GIE Webscrapping

Carlos Espeleta y Erick Quispe

05/06/2017

Contenido:

1.- Introducción:

1.1. Motivación.

En la actualidad, análisis textual es una herramienta muy útil para poder extraer información sobre temas sociales que se tratan día a día, descubrir posibles apariciones de tendencia o que piensa la gente sobre determinados temas.

En el marketing está técnica cada vez más es utilizada para poder realizar estrategias de ventas o lanzamientos de productos a partir de lo que las personas publican en redes sociales o diversas plataformas de internet.

En esta parte, el objetivo que perseguimos es poder hacer un análisis textual sobre tres temas.

1.- Comparación de información que publican 2 diferentes plataformas financieras en internet.

2.- Mercados financieros

3.- Economía.

En el primer, lo que queremos ver es en que información se centran más cada una de las plataformas analizadas, que en este caso es la cuenta oficial de los bancos de inversión Golman Sachs (@GoldmanSachs) y J.P. Morgan (@jpmorgan).

Mientras que en el caso de los temas de Mercados Financieros y Economía lo que pretendemos es ver con qué términos se relacionan estas dos palabras.

2.- Naturaleza de los datos.

Los datos para realizar el análisis son extraídos de la red social “Twiter”. Para el caso del análisis sobre Mercados Financieros y Economía, extraemos Tweets de diferentes partes que hablen exclusivamente sobre estos temas en concretos y ver cómo hemos comentado anteriormente, con que palabras la gente relaciona estos términos.

En cambio, para el caso de la comparación de la información que publican las 2 diferentes plataformas financieras, extraemos los tweets directamente de sus respectivas cuentas oficiales.

Figura 1. Imagen de cuenta oficial de la cuenta del banco de inverson J.P Morgan

3.- Extracción de los datos:

Actualmente, para poder extraer datos de Twitter, es necesario obtener unas credenciales, una vez se han adquirido las credenciales, podemos extraer los datos de la plataforma. En R trabajaremos con los paquetes "twitteR", "tm", "wordcloud".

```
consumer_key <- "gryb8NUTvfcmnDFcTOJB2V1"
consumer_secret <-"bP46GwVdRVh4VTtuhjEcZ5gNGwvDB8YYuC2fVOSRCbgMRra"
access_token <-"8355046-SsmIHAyMGwm21jh7b7r0JmUPatWL8WOVzmYF5lqK"
access_secret <-"d28iU79NWv6X15Qx8bFVTmCebwwFVnV5nZi8nLQA1qy"
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
```

Una vez activado las credenciales en R, procedemos a extraer primero los tweets referentes a las cuentas de J.P. Morgan y Goldman Sach y los tweets referente a los temas de mercados financieros "**Financial Market**" y a "**Economy**". Extraeremos un total de 10.000 tweets para cada tema. Una vez extraídos los ponemos en forma de vectores.

```
mach_tweets = searchTwitter("jpmorgan", n=10000, lang="en")
mach_text = sapply(mach_tweets, function(x) x$getText())
```

```
mach_tweets = searchTwitter("Economy", n=10000, lang="en")
mach_text = sapply(mach_tweets, function(x) x$getText())
```

Una vez hemos puestos los datos en un vector, empezamos la limpieza de los mismos removiendo símbolos de puntuación, números, link, Retweets etc.

```
txtclean = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", mach_text)
txtclean = gsub("@\\w+", "", txtclean)
txtclean = gsub("[[:punct:]]", "", txtclean)
txtclean = gsub("[[:digit:]]", "", txtclean)
txtclean = gsub("http\\w+", "", txtclean)
```

Una vez hecha la limpieza de los datos, creamos el cuerpo textual (corpus) y la matriz de documento (Term Document Matriz) aplicando algunas transformaciones, para que una nos elimine emoticonos y otros simboles raros.

