DOCUMENTACIÓN DE ELASTIC SEARCH

ENTREGABLES

13 DE DICIEMBRE DE 2022

MIGUEL GONZALEZ NAVARRO

UO282337

Contenido

1. Entregable 1

El objetivo de este ejercicio es generar un fichero donde se recojan por líneas los siguientes datos: la entidad de WikiData del trending topic, la lista de trending topics asociados a esa entidad (también denominados sinónimos) y el tipo de esa entidad (en el caso que lo tenga). Se han utilizados todos los tweets de castellano e inglés, como aspecto a destacar, aunque solo se trabajará sobre aquellos en español.

Se han entregado dos archivos de código fuente: indexer-ej1.py y query-ej1.py. El primero de ellos se encargaba de generar el índice con los tuits escritos en español utilizando shingles de dos y tres términos, una vez ejecutado este script, podemos ver en elasticvue (en el apartado de índices) que se ha generado correctamente (353 MB).

Name	Health	Status	UUID	Aliases	Shards	Lucene docs	Storage
tweets-miguel	yellow	open	Pc8fvqPfT0irWRySwG_AmA	[ ]	1 / 1	989998	353 MB

Para crear este fichero, se ha aprovechado código del fichero “bulk-indexer4.py” de las diapositivas de prácticas, solo se ha modificado que el idioma a seleccionar de los tuits sea el español, además de la estematizacion, que sea también español. Se ha tenido en cuenta la eliminación de palabras vacías que había ya en el código base (puesto que no aportan ninguna información relevante y no son necesarias, y así evitar “ruido”).

```
"palabras_vacias_ingles_porter": {
    "type": "stop",
    "stopwords": ["a","about","above","after","again",
                  "against","all","am","an","and","any","are",
                  "aren't","as","at","be","because","been",
                  "before","being","below","between","both",
                  "but","by","can't","cannot","could",
                  "couldn't","did","didn't","do","does",
                  "doesn't","doing","don't","down","during",
                  "each","few","for","from","further","had",
                  "hadn't","has","hasn't","have","haven't",
                  "having","he","he'd","he'll","he's","her",
                  "here","here's","hers","herself","him",
                  "himself","his","how","how's","i","i'd",
                  "i'll","i'm","i've","if","in","into","is",
                  "isn't","it","it's","its","itself","let's",
                  "me","more","most","mustn't","my","myself",
                  "no","nor","not","of","off","on","once",
                  "only","or","other","ought","our","ours",
                  "ourselves","out","over","own","same",
                  "shan't","she","she'd","she'll","she's",
                  "should","shouldn't","so","some","such",
                  "than","that","that's","the","their",
                  "theirs","them","themselves","then","there",
                  "there's","these","they","they'd","they'll",
                  "they're","they've","this","those","through",
                  "to","too","under","until","up","very","was",
                  "wasn't","we","we'd","we'll","we're","we've",
                  "were","weren't","what","what's","when",
                  "when's","where","where's","which","while",
                  "who","who's","whom","why","why's","with",
                  "won't","would","wouldn't","you","you'd",
                  "you'll","you're","you've","your","yours",
                  "yourself","yourselves","d","ll","m","re",
                  "s","t","ve","aren","can","couldn","didn",
                  "doesn","don","hadn","hasn","haven","he",
                  "here","how","i","isn","it","let","mustn",
                  "shan","she","shouldn","that","there","they",
                  "wasn","we","weren","what","when","where",
                  "who","why","won","wouldn","you"]
}
```

En el segundo fichero (query-ej1.py), se ha hecho una consulta al índice de los tweets (se puede previsualizar los resultados de esta aportando los argumentos correctos), este script está basado en los ejemplos de agregación que generaba trending topics, en este caso se quieren consultar 50 trending topics para cada hora.

