

Prueba Técnica Profesional
III Departamento de Datos
no Estructurados:
Modelado de Tópicos con
Tweets.

Manuel Alejandro Diaz Rubiano.

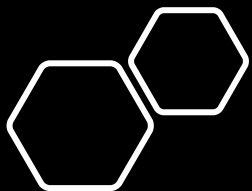
Cel: 3196371560

Email:

alejandromadr@gmail.com



DAVIVIENDA



Entendimiento del Negocio.

Según Gartner, en este momento, los datos no estructurados conforman el 80-90% del total de los datos que manejan las compañías. Dentro de ellos, la información que está contenida en texto desestructurado (correos, decisiones judiciales, evaluaciones de clientes, comentarios en redes sociales, etc) representa un enorme potencial para desarrollar analítica dentro de las empresas.

Un objetivo evidente de procesamiento de texto no estructurado se encuentra en la analítica de opiniones. Mediante ella, podemos obtener información de primera mano de lo que piensan nuestros clientes, que no esté mediada por encuestas u otros elementos logísticos que conformaban en un pasado reciente la única forma de tener información estructurada acerca de lo que opinan.

Objetivo.

El objetivo es desarrollar una analítica preliminar sobre un conjunto de tweets en los que se menciona de una u otra forma a nuestra organización. Mediante herramientas de procesamiento de lenguaje natural (NLP) buscamos que desarrollen el procesamiento de los textos contenidos en los tweets, que encuentren insights que puedan ser de interés para áreas como marketing, servicio al cliente, etc; y que visualicen estos datos de una forma clara para tomadores de decisiones. Además, en la medida de lo posible, buscamos una agrupación de estos tweets, en las categorías que ustedes consideren las mejores para hacer una buena analítica descriptiva de lo que opinan nuestros clientes en Twitter.

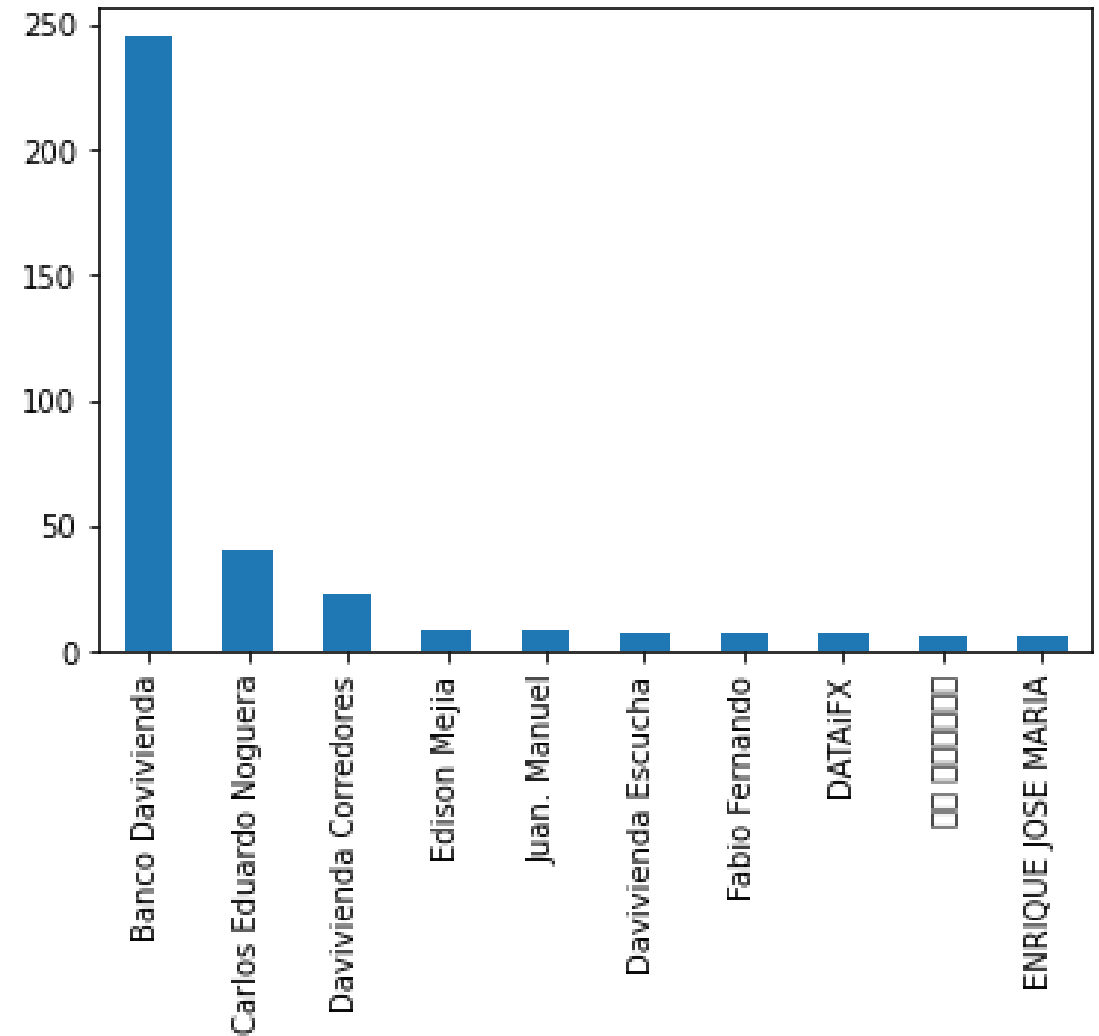
La base de Datos.