En stopword le digo que me elimine las palabras por las que he buscado los tweets para extraerlos ya que no me interesan, porque serán las que más se repitan y esto distorsionará el análisis, además para remover otras palabras le digo que use el “stopwords” en inglés, ya que es el idioma en los que he extraído los datos.

```
mach_corpus = Corpus(VectorSource(mach_text))
mytwittersearch_corpus<- tm_map(mach_corpus, content_transformer(function(x)
  iconv(enc2utf8(x), sub = "bytes")))
tdm = TermDocumentMatrix(mytwittersearch_corpus,
                         control = list(removePunctuation = TRUE,
                        stopwords = c("Market", "finalcial", stopwords("english")),
                          removeNumbers = TRUE, tolower = TRUE))
```

Posteriormente definimos el “tdm” como una matriz, la ordenamos de forma descendente, de manera que en primer lugar queden aquellas palabras que tienen una mayor frecuencia, por último transformo la matriz en un data frame.

```
m = as.matrix(tdm)
word_freqs = sort(rowSums(m), decreasing=TRUE)
dm = data.frame(word=names(word_freqs), freq=word_freqs)
```

Como al quedar palabras que no han sido eliminadas, o palabras que tiene símbolos extraños o palabras sin sentido, exportamos los datos a un fichero csv para poder terminar la limpieza de forma manual. Una vez realizada la última limpieza volvemos a importar el fichero a R y realizamos un Barplot y la nube de palabras.

```
write.table(dm, "dm.csv", sep=",")
dm<-read.csv2("dm.csv")
dm0<-dm[dm$freq>400,]
barplot(dm0$freq, las = 2, names.arg = dm0$word,
        col ="lightblue", main ="Most frequent words",
        ylab = "Word frequencies")

png("Cloud.png", width=7, height=8, units="in", res=500)
wordcloud(dm$word, dm$freq,min.freq = 50, random.order=FALSE,
          colors=brewer.pal(8, "Dark2"))
dev.off()
```

4.- Resultados de la comparación de la información de J.P. Morgan y Goldman Sachs:

Como podemos observar en la nube de palabras de los tweets extraídos de la cuenta de J.P. Morgan (grafico 2) vemos que en la gran mayoría de tweets se habla de persecución (chase), sobre la Libra, de Coalición y en menor medida sobre Venezuela y la crisis que está pasando.

Mientras que si vemos la nube de palabras con los tweets extraídos de Goldman Sachs (gráfico 3), observamos que en su mayoría se habla sobre la situación de Venezuela, los venezolanos, Maduro y Bonos que seguramente harán referencia a los bonos venezolanos que hace poco ha adquirido Goldman Sachs.

Por lo que podemos decir que 2 cuentas de empresas de un mismo sector emiten informaciones diferentes en internet, seguramente publiquen información según sea de su interés.

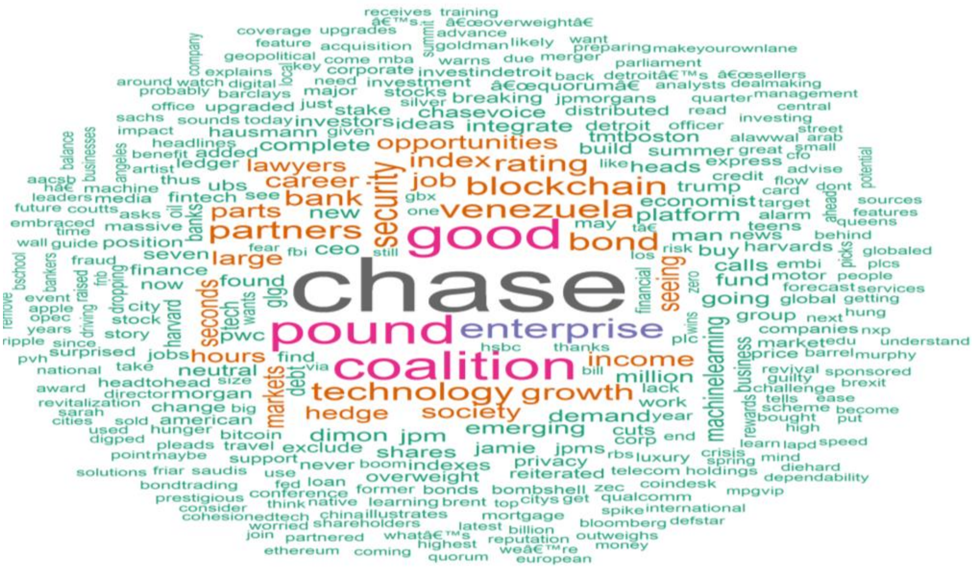


Gráfico 2: Nube de palabras con los datos de J.P. Morgan.