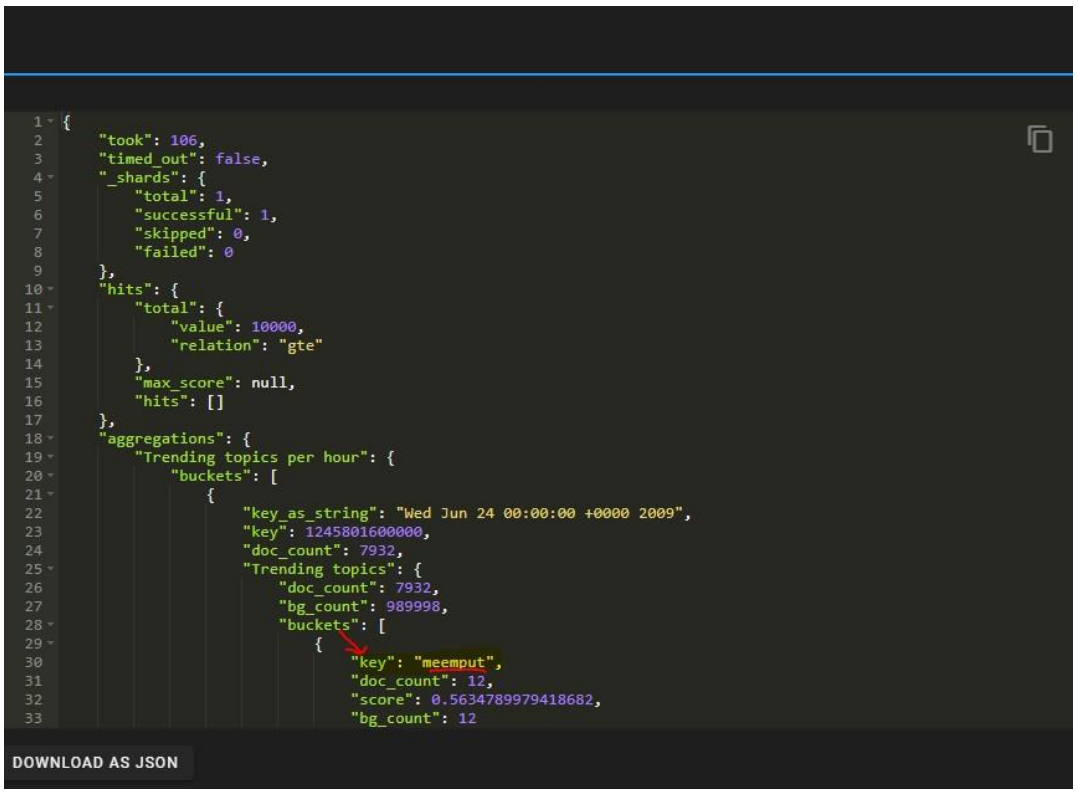
Me he basado en la siguiente diapositiva de prácticas:

Agregaciones

- Por último, podemos calcular varias agregaciones en una única consulta. Las agregaciones pueden estar al mismo nivel—*siblings*—o una agregación puede trabajar sobre la salida de la anterior—*parent*. En este ejemplo vamos a **calcular 10 "trending topics" por hora.**

```
tweets-20090624-20090626-en_es-10percent/_search
{
  "size": 0,
  "query": {
    "query_string": {
      "query": "lang:en"
    }
  },
  "aggs": {
    "Trending topics per hour": {
      "date_histogram": {
        "field": "created_at",
        "fixed_interval": "1h"
      },
      "aggs": {
        "Trending topics": {
          "significant_terms": {
            "field": "text",
            "size": 10,
            "gnd": {}
          }
        }
      }
    }
  }
}
```

Pero he cambiado detalles, como que el idioma sea español y que se generen 50 trending topics. Para visualizar los resultados en elasticvue, vamos al apartado de REST y seleccionamos el método POST, solo queda poner nuestro Path donde se encuentra el índice, colocar los argumentos de la consulta en la pantalla y darle al SEND REQUEST, nos saldrá algo tal que así:



El uso de esta herramienta me ha ayudado a visualizar los datos y poder entender la agregación y demás. Los trending topics serán los valores que nos aparecen en el campo key, por lo que en Python debemos de recoger esos datos a través del json que nos devuelve la consulta, iremos navegando a través de este y guardaremos en una lista los trending topics por hora.

Una vez que tenemos esta lista, tenemos que encontrar los sinónimos (esto era una de las ampliaciones), para ello se utilizará WikiData, que proporciona una API REST, en la que a través de Requests, accederemos a datos que nos proporciona (los datos están representados en un JSON al igual que en elasticvue).

Para acceder a WikiData tenemos que importar requests, y hacer una consulta a través de los enlaces (en este caso se utiliza el segundo, sustituyendo “michael+jackson” por el trending topic de la lista) que se proporciona en la diapositiva del entregable:

Pistas:

- Wikidata proporciona un API REST que **debería usarse** en este ejercicio (en otras palabras, no utilices SPARQL).
- https://pypi.org/project/requests/
- https://www.wikidata.org/w/api.php?action=wbsearchentities&language=en&format=json&search=michael+jackson
- https://www.wikidata.org/w/api.php?action=wbgetentities&ids=Q2831&languages=en&format=json

Los datos que nos proporciona (yo los he podido visualizar de una forma muy cómoda y estructurada en Firefox ya que en Edge se veían todas las líneas del JSON juntas) es un JSON con información acerca de ese trending topic (entidad), tenemos que encontrar la primera id que encontremos, tendrá el formato “Q…”.

