


Process Documents from Data

En el wordList resultante vemos el conglomerado de ocurrencias de las palabras relevantes encontradas en la columna de descripción. Para un total de 2220 documentos (registros) la palabra loan (préstamo) aparece en 787 documentos (registros) es la más ocurrente en el conjunto completo y por documento.

Result History

WordList (Process Documents from Data)

ExampleSet


Data

Word	Attribute Name	Total Occurences	Document Occurences
busi	busi	333	248
card	card	516	429
consolid	consolid	293	267
credit	credit	594	476
debt	debt	395	315
expens	expens	253	224
help	help	271	225
loan	loan	1004	787
pay	pay	764	578
payment	payment	343	275
purchas	purchas	313	280
year	year	270	232

En example set vemos el resultado de la búsqueda binaria de cada palabra por documento (registro), por ejemplo, en el registro 10 vemos que se encuentran tres palabras consolidar-deuda-pagar, la cual a simple vista podríamos discernir el motivo de este préstamo el cual es una consolidación de deudas.

Result History

WordList (Process Documents from Data)

ExampleSet (Process Documents from Data)

Data

Statistics

Visualizations

Annotations

Open in

Turbo Prep

Auto Model

Filter (2,220 / 2,220 examples):

all

Row No.	busi	card	consolid	credit	debt	expens	help	loan	pay	payment	purchas	year
1	0	0	0	0	0	0	0	0	0	0	1	0
2	0	0	0	0	0	0	1	0	0	0	1	0
3	0	0	0	1	0	0	0	1	1	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	0	0	1	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	1	0
8	0	0	0	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	0	0	1	0
10	0	0	1	0	1	0	0	0	1	0	0	0
11	0	0	0	0	0	1	0	1	0	0	0	0
12	0	0	0	0	0	0	0	1	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0

Para el método TF-IDF se muestra la relevancia de cada palabra dentro de cada documento, y para el mismo caso del documento 10 vemos que la palabra con menos peso es pagar, para la cual si extrajéramos esta nos quedaría consolidar-deuda que nos seguiría brindando el mismo significado.

WordList (Process Documents from Data) × ExampleSet (Process Documents from Data) ×													
<div> <div>Open in</div> <div>Turbo Prep</div> <div>Auto Model</div> </div> <div>Filter (2,220 / 2,220 examples): all</div>	Row No.	busi	card	consolid	credit	debt	expens	help	loan	pay	payment	purchas	year
	1	0	0	0	0	0	0	0	0	0	0	1	0
	2	0	0	0	0	0	0	0.742	0	0	0	0.671	0
	3	0	0	0	0.672	0	0	0	0.452	0.587	0	0	0
	4	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	1	0	0	0	0
	6	0	0	0	0	0	0	0	0.676	0	0	0	0.737
	7	0	0	0	0	0	0	0	0	0	0	1	0
	8	0	0	0	0	0	0	0	0	0	0	1	0
	9	0	0	0	0	0	0	0	0	0	0	1	0
	10	0	0	0.666	0	0.614	0	0	0	0.423	0	0	0
	11	0	0	0	0	0	0.911	0	0.412	0	0	0	0
	12	0	0	0	0	0	0	0	1	0	0	0	0

En el caso de Term Occurrences, pienso que no nos brinda mayor significancia que TF-IDF ya que nos da un resumen de conteo de palabras por documento, aunque si estableceríamos la importancia de una palabra dentro de un documento basado en su número de apariciones los resultados serian muy distintos a los encontrados en TF-IDF, por ejemplo el documento 10, en este caso las ocurrencias son iguales para las tres palabras por lo cual el peso de cada una seria de 0,33

WordList (Process Documents from Data) ExampleSet (Process Documents from Data)												
Open in		Turbo Prep	Auto Model		Filter (2,220 / 2,220 examples): all							
Row No.	busi	card	consolid	credit	debt	expens	help	loan	pay	payment	purchas	year
1	0	0	0	0	0	0	0	0	0	0	1	0
2	0	0	0	0	0	0	1	0	0	0	1	0
3	0	0	0	1	0	0	0	1	1	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	0	0	2	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	1	0
8	0	0	0	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	0	0	2	0
10	0	0	1	0	1	0	0	0	1	0	0	0
11	0	0	0	0	0	1	0	1	0	0	0	0
12	0	0	0	0	0	0	0	1	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	1	0	1	0	0	2	0	0	1	0
17	0	0	0	0	0	0	0	0	0	0	1	0

¿Crees que binary term occurrences es la mejor forma de representar los datos?

