 2	0.02	0.02	0.01	...	0.03
word 3	0.05	0.12	0.08	...	0.02
...	...	...	...	...	...
word N	0.04	0.01	0.03	...	0.07

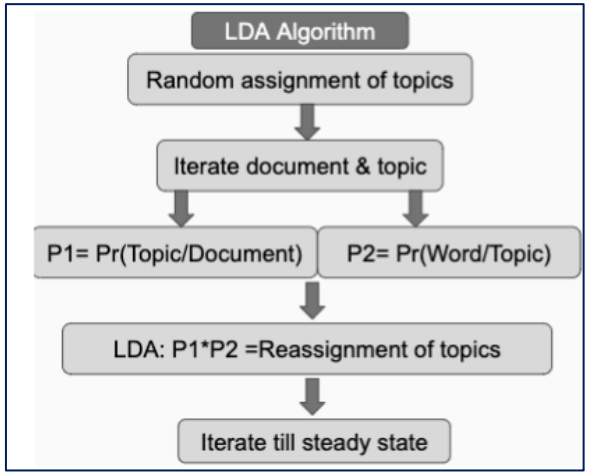
Data Science

Entendiendo LDA

Proceso iterativo



LDA asume que los documentos son una mezcla de temas y los temas son una mezcla de palabras.



LDA hace otra suposición: todos los temas que se han asignado son correctos excepto la palabra actual. Entonces, en base a esas asignaciones de palabras-tema ya correctas, LDA intenta corregir y ajustar la asignación de tema de la palabra actual con una nueva asignación para lo cual LDA iterará sobre cada documento 'D' y cada palabra 'w'.

Usando estas probabilidades p1 y p2, LDA estima una nueva probabilidad, que es el producto de (p1*p2), y a través de esta probabilidad producto, LDA identifica el nuevo tema, que es el tema más relevante para la palabra actual.

Data Science

Metodología {

FASE 1

SELECCIÓN DE DATOS:

- Recopilar data de reviews de restaurantes.

FASE 2

PRE-PROCESAMIENTO DE REVIEWS:

- Retirar caracteres extraños de la data, es decir, eliminar el "ruido"
- Tokenizar y Remover Stop Words.

FASE 3

SELECCIONAR SCORES NEGATIVOS Y ANALIZAR FRECUENCIAS DE PALABRAS

PROCESADO FINAL DE REVIEWS

- Lematizar
- Colocar los tokens en listas por cada review

FASE 4

CONSTRUCCIÓN Y EVALUACION DE MODELOS.

- Creación de modelos LDA y LDA mallet.
- Evaluación y entendimiento de tópicos hallados.

FASE 5

Unión de Tópicos hallados por review y restaurante correspondiente.

- Analizar Tópicos relevantes por restaurante.

}

SELECCIÓN DE DATOS

- **El dataset utilizado en el presente proyecto es recolectado de Kaggle Contiene información extraída de GooglePlaces y Tripadvisor utilizando Selenium, Requests, BeautifulSoup y Rvest.**

	id_review	review	title	score
0	R1245	Muy buena presentación y servicio sin embargo ...	Muy buena presentación y servicio	3.0
1	R1246	Desde la presentación de los platos a la calid...	Una experiencia	5.0
2	R1247	El mejor lugar para reencontrame con mis amigo...	Felicitaciones a Statera!	5.0
3	R1248	Excelente experiencia Comida maravillosa con e...	Aniversario de boda	5.0
4	R1249	Mi pasión es viajar y disfrutar de las grandez...	No te lo querrás perder!	5.0