Para esta implementación, a medida que vamos encontrando la entidad de cada trending topic las vamos añadiendo a un diccionario, cade destacar que un trending topic es sinónimo de otro si tiene su misma entidad. En el caso de que el trending topic no tenga ninguna Q asociada, se añade igualmente al volcado, pero no tiene entidad asociada (se indicará en el fichero de salida con toda la información). Iremos acumulando todos estos datos por pasos, en una lista de cadenas, donde cada elemento representa una línea de un fichero (tenemos una por cada trending topic asociado). En el caso de que un trending topic se repita más de una vez, será representado en el fichero dos veces con la misma entidad, sinónimos y tipo; pero con distinta fecha y hora. Se han incluido comentarios en el código para entender esta implementación.

Ejemplo: Formato del diccionario

{“Michael Jackson”: “Q2831”, “Rey del pop”: “Q2831”} [Michael Jackson y rey del pop son sinónimos]

No se trata de buscar sinónimos en elastic search, sino tratar de seguir esta mecánica.

Por último, para encontrar el tipo de la entidad, recorremos este diccionario, y a través de WikiData (utilizando el tercer enlace de la pantalla anterior) podremos acceder al tipo, sustituyendo la Q por la de cada entidad del diccionario, y lo vamos añadiendo a un nuevo diccionario, donde tenemos por clave la entidad y como valor el tipo (en el caso de que no exista el tipo en español, es decir, la etiqueta “es”, se pondrán en inglés).

Ahora mismo a partir, luego para unificar la información, creamos un fichero de texto (“trendingTopics.txt”) donde se ha escrito por pantalla la fecha y hora, el trending topic, la entidad, el tipo que instancia esa entidad concreta y sus sinónimos.

En el caso de que una entidad no tenga sinónimos ni tipo, no aparecerán las etiquetas en la línea correspondiente del fichero, es decir solo aparecerá el trending topic y su entidad asociada, y si tampoco tiene entidad pues solo el trending topic.

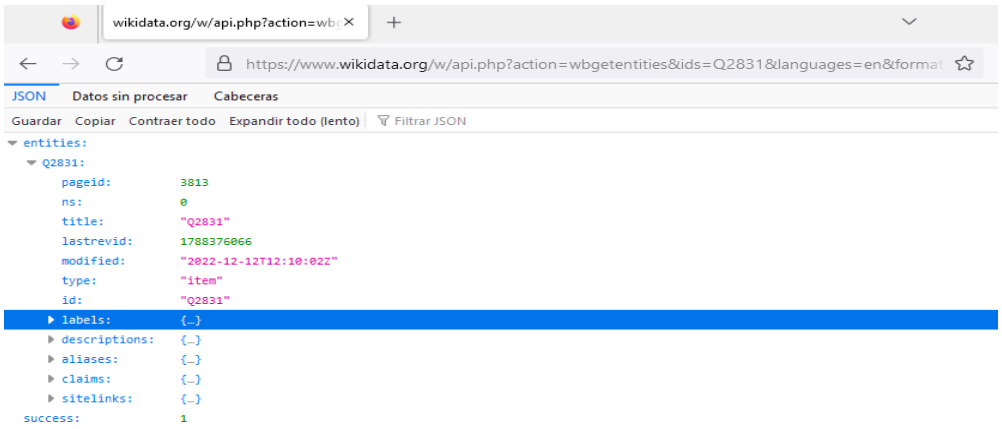
Ejemplo generado en el fichero:

Formato [fecha y hora, trendingTopic, entidad: ?, tipo: ?, sinónimos: ?]

```
fecha y hora: Wed Jun 24 20:00:00 +0000 2009, trending topic: usa, entidad: Q30, tipo: estado soberano, sinónimos: eua
```

En el siguiente fichero, se escribirán todos los trending topics con sus datos ordenados por fecha y hora.

Ejemplo de búsqueda de WikiData, entidad Q2831 (“Michael Jackson”):



Ya tendríamos todos los resultados pedidos en el entregable.

2. Entregable 2

2.1. ¿Cómo he encontrado los resultados?