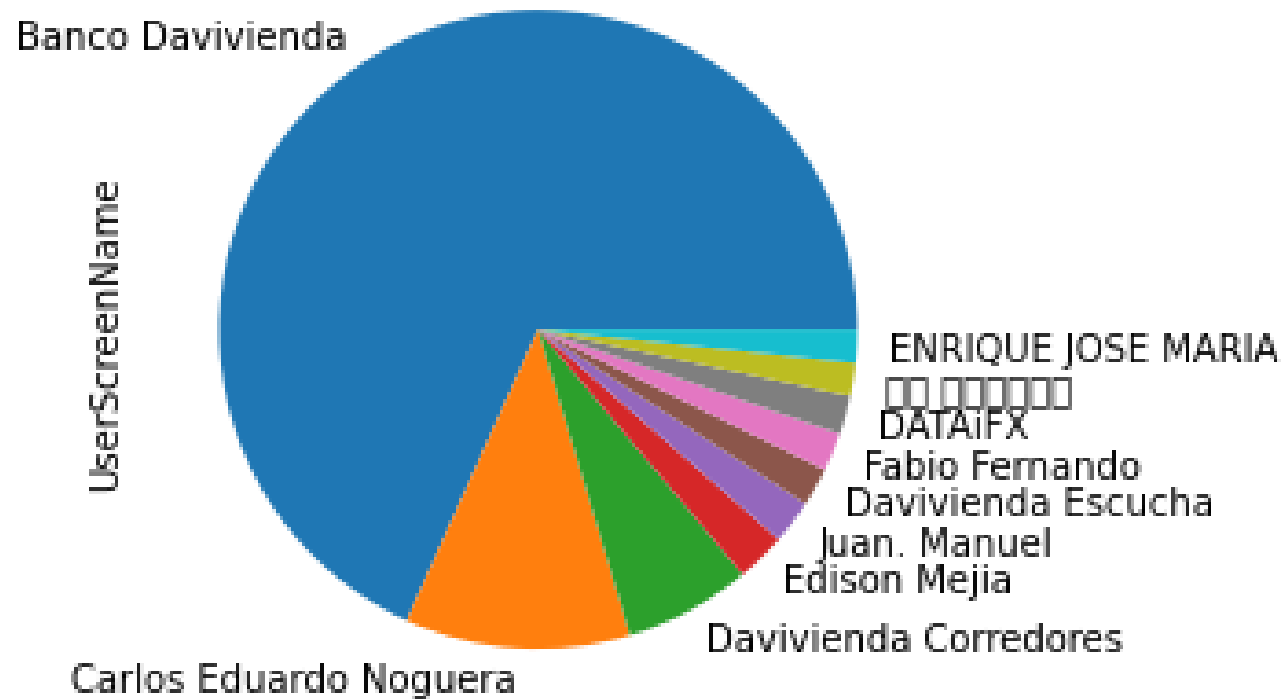
- Se tiene una base de datos, en la cual se tiene comentarios formulados en la red social de Twitter.
- La base de Datos tiene 1811 filas, y 12 columnas.