No, binary term occurrences nos brinda un análisis muy básico en el que únicamente podemos ver si la palabra se encuentra en un documento o no.

¿Cuál es la diferencia entre las dos vistas de los resultados (Example Set y Word List)?

WordList muestra un resumen del conteo por palabra, mientras que ExampleSet muestra un detalle de conteo por documento.

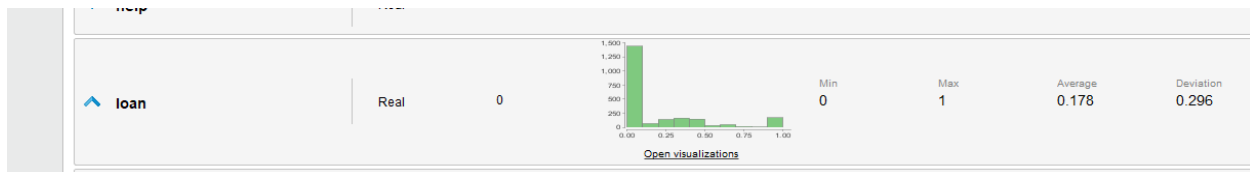
¿Cuál es la diferencia entre Total Occurences y Document Occurences?

Total occurrences es la cantidad de veces que una palabra se encuentra en el fichero, y document occurrences es el numero de veces que la palabra se encuentra en el fichero agrupada por documentos, es decir, si la palabra PAY se encuentra dos veces en un documento, suma 1 al document occurrences para PAY.

¿Cuáles serían los valores óptimos de podado (pruning) para quedarnos con palabras representativas?

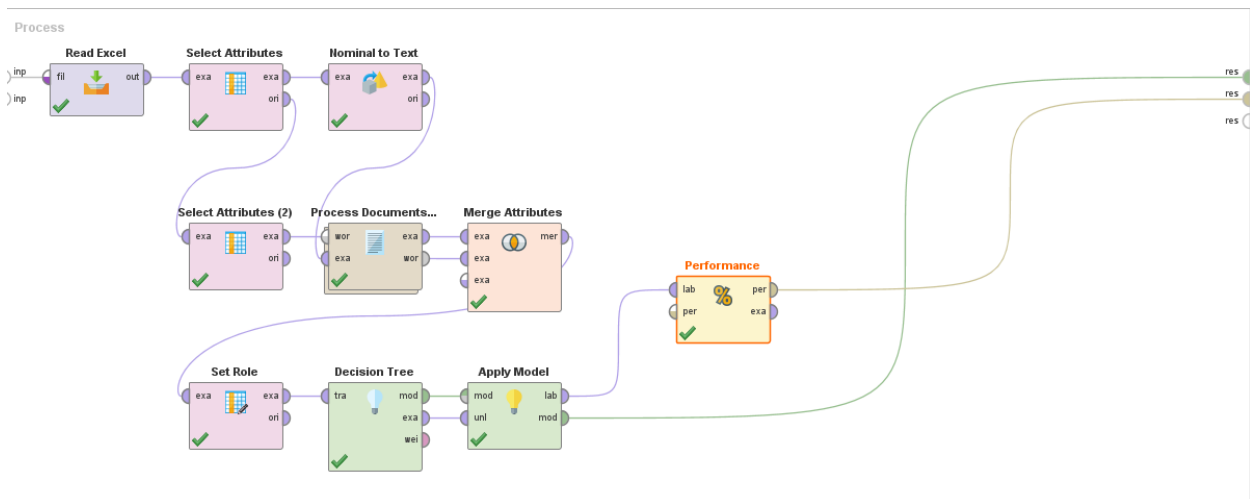
Considero que podar las palabras menores que 3 caracteres es un buen acercamiento, ya que por ejemplo hay palabras de 3 caracteres que, si pueden llegar a brindar mayor relevancia a alguna palabra dentro del documento, por ejemplo “aun” podría darle mayor relevancia a una palabra predecesora, y “por” a una palabra sucesora.