Para generar un volcado con todos los tweets relativos al tema: venta de cristiano Ronaldo del Manchester United al Madrid, (la consulta inicial será “cristiano ronaldo”) mediante una consulta en un inglés, he utilizado en primer lugar, el fichero bulk-indexer4.py para generar el índice en inglés (con el 10% de los tuits, ya que considero que son suficientes para buscar información acerca del tema en cuestión). El volcado incluirá el autor, la fecha de creación y el texto del tuit en un fichero TSV.

El mecanismo utilizado es el siguiente, utilizar dos consultas con dos métricas diferentes (gnd y jlh), la explicación de su uso y del tema elegido se explicará más adelante. Hemos ido ejecutando estas consultas y viendo cuántos términos nos interesaban del total, es decir, que tuits contenían información más relevante, cuáles nos interesaban más para nuestro tema, es decir, evitar el mayor ruido posible.

```
"gnd": {} # en este caso, hemos escogido esta métrica
```

```
"jlh": {} # en este caso, hemos escogido esta métrica
```

Después de las consultas, realizar un escaneo de ambas, haciendo una consulta expandida con los términos más significativos (se utiliza agregación) de la anterior consulta, más el tema en cuestión. Para ello, añadimos en una lista todos los términos y los vamos acumulando en una cadena para pasársela a la query del escaneo.

```
resultadosGnd = [] # vamos añadiendo los términos significativos con dicha métrica
infoGnd = "(cristiano ronaldo)"
for elem in results["aggregations"]["Most significant users"]["buckets"]:
    if (str(elem["key"]) not in resultadosGnd):
        resultadosGnd.append(str(elem["key"]))
        infoGnd = infoGnd + " OR (" + str(elem["key"]) + ")" # añadimos el término a la cadena para escanearlo
```

```
"query_string" : {
    "query": infoGnd, # le pasamos los datos para expandir la consulta
    "default_operator" : "AND" # los espacios en blanco sino los interpreta como operadores OR
}
```

También hemos incluido en el escaneo el operador por defecto AND, ya que, sino elastic los espacios en blanco entre términos los asume como un OR, y no es lo que queremos, queremos información en conjunto de los términos, no información de cada uno por separado.

A partir de escanear, obtenemos la fecha, autor y tuit a partir de los hits de este, y los escribimos en un fichero tsv. De esta forma, se crean dos volcados (cada uno con su métrica) lo más exhaustivos posibles.

```
# abrimos el primer fichero
file=open("terminosSignificativosGnd.tsv","w") # creamos un

# para cada resultado guardamos el autor, la fecha de creac
for hit in escaneo:
    fecha = str(hit["_source"]["created_at"])
    autor = str(hit["_source"]["user_id_str"])
    tuit = str(hit["_source"]["text"])
    file.write(fecha + "\t" + autor + "\t" + tuit + "\n")

file.close() # cerramos el fichero
```

Ahora analizaremos cada fichero con los tuits (20 de cada métrica, los 20 primeros), y aportaremos una reflexión crítica sobre los resultados de ambas métricas y distintos números de términos. Lo suyo sería aportar una muestra aleatoria, pero no se especifica en ningún lado cómo realizarlo, así que he optado por esta estrategia.

También vamos a realizar una evaluación sistemática de los resultados, incluyendo la precisión lograda con cada configuración (en porcentaje), ya que es un tema subjetivo. En este caso no es buena idea hacer una selección de más de 20 documentos.

2.2. Detalles sobre los resultados y el tema inicial

En primer lugar, he elegido el tema de la venta de Cristiano Ronaldo del Manchester United al Real Madrid, ya que es un tema que me interesa mucho y tuvo mucha repercusión en su día, luego pensé que iba a haber mucha información acerca de ello, y he confirmado la hipótesis buscando términos significativos relativos a “cristiano ronaldo” (consulta inicial es “cristiano ronaldo”) en elasticvue. Me han salido muchos resultados, la idea inicial era buscar “cristiano ronaldo deal” pero no me salía ningún resultado ya que es una búsqueda muy concreta, luego se tendrá que poner simplemente “cristiano Ronaldo” y concretar con la expansión de consultas. El idioma elegido es el inglés, ya que en este idioma hay mucha más información que en español, y uno de los equipos es de Inglaterra, luego habrá más tuits acerca de ello.

Para la primera métrica (gnd) he elegido que el tamaño máximo de términos significativos sea 22, ya que a partir de aquí empiezan a haber mucho ruido y hay términos muy repetitivos; por ejemplo: ya ha salido el topic “star cristiano ronaldo” y justo después aparece”unit star cristiano” o “six year real” por “agree six year” que significa lo mismo, puesto que ha habido un acuerdo con el Real Madrid, que es el club que lo ficha. Se han incluido los términos marcados con una flecha en la captura y se han desechado aquellos con una cruz, no se han escogido a mano, simplemente se han dejado fuera con el número máximo de términos.