UserScreenName	UserName	Timestamp	Text	Embedded_text	Emojis	Comments	Likes	Retweets
Andrés Langebaek	@ALangebaek	2021-12-01T20:43:12.000Z	Langebaek\n@ALangebaek\n\n1 dic.	Andrés 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	Martinez\n@JulianM998\n\n1 dic.	Julián Señores \n@Davivienda\nno 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

Exploración de Datos.

Se puede observar que la mayoría de tweets que se tienen, corresponden al usuario de “Banco Davivienda”, por lo cual sería obvio que la palabra con mas frecuencia que apareciera, correspondiera al nombre del banco.





Mas del 13% de los tweets que se tienen en la base de datos, corresponde al usuario de Banco Davivienda, que tiene una representación bastante grande, a comparación de los otros usuarios.

Dificultades.

Existen varios caracteres especiales.

Demasiados quiebres en la continuidad del texto y de líneas.

Signos de puntuación.

Palabras mal escritas.

Técnicas utilizadas para la preparación de datos.

1. Primero se realiza un proceso de limpieza, en la cual se remueven de los tweets cosas como números, caracteres especiales, emoticones puntos, comas y demás, y luego pasar todas las palabras a forma minúscula.
2. Posteriormente se realiza el proceso de tokenización, el cual consiste en separar palabra por palabra cada tweet, en una lista para cada tweet.
3. Después de esto, se remueven las stopwords (palabras no relevantes como “el”, “las”, “con”); además de estas palabras StopWords en español, se pueden agregar más palabras que no se deseen observar.
4. Por último, se realiza el proceso de lematización, para evitar la redundancia de palabras, y su significado.



	list_untoken	list_tokenize	word_lemma	at	hashtags2
0	la confianza se afectó el indicador de confian...	[confianza, afectó, indicador, confianza, davi...	[confianza*, afectó*, indicador*, confianza*, ...	[]	[]
1	buscamos la accesibilidad y mejor atención en ...	[buscamos, accesibilidad, mejor, atención, trá...	[buscamos*, accesibilidad*, mejor*, atención*, ...	[]	[]
2	señores davivienda no he podido ingresar a mi ...	[señores, davivienda, podido, ingresar, app, d...	[señores*, davivienda*, podido*, ingresar*, ap...	[@Davivienda]	[]
3	llevo toda una semana sufriendo intento de hur...	[llevo, toda, semana, sufriendo, intento, hurt...	[llevo*, toda*, semana*, sufriendo*, intento*, ...	[@Davivienda, @Davivienda]	[]
4	hemos retrocedido tanto en este país con este ...	[retrocedido, país, gobierno, malparidos, caje...	[retrocedido*, país*, gobierno*, malparidos*, ...	[@Davivienda]	[]