Data Science

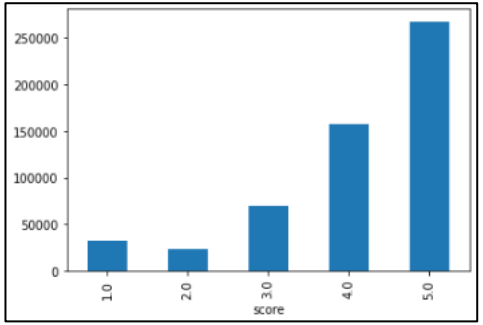
Data Science

Pre-Procesamiento de datos

```
random.seed(42)

spanish_stopwords = stopwords.words('spanish')
all_stopwords=list(set(spanish_stopwords+list(STOP_WORDS)+['q','pq','xd','xq']))

wlm = WordNetLemmatizer()
```

```
def preprocesamiento_reviews(text):
  text = text.lower()
  #text = normalize('NFKD', text)

  text = re.sub('[áàä]', 'a', text)
  text = re.sub('[éèë]', 'e', text)
  text = re.sub('[íìï]', 'i', text)
  text = re.sub('[óòö]', 'o', text)
  text = re.sub('[úùü]', 'u', text)

  text = re.sub('[.,:;]', '', text)
  text = re.sub(r'[`´"¿?!¡\)\(%\\\/''\[\]|]', '', text)
  text = re.sub(r'\d{1,}', '', text)
  text = re.sub(r'[---]', ' ', text)
  text = re.sub('```', '', text)
  text = re.sub(r' {2,}', ' ',text)

  return text

def procesamiento_reviews(texto_original, stopwords=all_stopwords):
  tokens = word_tokenize(texto_original)
  texto_limpio = ' '.join([t for t in tokens if t not in stopwords])

  return texto_limpio

def text_preprocesor(sentence: str):
    sentence = sentence.translate(str.maketrans('', '', string.punctuation))
    tokens = [wlm.lemmatize(word) for word in nltk.word_tokenize(sentence.lower()) if (word not in all_stopwords) and re.search('[a-zA-Z]', word)]

    return tokens
```

Seleccionar Reviews negativos, observar frecuencias



Data Science

Procesamiento final de datos

```
def text_preprocesor(sentence: str):
    sentence = sentence.translate(str.maketrans('', '', string.punctuation))
    tokens = [wlm.lemmatize(word) for word in nltk.word_tokenize(sentence.lower()) if (word not in all_stopwords) and re.search('[a-zA-Z]', word)]

    return tokens
```

```
[ ] doc_list = list(df_reviews_negativos.review_final.apply(text_preprocesor))

[ ] doc_list[:2]

[ ] df_reviews_negativos['review_final_token']=doc_list

[ ] df_reviews_negativos.head()
```

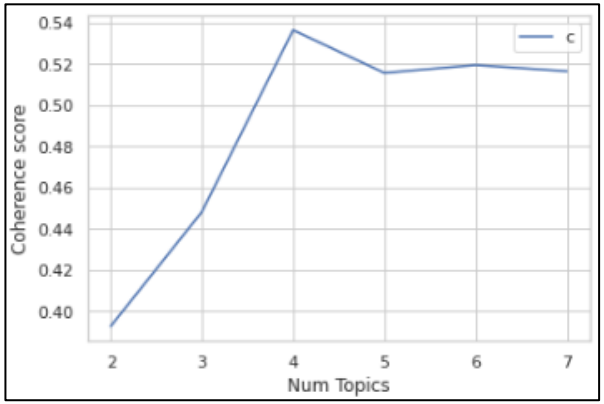
	id_review	review	title	score	likes	id_nick	service	date	platform	review_prep	review_final	review_final_token
215	R3005	Los acompañamientos muy malos. Los cortes de c...	Una mala experiencia	1.0	0	87carlosi	35956.0	1 years ago	tripadvisor	los acompañamientos muy malos los cortes de ca...	acompañamientos malos cortes carne vienen pone...	[acompañamientos, malos, cortes, carne, vienen...
225	R3015	Trabajan dos versiones diferentes de la misma ...	Facturan de más sin avisar.	1.0	0	ervas10	35956.0	1 years ago	tripadvisor	trabajan dos versiones diferentes de la misma ...	trabajan versiones etiqueta cepa vino caso tra...	[trabajan, versiones, etiqueta, cepa, vino, ca...
247	R3037	Llegue a este restaurante por recomendación de...	Un restaurante de Buda y despedida	1.0	0	osé M	35956.0	2 years ago	tripadvisor	llegue a este restaurante por recomendacion de...	llegue restaurante recomendacion youtuber loca...	[llegue, restaurante, recomendacion, youtuber,...
268	R3058	Los 330 soles (US$100) peor invertidos. Fui co...	Malísimo !	1.0	0	abiioooo	35956.0	2 years ago	tripadvisor	los soles us$ peor invertidos fui con mi hija ...	soles us $ invertidos hija cabrera miraflores ...	[sol, u, invertidos, hija, cabrera, miraflores...