Para la segunda métrica (jlh) he elegido que el tamaño máximo de términos significativos sea 14, ya que a partir de aquí empiezan a haber mucho ruido y hay términos muy repetitivos y datos que no nos interesan.

En la consulta expandida de ambas métricas, estos son los términos para buscar en el escaneo:

Topics con la métrica gnd:

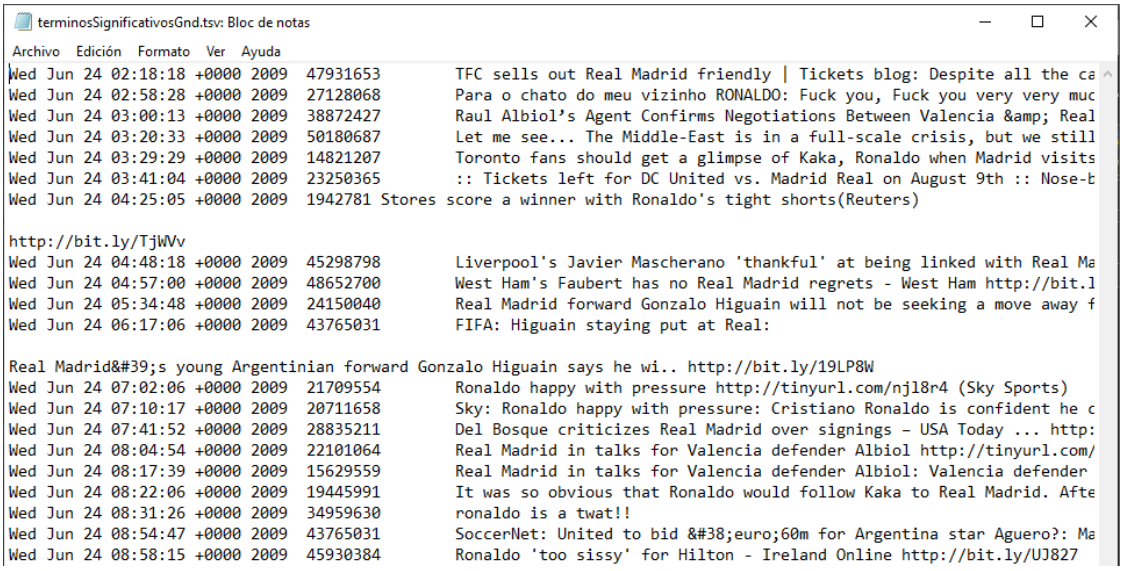
```
(cristiano ronaldo) OR (cristiano ronaldo) OR (cristiano) OR (cristiano ronaldo _) OR (ronaldo _) OR (ronaldo) OR (_
cristiano ronaldo) OR (_ cristiano) OR (_ real madrid) OR (ronaldo _ _) OR (deal cristiano) OR (unit star cristiano) OR
(year real deal) OR (real deal cristiano) OR (agre six) OR (six year real) OR (manchest unit star) OR (done manchest unit)
OR (deal cristiano ronaldo) OR (star cristiano) OR (_ done manchest) OR (ronaldo agre six) OR (star cristiano ronaldo)
```

Topics con la métrica mutual_information:

```
(cristiano ronaldo) OR (cristiano ronaldo) OR (cristiano) OR (cristiano ronaldo _) OR (ronaldo) OR (ronaldo _) OR (_
cristiano ronaldo) OR (_ cristiano) OR (ronaldo _ _) OR (manchest unit star) OR (year real deal) OR (deal cristiano ronaldo)
OR (deal cristiano) OR (real deal cristiano) OR (star cristiano)
```

Podemos comprobar a simple vista, que ambas métricas devuelven resultados distintos, esto se podrá reflejar en la selección final de tuits y la valoración de porcentaje de documentos relevantes.

Para ello hemos hecho una elección de los primeros 20 tuits de cada volcado, podemos ver que el contenido es distinto, puesto que hemos aplicado dos métricas diferentes y elegido un cierto tamaño de términos en la consulta inicial.