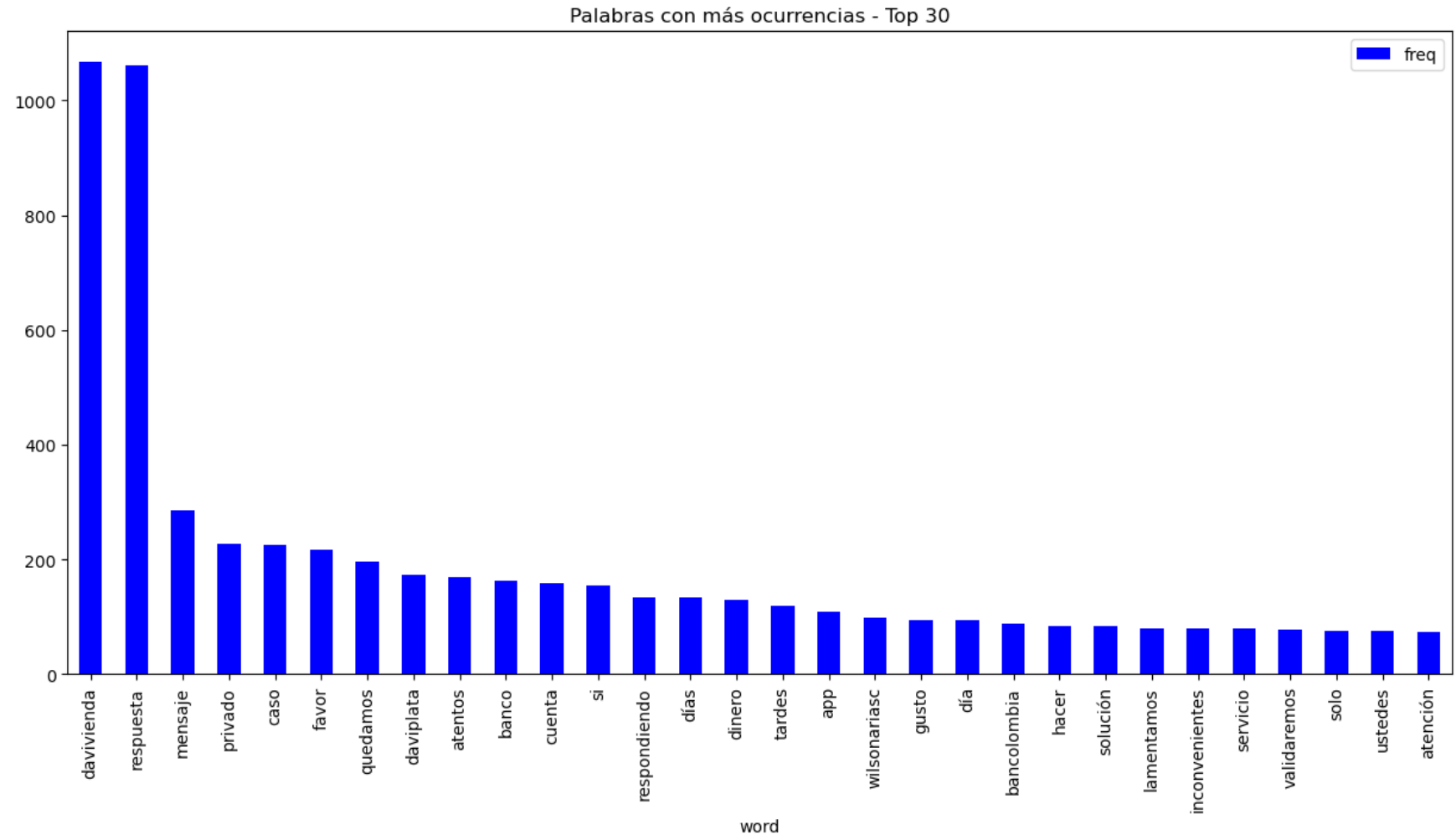
La base de datos resultante después de aplicar el pre procesamiento, luce de la siguiente manera:



La base resultante contiene:

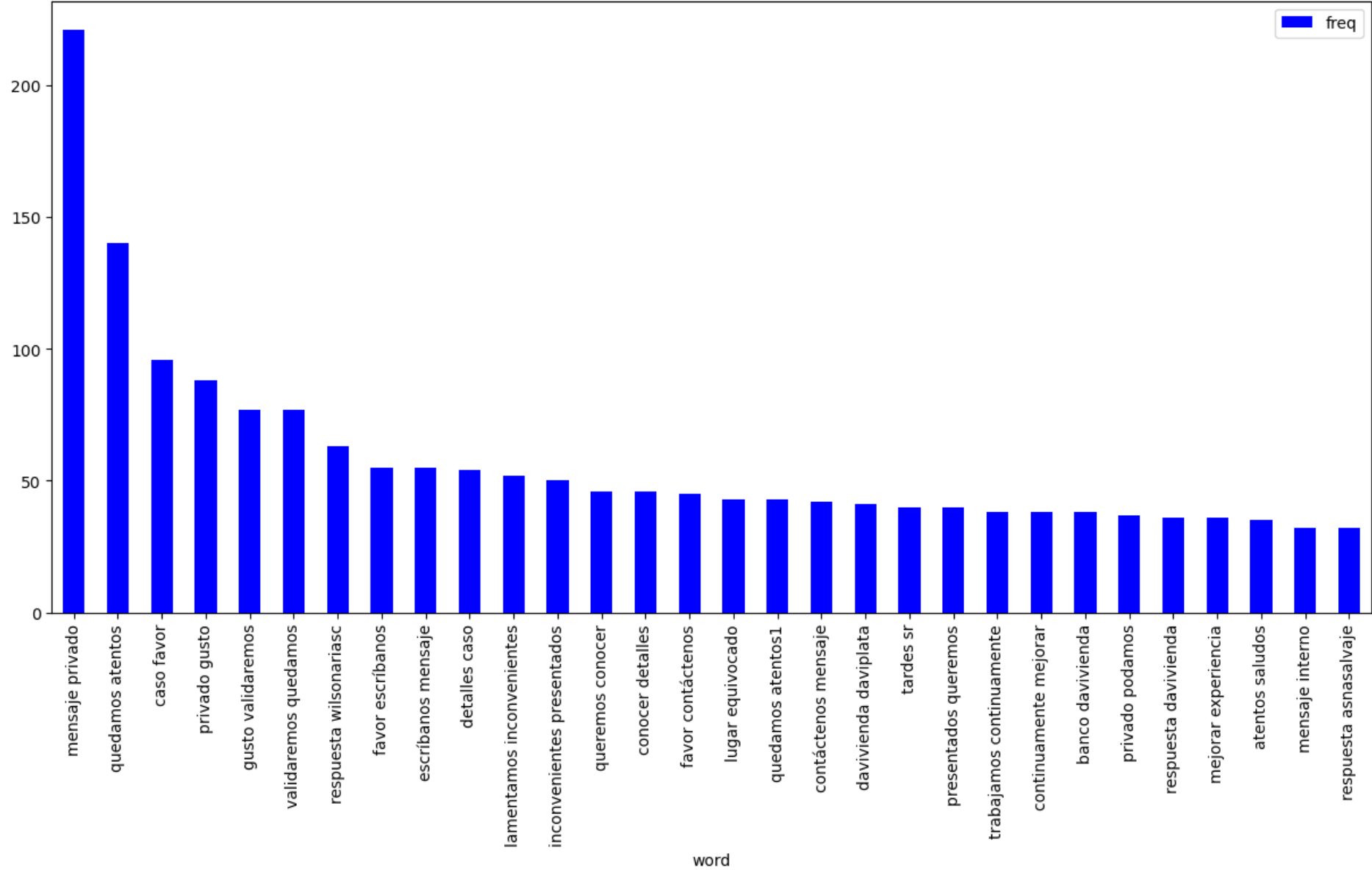
- el texto tratado
- la tokenización de cada texto
- las palabras lematizadas
- una columna con los hashtags
- una columna con los arrobas.

Histogramas de N gramas.