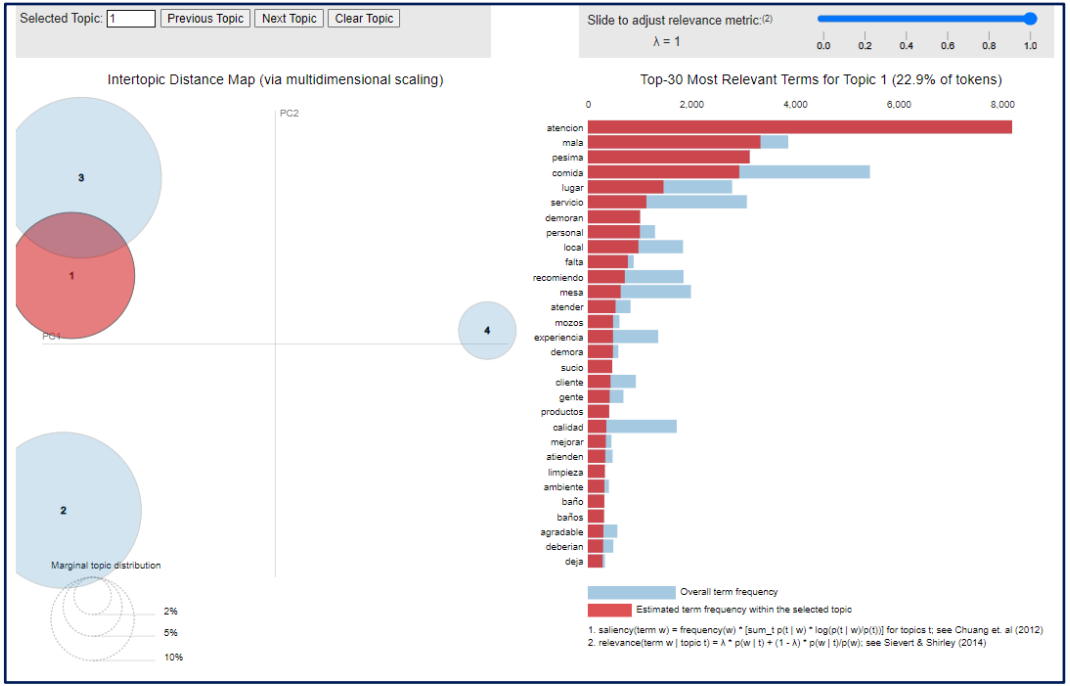
Data Science

Construcción y evaluación de modelos: Gensim LDA

```
dictionary = corpora.Dictionary(df_reviews_negativos['review_final_token'])
corpus = [dictionary.doc2bow(text) for text in df_reviews_negativos['review_final_token']]
```

```
def compute_coherence_values(dictionary, corpus, texts, limit, start=2, step=1):

  coherence_values = []
  model_list = []
  for num_topics in range(start, limit, step):
      model= gensim.models.ldamodel.LdaModel(corpus, num_topics = num_topics, id2word=dictionary, random_state=100, passes=15)
      model_list.append(model)
      coherencemodel = CoherenceModel(model=model, texts=texts, dictionary=dictionary, coherence='c_v')
      coherence_values.append(coherencemodel.get_coherence())

  return model_list, coherence_values
```