En la vista Example Set, elige algún término representativo y abriendo la pestaña statistics y luego visualization, muestra el histograma de ese término ¿Qué representa?



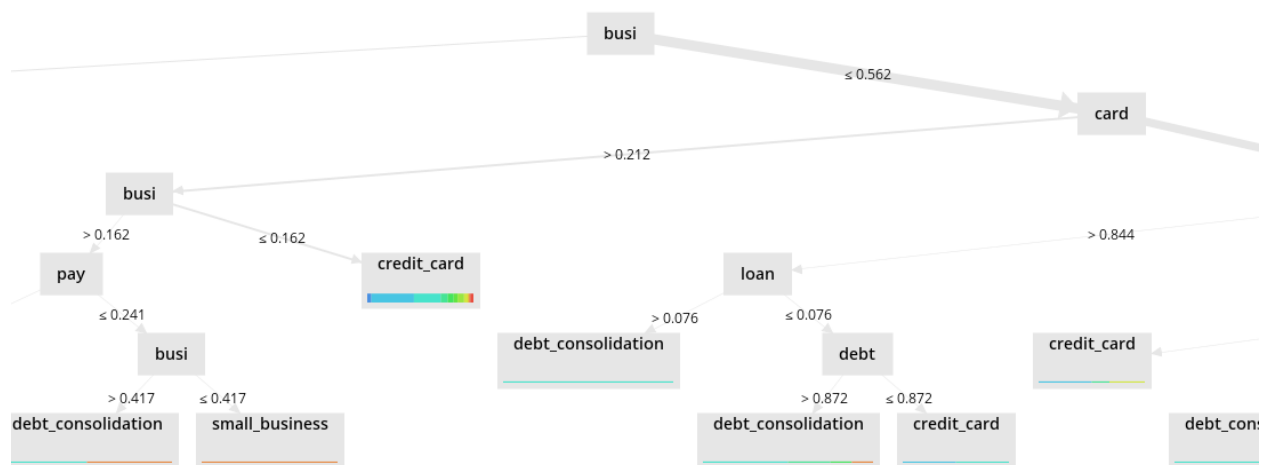
En el caso de TF-IDF muestra un resumen de relevancia por palabra, mostrando su mínima y máxima relevancia junto con la media de relevancia con la cual podríamos determinar la palabra más relevante en el fichero.

Entrega una impresión de pantalla del proceso configurado.



¿Qué resultados de clasificación obtienes?

Como resultados tenemos una matriz y un árbol donde se clasifican las relaciones entre las palabras en la descripción del préstamo, según el árbol, la palabra padre con la que se llega a una clasificación optima para este caso es busi, siendo sus hijos inmediatos card y payment. Adicional obtenemos el un acierto en la predicción del modelo que llega al 33.24% considerado como bajo ya que de cada 3 hojas acierta la rama padre de una sola.



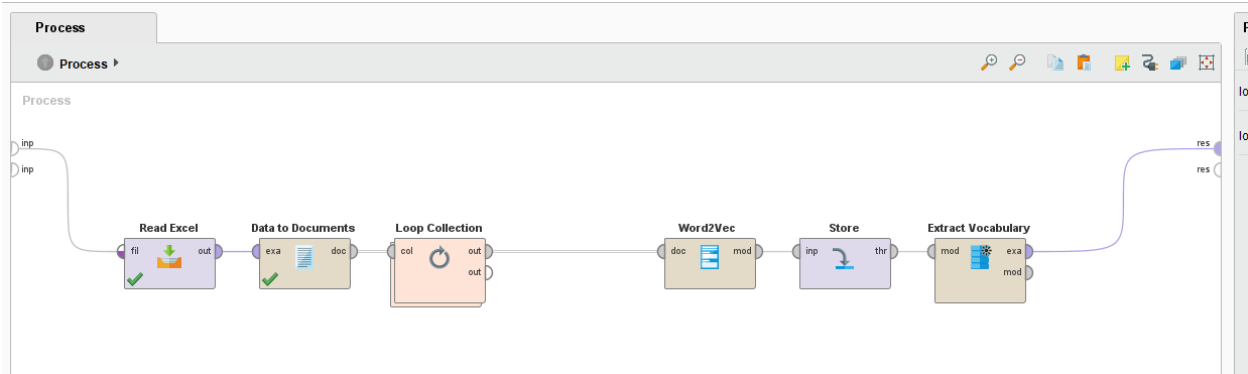
Prueba diferentes representaciones de los documentos (configuración del operador process text) e indica con cuál de ellas se obtienen los mejores resultados ¿Por qué crees que es así?