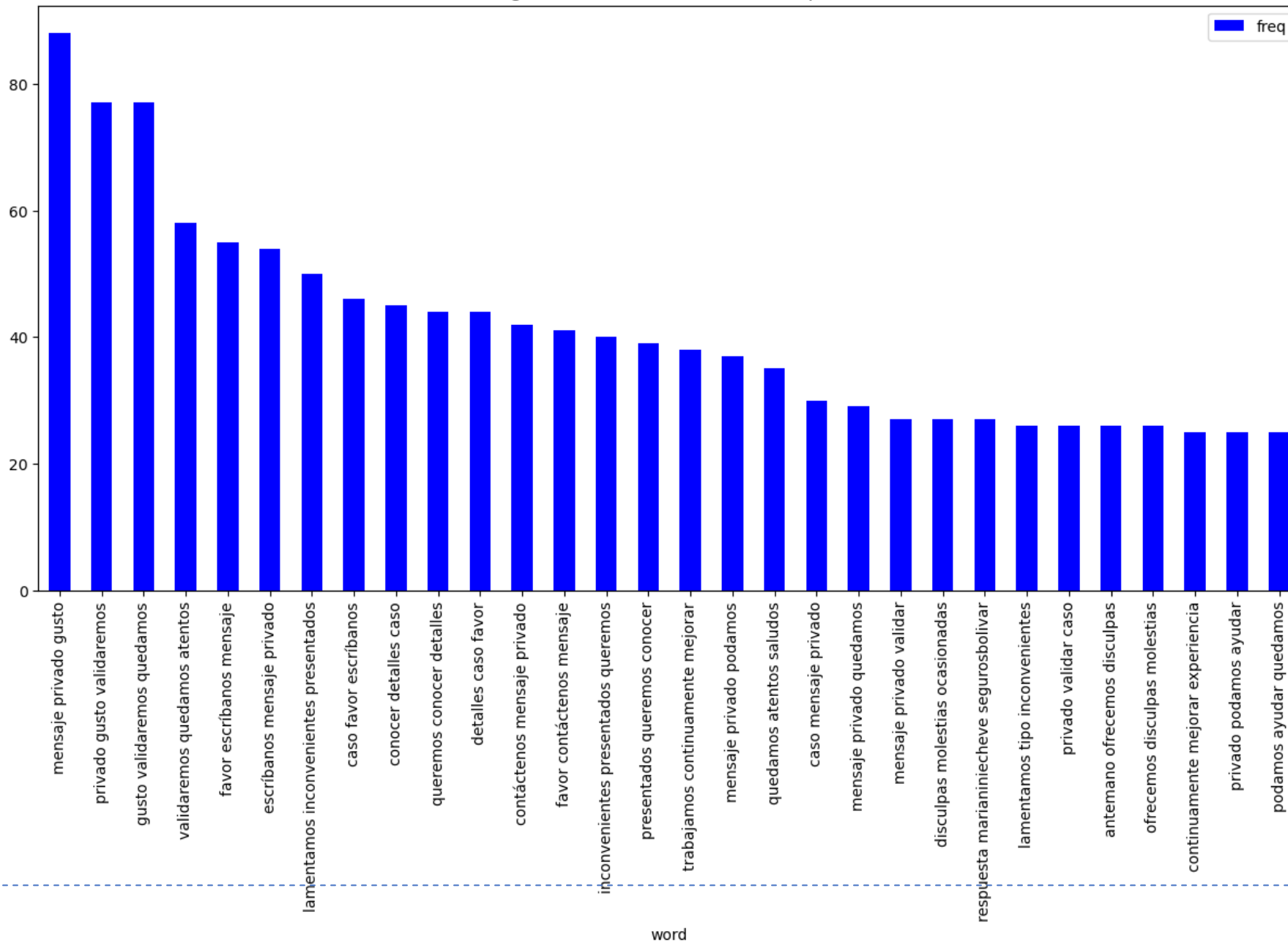
Se puede observar la frecuencia de las 30 palabras o unigramas mas usadas. Se puede observar que por mucho, las palabras davivienda y respuesta son las mas usadas en el set de datos.

Bigramas con más ocurrencias - Top 30



Se pueden observar los bigramas (composición de dos palabras) con mas ocurrencias en el set de datos.

Trigramas con más ocurrencias - Top 30



Se puede observar los trigramas (composicion de tres palabras) mas comunes.

WordCloud.

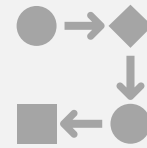


Se puede observar como se crea un nube de palabras o WordCloud; se puede ver que las palabras que mas resaltan son davivienda y respuesta.

Modelación.



Para la modelación, se crea la matriz de frecuencias, y el corpus. Para esto, se utiliza las librerías de Python *corpora* y *id2word*.



Posteriormente, se crea una iteración, en la cual se la métrica de coherencia de tópicos ,con distintos valores de numero de tópicos.



Luego que se tenga el valor optimo de numero de tópicos, se procede a crear el modelo LDA con el determinado numero de tópicos.