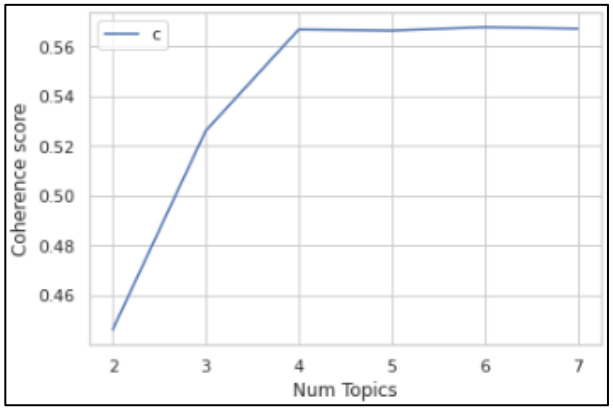


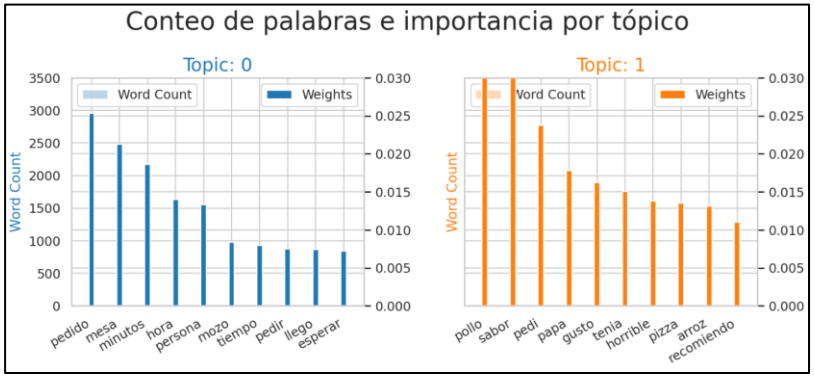


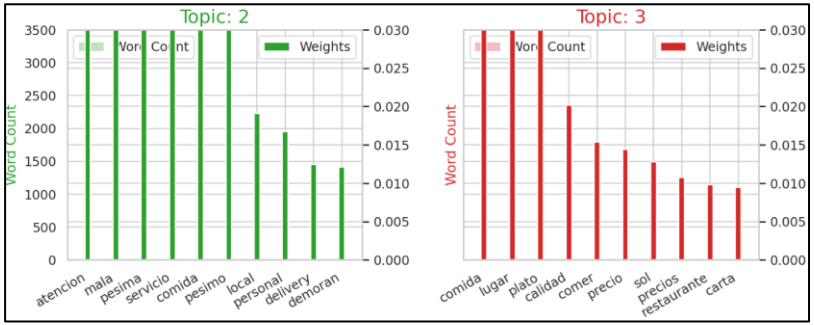
Construcción y evaluación de modelos:

LDA Mallet

```
import os.path
os.environ['MALLET_HOME'] = '/content/mallet-2.0.8'
mallet_path='/content/mallet-2.0.8/bin/mallet'
ldamallet = gensim.models.wrappers.LdaMallet(mallet_path, corpus=corpus, num_topics=4, id2word=dictionary, workers=8)
```







Data Science

Tópicos

Inconformidad con el tiempo de llegada del pedido.

0

1

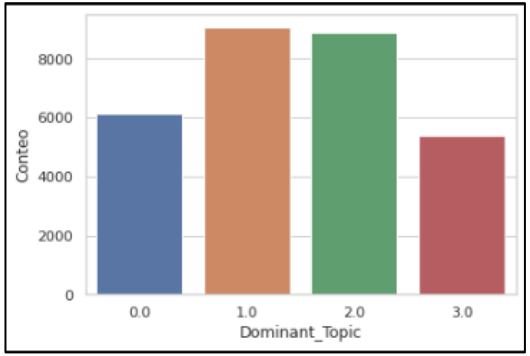
Mal sabor de la comida, platillos de mal gusto.

Atención mala y/o servicio pésimo.

2

3

Lugar o forma de servir la comida poco agradable, mala relación calidad precio.



Data Science

¿Qué hacemos con los tópicos hallados?

review_final_token	Topic Distribution	Dominant_Topic
['pesimo', 'servicio', 'franquicia', 'deberia', 'ordenar', 'emparejen', 'plato', 'verguenza', 'sirve', 'leche', 'tigre', 'vaso', 'onz', 'plastico', 'plato', 'quiñados', 'comida', 'presentada', 'forma', 'terrible']	[[0. 0.221] [1. 0.239] [2. 0.221] [3. 0.318]]	3.0
['arroz', 'mariscos', 'parecia', 'mazamorra', 'sabor', 'tenia', 'color', 'rojizo', 'ceviche', 'casa', 'malisimo', 'recomiendo']	[[0. 0.218] [1. 0.375] [2. 0.203] [3. 0.203]]	1.0
['mala', 'atencion', 'plato', 'llegaron', 'hora', 'retraso']	[[0. 0.271] [1. 0.225] [2. 0.265] [3. 0.239]]	0.0