Gráfico 3: Nube de palabras con los datos de Goldman Sachs.

5.- Resultados de los temas “Economy” y “Financial Market”:

En el Barplot podemos observar algunas de las palabras que poseen una mayor frecuencia. Por ejemplo, “money” es la que más frecuencia posee, con un recuento de 852 repeticiones, seguido por la palabra “buying”.

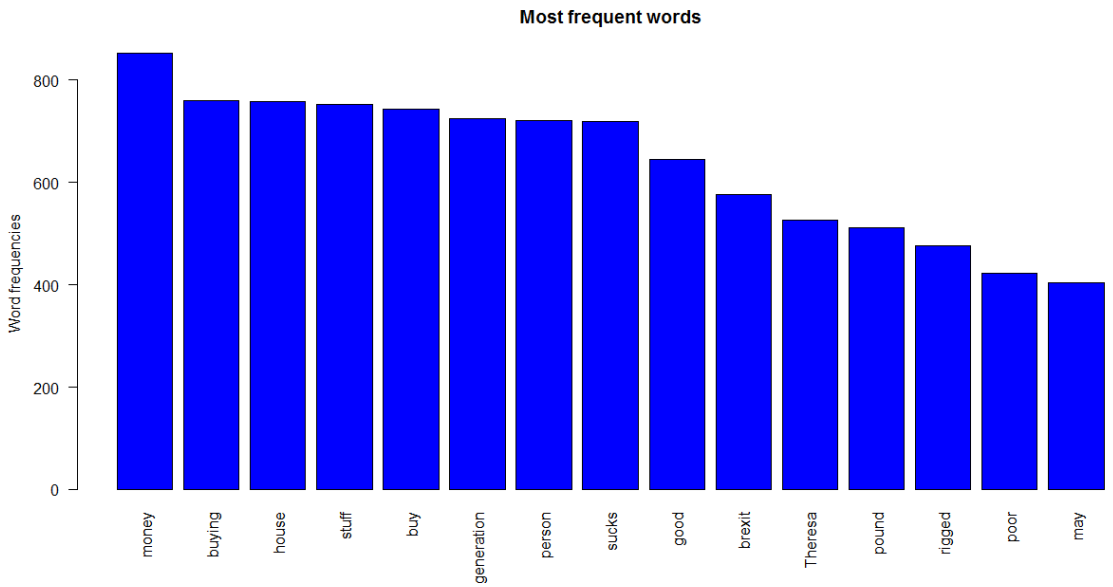


Gráfico de barras

Si observamos más abajo la nube de palabras (grafico 4) construida a partir del análisis textual del tema “Economy” podemos sacar las siguientes conclusiones:

Cuando se habla de economía, en un gran porcentaje de veces se lo asocia con las siguientes palabras: dinero, comprar, mercaderías, generación, pobreza, tasa de interés, negocio, trabajo, gobierno, inflación etc.

Por otro lado, observamos también que cuando se habla de economía, se lo relaciona con eventos y personajes políticos como en el caso de Theresa (May), Trump o el tema del Brexit, así como la libra.

Otro punto a observar, para algunos el tema economía es un tema un poco espinoso, ya que como vemos la palabra “sucks” aparece un gran porcentaje de veces.