¿Qué es la coherencia de tópicos?

Las medidas de coherencia de tópicos califican un solo tema midiendo el grado de similitud semántica entre las palabras de alto puntaje en el tema. Estas medidas ayudan a distinguir entre temas que son temas interpretables semánticamente y temas que son artefactos de inferencia estadística. Se dice que un conjunto de enunciados o hechos es coherente, si se apoyan entre sí. Por lo tanto, un conjunto de hechos coherente puede interpretarse en un contexto que cubre todos o la mayoría de los hechos.

[Evaluate Topic Models: Latent Dirichlet Allocation \(LDA\) | by Shashank Kapadia | Towards Data Science](#)



Resultado de la Iteración.

Validation Set	Topics	Coherence
100% Corpus	2	0.363943
100% Corpus	3	0.386662
100% Corpus	4	0.420743
100% Corpus	5	0.414722
100% Corpus	6	0.368764
100% Corpus	7	0.365602
100% Corpus	8	0.331858
100% Corpus	9	0.346982
100% Corpus	10	0.365711
100% Corpus	11	0.371927
100% Corpus	12	0.349965
100% Corpus	13	0.343371
100% Corpus	14	0.357092
100% Corpus	15	0.368883
100% Corpus	16	0.380766
100% Corpus	17	0.360046
100% Corpus	18	0.360737
100% Corpus	19	0.336970
100% Corpus	20	0.348954
100% Corpus	21	0.373521
100% Corpus	22	0.351964
100% Corpus	23	0.367800
100% Corpus	24	0.350313
100% Corpus	25	0.350163
100% Corpus	26	0.336554
100% Corpus	27	0.358122
100% Corpus	28	0.354628
100% Corpus	29	0.372838

Teniendo en cuenta los resultados, se puede observar que según la iteración, los 3 números óptimos de tópicos, son 4, 5 y 16. Se correrá el modelo LDA utilizando los tres numero de tópicos.




Asignación de Dirichlet Latente.



La Asignación de Dirichlet Latente (LDA por sus siglas en inglés), es uno de los algoritmos generativos más usados y con mejor desempeño en cuanto a modelado de tópicos.

Este algoritmo tiene sus bases probabilísticas en la estadística bayesiana e inferencia estadística bayesiana, pero que gracias a su gran flexibilidad, es fácilmente implementable en lenguaje de maquina y al igual que el algoritmo de Bayes Ingenuo, su asunción de independencia facilita la computación.

En la Asignación de Dirichlet Latente, se considera que cada documento, o bolsa de palabras, tiene una mezcla de varios temas que son asignados a través de este algoritmo; en esta parte se supone que la distribución de cada tema proviene de una distribución de probabilidad de Dirichlet.



Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

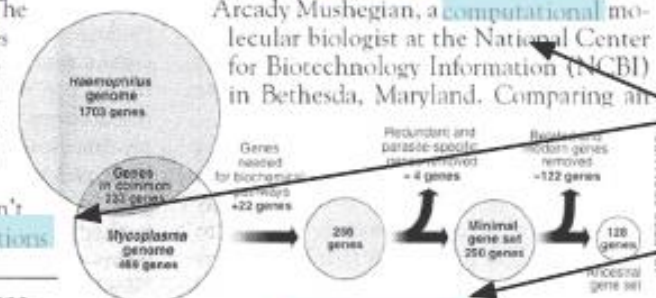
Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

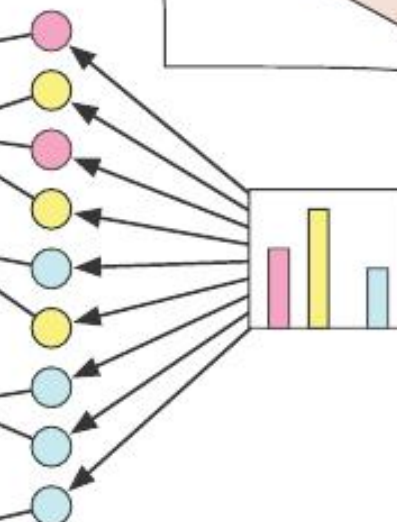


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. **Computer analysis** yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