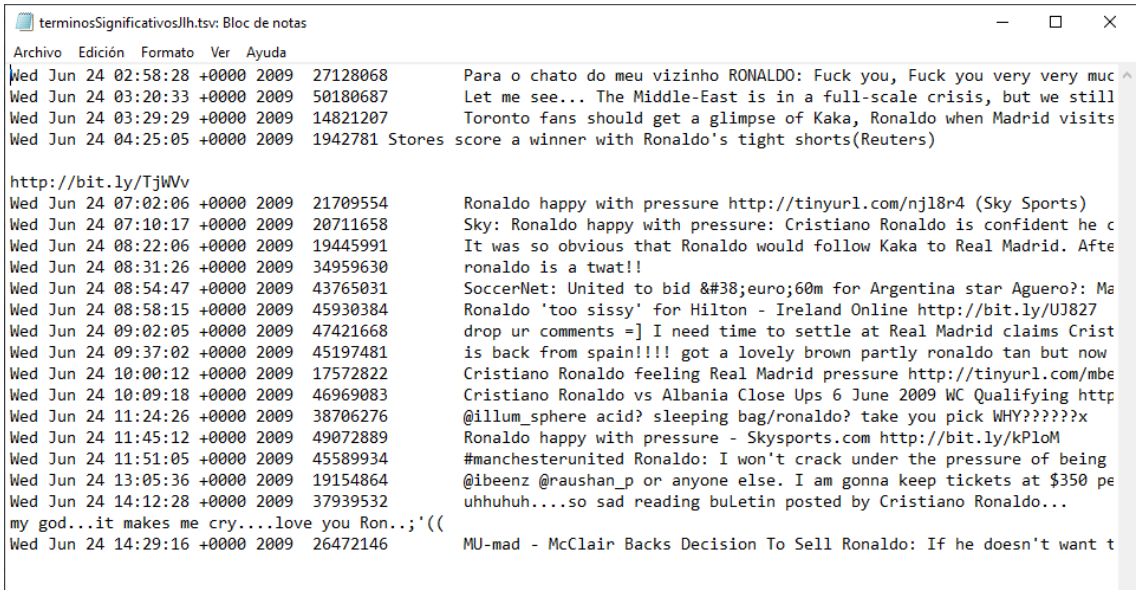

terminosSignificativosGnd.tsv: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
Wed Jun 24 02:18:18 +0000 2009  47931653        TFC sells out Real Madrid friendly | Tickets blog: Despite all the ca
Wed Jun 24 02:58:28 +0000 2009  27128068        Para o chato do meu vizinho RONALDO: Fuck you, Fuck you very very muc
Wed Jun 24 03:00:13 +0000 2009  38872427        Raul Albiol's Agent Confirms Negotiations Between Valencia &amp;amp; Real
Wed Jun 24 03:20:33 +0000 2009  50180687        Let me see... The Middle-East is in a full-scale crisis, but we still
Wed Jun 24 03:29:29 +0000 2009  14821207        Toronto fans should get a glimpse of Kaka, Ronaldo when Madrid visits
Wed Jun 24 03:41:04 +0000 2009  23250365        :: Tickets left for DC United vs. Madrid Real on August 9th :: Nose-t
Wed Jun 24 04:25:05 +0000 2009  1942781 Stores score a winner with Ronaldo's tight shorts(Reuters)

http://bit.ly/TjWVv
Wed Jun 24 04:48:18 +0000 2009  45298798        Liverpool's Javier Mascherano 'thankful' at being linked with Real Ma
Wed Jun 24 04:57:00 +0000 2009  48652700        West Ham's Faubert has no Real Madrid regrets - West Ham http://bit.l
Wed Jun 24 05:34:48 +0000 2009  24150040        Real Madrid forward Gonzalo Higuain will not be seeking a move away f
Wed Jun 24 06:17:06 +0000 2009  43765031        FIFA: Higuain staying put at Real:

Real Madrid&#39;s young Argentinian forward Gonzalo Higuain says he wi.. http://bit.ly/19LP8W
Wed Jun 24 07:02:06 +0000 2009  21709554        Ronaldo happy with pressure http://tinyurl.com/nj18r4 (Sky Sports)
Wed Jun 24 07:10:17 +0000 2009  20711658        Sky: Ronaldo happy with pressure: Cristiano Ronaldo is confident he c
Wed Jun 24 07:41:52 +0000 2009  28835211        Del Bosque criticizes Real Madrid over signings – USA Today ... http:
Wed Jun 24 08:04:54 +0000 2009  22101064        Real Madrid in talks for Valencia defender Albiol http://tinyurl.com/
Wed Jun 24 08:17:39 +0000 2009  15629559        Real Madrid in talks for Valencia defender Albiol: Valencia defender
Wed Jun 24 08:22:06 +0000 2009  19445991        It was so obvious that Ronaldo would follow Kaka to Real Madrid. Afte
Wed Jun 24 08:31:26 +0000 2009  34959630        ronaldo is a twat!!
Wed Jun 24 08:54:47 +0000 2009  43765031        SoccerNet: United to bid &#38;euro;60m for Argentina star Aguero?: Ma
Wed Jun 24 08:58:15 +0000 2009  45930384        Ronaldo 'too sissy' for Hilton - Ireland Online http://bit.ly/UJ827
```



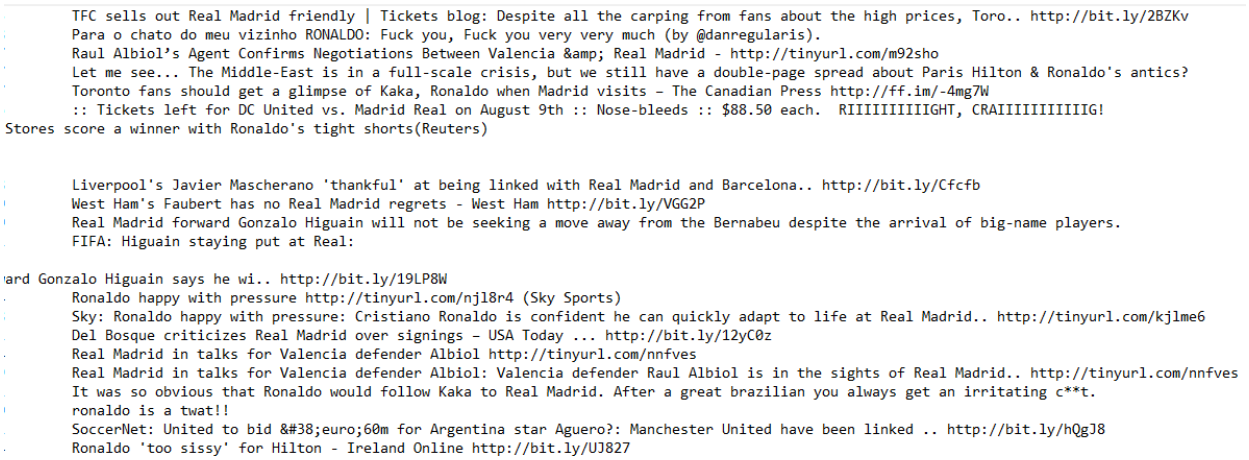
terminosSignificativosJlh.tsv: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
Wed Jun 24 02:58:28 +0000 2009  27128068        Para o chato do meu vizinho RONALDO: Fuck you, Fuck you very very muc
Wed Jun 24 03:20:33 +0000 2009  50180687        Let me see... The Middle-East is in a full-scale crisis, but we still
Wed Jun 24 03:29:29 +0000 2009  14821207        Toronto fans should get a glimpse of Kaka, Ronaldo when Madrid visits
Wed Jun 24 04:25:05 +0000 2009  1942781 Stores score a winner with Ronaldo's tight shorts(Reuters)

http://bit.ly/TjWVv
Wed Jun 24 07:02:06 +0000 2009  21709554        Ronaldo happy with pressure http://tinyurl.com/njl8r4 (Sky Sports)
Wed Jun 24 07:10:17 +0000 2009  20711658        Sky: Ronaldo happy with pressure: Cristiano Ronaldo is confident he c
Wed Jun 24 08:22:06 +0000 2009  19445991        It was so obvious that Ronaldo would follow Kaka to Real Madrid. Afte
Wed Jun 24 08:31:26 +0000 2009  34959630        ronaldo is a twat!!
Wed Jun 24 08:54:47 +0000 2009  43765031        SoccerNet: United to bid &#38;euro;60m for Argentina star Aguero?: Ma
Wed Jun 24 08:58:15 +0000 2009  45930384        Ronaldo 'too sissy' for Hilton - Ireland Online http://bit.ly/UJ827
Wed Jun 24 09:02:05 +0000 2009  47421668        drop ur comments =] I need time to settle at Real Madrid claims Crist
Wed Jun 24 09:37:02 +0000 2009  45197481        is back from spain!!!! got a lovely brown partly ronaldo tan but now
Wed Jun 24 10:00:12 +0000 2009  17572822        Cristiano Ronaldo feeling Real Madrid pressure http://tinyurl.com/mbe
Wed Jun 24 10:09:18 +0000 2009  46969083        Cristiano Ronaldo vs Albania Close Ups 6 June 2009 WC Qualifying http
Wed Jun 24 11:24:26 +0000 2009  38706276        @illum_sphere acid? sleeping bag/ronaldo? take you pick WHY??????x
Wed Jun 24 11:45:12 +0000 2009  49072889        Ronaldo happy with pressure - Skysports.com http://bit.ly/kPloM
Wed Jun 24 11:51:05 +0000 2009  45589934        #manchesterunited Ronaldo: I won't crack under the pressure of being
Wed Jun 24 13:05:36 +0000 2009  19154864        @ibeenz @raushan_p or anyone else. I am gonna keep tickets at $350 pe
Wed Jun 24 14:12:28 +0000 2009  37939532        uhhuhuh....so sad reading buLetin posted by Cristiano Ronaldo...
my god...it makes me cry....love you Ron..;'((
Wed Jun 24 14:29:16 +0000 2009  26472146        MU-mad - McClair Backs Decision To Sell Ronaldo: If he doesn't want t
```