Data Science

ANÁLISIS DE TÓPICOS PARA SEÑOR LIMÓN

EJEMPLOS DE REVIEWS Y TÓPICOS PRINCIPALES

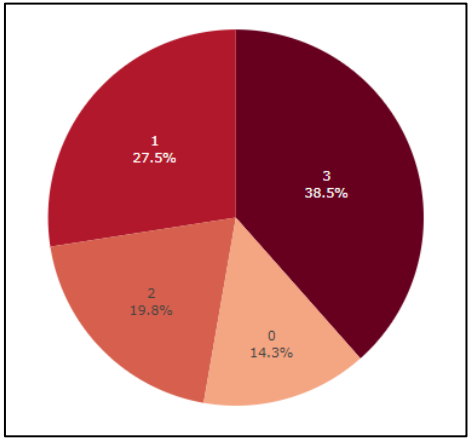
review	Dominant_Topic	district
Fui el día Domingo. El servicio fue muy lento y la comida de baja calidad. Las bebidas y los platos muy caros. Definitivamente no hay relación precio calidad y no vuelvo mas.	3.0	SAN MIGUEL
Buena atención , pésima comida. Realmente una pena que uno venga con ganas de disfrutar de algo agradable y que se vaya tan decepcionado por tan mala comida servida. Señor limón nunca más.	3.0	SAN MIGUEL
Pésimo el servicio, la franquicia debería ordenar que emparejen los platos, aquí una vergüenza donde se sirve la leche de tigre, vaso de 8 onz plástico, platos quiñados, y la comida presentada de forma terrible.	3.0	SAN ISIDRO
El arroz con mariscos parecía una mazamorra sin sabor tenía un color rojizo, el ceviche de la casa malísimo. No lo recomiendo	1.0	SAN ISIDRO
Me decepcionó el camote del ceviche estaba muy pero muy dulce Pero la atención y lo demás estaba excelente esa sería mi única queja	1.0	SAN ISIDRO

TÓPICO DE QUEJA PRINCIPAL POR SEDE

```
df_srLimon.groupby(['district'])['Dominant_Topic'].agg(pd.Series.mode)
```

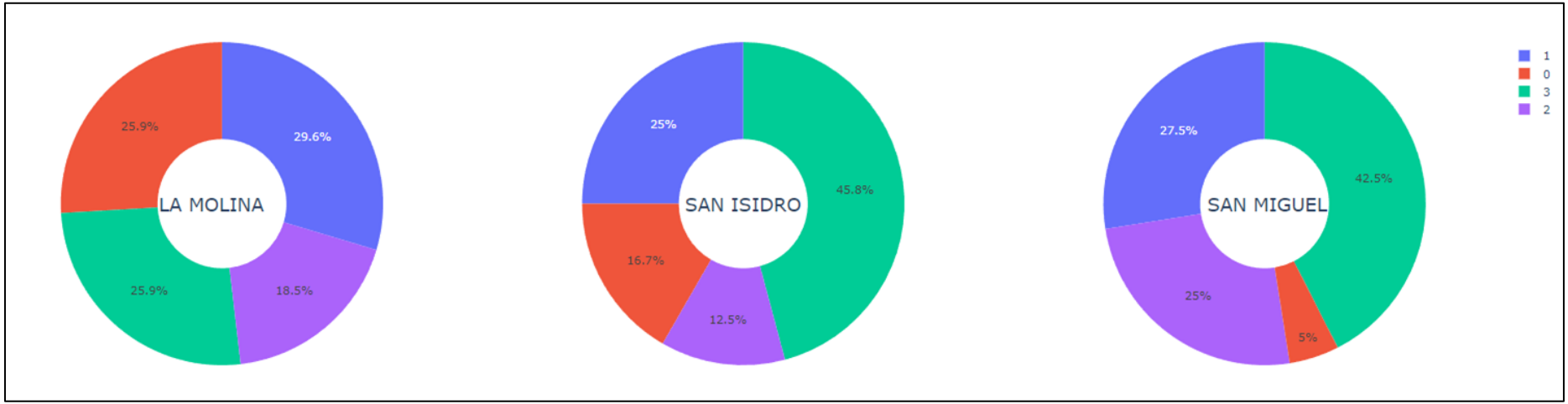
```
district
LA MOLINA     1.0
SAN ISIDRO    3.0
SAN MIGUEL    3.0
Name: Dominant_Topic, dtype: float64
```

PROPORCIÓN DE QUEJAS A NIVEL GENERAL PARA LA CADENA DE RESTAURANTES



ANÁLISIS DE TÓPICOS PARA SEÑOR LIMÓN

Proporción de tópicos de quejas por sede



GRACIAS

www.datascience.pe