Imprimir el resultado del modelo con 4 tópicos

```
[(0,
  '0.039*"respuesta*" + 0.026*"davivienda*" + 0.011*"mensaje*" + '
  '0.011*"privado*" + 0.008*"favor*" + 0.008*"caso*" + 0.008*"quedamos*" +
  '0.006*"atentos*" + 0.006*"cuenta*" + 0.005*"dinero*"''),
(1,
  '0.045*"davivienda*" + 0.025*"respuesta*" + 0.010*"daviplata*" + '
  '0.006*"cuenta*" + 0.006*"banco*" + 0.005*"q*" + 0.005*"respondiendo*"
  + '
  '0.004*"1*" + 0.004*"si*" + 0.004*"solución*"''),
(2,
  '0.030*"respuesta*" + 0.019*"davivienda*" + 0.015*"mensaje*" +
  0.014*"caso*" '
  '+ 0.012*"quedamos*" + 0.011*"favor*" + 0.011*"privado*" +
  0.010*"atentos*" '
  '+ 0.008*"lamentamos*" + 0.008*"inconvenientes*"''),
(3,
  '0.028*"davivienda*" + 0.015*"respuesta*" + 0.005*"daviplata*" + '
  '0.005*"banco*" + 0.005*"si*" + 0.004*"app*" + 0.004*"mismo*" + '
  '0.004*"solución*" + 0.004*"respondiendo*" + 0.003*"bancolombia*"'')]
```

Resultados y Evaluación.

Imprimir el resultado del modelo con 5 tópicos

```
[(0,
  '0.036*"respuesta*" + 0.026*"davivienda*" + 0.012*"mensaje*" + '
  '0.011*"privado*" + 0.009*"favor*" + 0.008*"quedamos*" + 0.008*"caso*" + '
  '0.008*"cuenta*" + 0.007*"atentos*" + 0.006*"dinero*"''),
(1,
  '0.045*"davivienda*" + 0.023*"respuesta*" + 0.011*"daviplata*" + '
  '0.008*"cuenta*" + 0.005*"banco*" + 0.005*"respondiendo*" + 0.004*"q*" + '
  '0.004*"1*" + 0.004*"solución*" + 0.004*"dinero*"''),
(2,
  '0.030*"respuesta*" + 0.018*"mensaje*" + 0.017*"caso*" + 0.017*"davivienda*" '
  '+ 0.015*"quedamos*" + 0.014*"privado*" + 0.014*"favor*" + 0.013*"atentos*" '
  '+ 0.010*"lamentamos*" + 0.010*"inconvenientes*"''),
(3,
  '0.028*"davivienda*" + 0.012*"respuesta*" + 0.007*"daviplata*" + '
  '0.005*"banco*" + 0.005*"app*" + 0.005*"si*" + 0.004*"mismo*" + '
  '0.004*"respondiendo*" + 0.004*"solución*" + 0.004*"problema*"''),
(4,
  '0.036*"respuesta*" + 0.031*"davivienda*" + 0.005*"wilsonariasc*" + '
  '0.005*"tarjeta*" + 0.005*"días*" + 0.005*"banco*" + 0.004*"q*" + '
  '0.004*"día*" + 0.004*"servicio*" + 0.004*"mensaje*"'')]
```

Interpretación de Resultados.

Se puede observar que las palabras que mas aparecen son “davivienda” y “respuesta”.

Se puede interpretar, que los usuarios que acuden a la red social Twitter, y mencionan al banco Davivienda, lo hacen principalmente con la intención de plantear un requerimiento, y buscando una respuesta. Se puede observar esto en los primeros tópicos que se forman al momento de imprimir el resultado de tópicos, del modelo LDA.

También se puede observar que la palabra privado aparece también en el primer tópico formado por el modelo, lo que haría referencia a que por favor se envíe el requerimiento por un mensaje privado.

En cuanto al segundo tópico, se puede observar que tiene la palabra “daviplata”, “cuenta”, “problema”, lo cual se puede de interpretar con que varios usuarios están presentando inconvenientes con la plataforma de Daviplata.