Generación de la representación con Word embeddings

Al correr el proceso sale un error que pide tokenizar los documentos, pero dentro del loop collection estoy aplicando el operador de tokenize y he probado los diferentes parámetros de dicho operador:



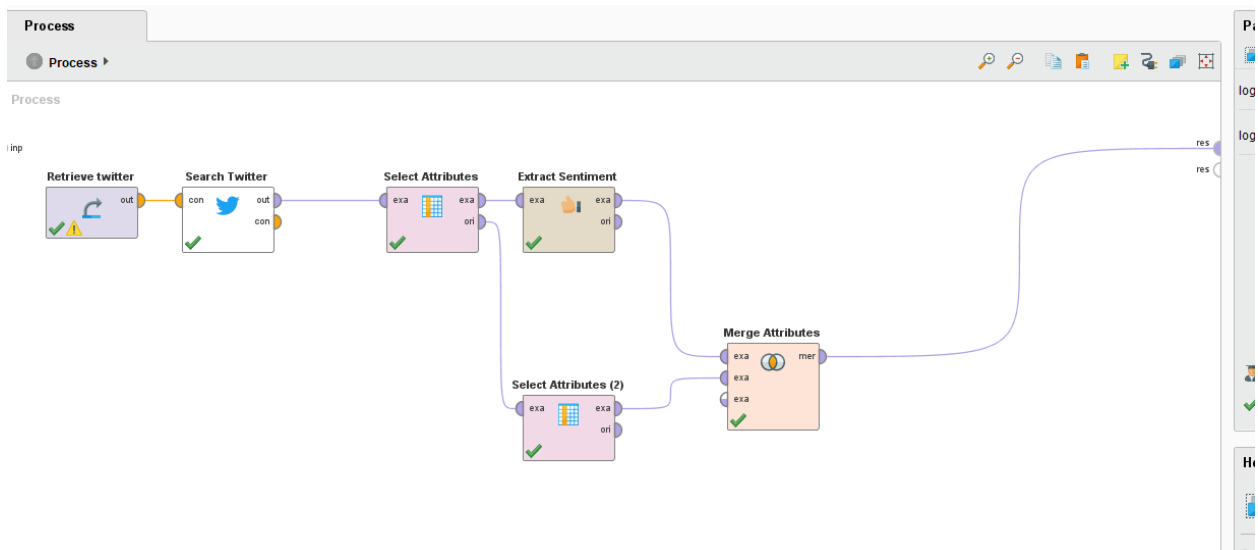
Finalmente, el proceso se visualiza de la siguiente forma:



Según entiendo con este proceso el objetivo es generar un corpus a través de la entrada de motivos de préstamo, para que posteriormente se puedan generar cálculos en sentencias procedentes de otros motivos de préstamo.

Obteniendo datos de twitter

Se busca la palabra IBEX dentro de los Tweets con el objetivo de hacer un análisis de sentimiento sobre la cotizada quedando el siguiente modelo:



Se extrae el score del sentimiento configurado de la siguiente manera:

Muy positivo = 1, Positivo = 0.75, Neutral = 0, Negativo = -0.75 and Muy negativo = -1

Adicional se selecciona el atributo de conteo de retweets para posteriormente analizar la importancia y pos tanto un posible impacto de la subida o bajada del precio del índice. Se obtienen los siguientes resultados:

Row No.	Id_1	Score	Text	Created-At	Retweet-Co...	Id_2
1	1252700453...	0	Muy posible publicidad progubernamental que seguro que bancos e Ibex fl...	Apr 21, 2020 10:47:16 PM CEST	89	1252700453...
2	1252159868...	0	El Ibex sufre la mayor revisión a la baja de estimaciones de beneficios de I...	Apr 20, 2020 10:59:11 AM CEST	57	1252159868...
3	1252141210...	-0.308	Casado, a empresarios del Ibex: "Sánchez quiere lo mismo que Zapatero ...	Apr 20, 2020 9:45:02 AM CEST	27	1252141210...
4	1252710489...	0	RT @MarcelBL21: IBEX-35 y CEOE, renuncien a los paraísos fiscales y pa...	Apr 21, 2020 11:27:09 PM CEST	52	1252710489...
5	1252710406...	0	RT @rimbaudarth: Hola, ☐@davidbroncano☐	Apr 21, 2020 11:26:49 PM CEST	336	1252710406...
6	1252710214...	0	RT @hermannntersch: Muy posible publicidad progubernamental que segu...	Apr 21, 2020 11:26:04 PM CEST	89	1252710214...
7	1252710193...	-0.308	RT @FrancoZombi: Y si queréis saber mejor por qué esto no sale en los ...	Apr 21, 2020 11:25:58 PM CEST	61	1252710193...
8	1252709416...	0	RT @hermannntersch: Muy posible publicidad progubernamental que segu...	Apr 21, 2020 11:22:53 PM CEST	89	1252709416...
9	1252709361...	0	RT @MarcelBL21: IBEX-35 y CEOE, renuncien a los paraísos fiscales y pa...	Apr 21, 2020 11:22:40 PM CEST	52	1252709361...
10	1252709094...	-0.308	RT @FrancoZombi: Y si queréis saber mejor por qué esto no sale en los ...	Apr 21, 2020 11:21:36 PM CEST	61	1252709094...
11	1252709077...	0	RT @LouMonth: Florentino Pérez, el padrino del IBEX y de los medios de c...	Apr 21, 2020 11:21:32 PM CEST	2	1252709077...
12	1252709010...	0	RT @hermannntersch: Muy posible publicidad progubernamental que segu...	Apr 21, 2020 11:21:16 PM CEST	89	1252709010...
13	1252708944...	0	RT @javersastre_: Comentario de Cierre de #Mercados (21 Abril).	Apr 21, 2020 11:21:01 PM CEST	1	1252708944...
14	1252708877...	-0.308	Comentario de Cierre de #Mercados (21 Abril).	Apr 21, 2020 11:20:45 PM CEST	1	1252708877...
15	1252708696...	0	RT @veranoaz: https://t.co/pCXBqIfScl para reducir el pago de impuestos. E...	Apr 21, 2020 11:20:02 PM CEST	2	1252708696...
16	1252708599...	0	RT @MarcelBL21: IBEX-35 y CEOE, renuncien a los paraísos fiscales y pa...	Apr 21, 2020 11:19:39 PM CEST	52	1252708599...
17	1252708436...	0	RT @NatureIsMetal: Arabian Wolf hunting ibex (wild goat) on a desert cliff ht...	Apr 21, 2020 11:19:00 PM CEST	225	1252708436...
18	1252708346...	0	RT @LouMonth: Florentino Pérez, el padrino del IBEX y de los medios de c...	Apr 21, 2020 11:18:38 PM CEST	2	1252708346...
19	1252708323...	0	Now religious extremists are claiming that the coronavirus pandemic starte...	Apr 21, 2020 11:18:33 PM CEST	0	1252708323...
20	1252708262...	0	RT @hermannntersch: Muy posible publicidad progubernamental que segu...	Apr 21, 2020 11:18:18 PM CEST	89	1252708262...
21	1252708201...	0	RT @elEconomistaes: #Apertura📈 Las bolsas europeas retroceden más ...	Apr 21, 2020 11:18:04 PM CEST	4	1252708201...
22	1252708178...	0	RT @elEconomistaes: #Actualización📈 Las bolsas europeas retroceden ...	Apr 21, 2020 11:17:58 PM CEST	2	1252708178...
23	1252708058...	0.359	RT @matthewbennett: 16. El IBEX 35 y el Coronavirus: está entre los 6.000 ...	Apr 21, 2020 11:17:30 PM CEST	25	1252708058...

Como se aprecia la mayor cantidad de resultados son de tipo neutral por lo que considero se podrían podar para quedarnos con los positivos y negativos.

El modelo aquí presente se complementaría con un proceso de extracción de data financiera, ya sea de Google finance o yahoo finance para obtener el precio de cierre y calcular su volatilidad para que fuese comparada con una serie temporal de score de sentimientos y así poder verificar el impacto que tienen los tweets relevantes en el precio de una cotizada.