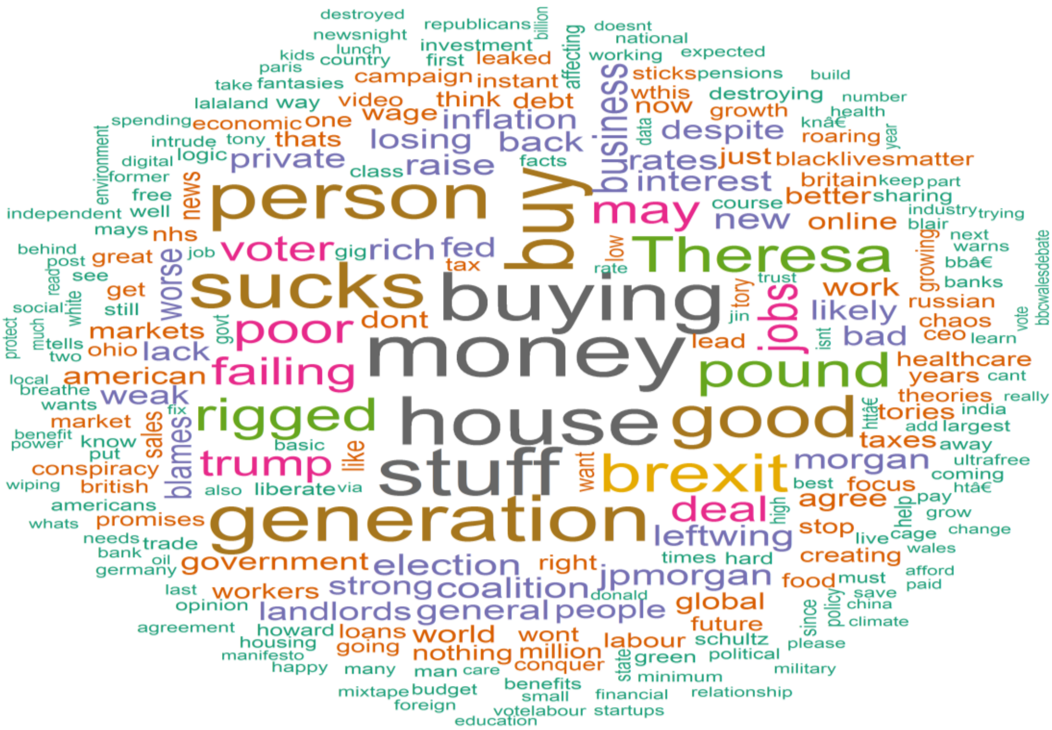
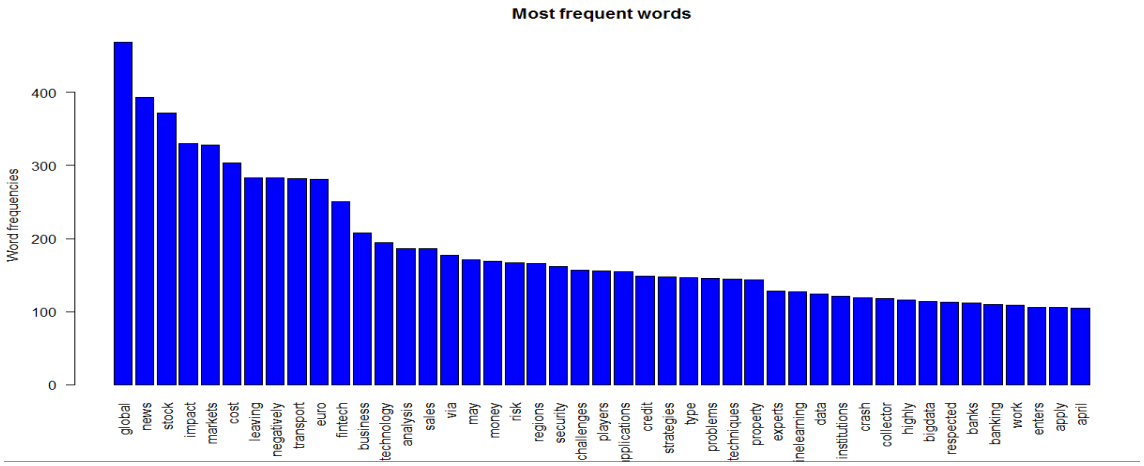


Gráfico 4: Nube de palabras con los datos recogidos sobre el teme de “economy”

En el caso de “Financial Market” vemos que las palabras que más se repiten es global, noticias(news), acciones (stocks), impacto (impact) entre otras.



Como observamos en la nube de palabras, cuando se habla de mercados financieros, se lo asocia con palabras como acciones, euro, negocios, riesgo, prosperidad.

Gráfico 4: Nube de palabras con los datos recogidos sobre el teme de “Financial Market”