Vamos a analizar el primer volcado, el de términos significativos con gnd:



```
        TFC sells out Real Madrid friendly | Tickets blog: Despite all the carping from fans about the high prices, Toro.. http://bit.ly/2BZKv
        Para o chato do meu vizinho RONALDO: Fuck you, Fuck you very very much (by @danregularis).
        Raul Albiol's Agent Confirms Negotiations Between Valencia &amp;amp; Real Madrid - http://tinyurl.com/m92sho
        Let me see... The Middle-East is in a full-scale crisis, but we still have a double-page spread about Paris Hilton & Ronaldo's antics?
        Toronto fans should get a glimpse of Kaka, Ronaldo when Madrid visits - The Canadian Press http://ff.im/-4mg7W
        :: Tickets left for DC United vs. Madrid Real on August 9th :: Nose-bleeds :: $88.50 each.  RIIIIIIIIIIGHT, CRAIIIIIIIIIIG!
Stores score a winner with Ronaldo's tight shorts(Reuters)


        Liverpool's Javier Mascherano 'thankful' at being linked with Real Madrid and Barcelona.. http://bit.ly/Cfcfb
        West Ham's Faubert has no Real Madrid regrets - West Ham http://bit.ly/VGG2P
        Real Madrid forward Gonzalo Higuain will not be seeking a move away from the Bernabeu despite the arrival of big-name players.
        FIFA: Higuain staying put at Real:

ard Gonzalo Higuain says he wi.. http://bit.ly/19LP8W
        Ronaldo happy with pressure http://tinyurl.com/njl8r4 (Sky Sports)
        Sky: Ronaldo happy with pressure: Cristiano Ronaldo is confident he can quickly adapt to life at Real Madrid.. http://tinyurl.com/kjlme6
        Del Bosque criticizes Real Madrid over signings - USA Today ... http://bit.ly/12yC0z
        Real Madrid in talks for Valencia defender Albiol http://tinyurl.com/nnfves
        Real Madrid in talks for Valencia defender Albiol: Valencia defender Raul Albiol is in the sights of Real Madrid.. http://tinyurl.com/nnfves
        It was so obvious that Ronaldo would follow Kaka to Real Madrid. After a great brazilian you always get an irritating c**t.
        ronaldo is a twat!!
        SoccerNet: United to bid &#38;euro;60m for Argentina star Aguero?: Manchester United have been linked .. http://bit.ly/hQgJ8
        Ronaldo 'too sissy' for Hilton - Ireland Online http://bit.ly/UJ827
```

En primer lugar, vamos a explicar por qué un tuit es pertinente o no, los criterios para aceptar un tuit son muy sencillos, que se hable acerca del fichaje de Cristiano Ronaldo del Manchester United al Madrid, no nos vale que hablen solo de compañeros de fútbol, ni de ambos clubes por separado sin la figura de Cristiano Ronaldo.

El primer tuit es irrelevante, ya que solo habla del Real Madrid sin la figura de Cristiano Ronaldo. El segundo no habla de Cristiano Ronaldo directamente, habla de otro Ronaldo. El tercero habla sobre el Real Madrid y un posible fichaje, pero no es el de Cristiano Ronaldo luego tampoco. En el 4, Ronaldo está fuera de contexto del fútbol. En el quinto la información empieza a ser más importante pero no es lo que queremos, ya que Kaka y Ronaldo van a ser compañeros en el Madrid, pero no comenta nada del fichaje. En el sexto no, ya que habla sobre un partido entre el Madrid y United. El séptimo está fuera de contexto. El octavo y el noveno hablan del Madrid, pero no nos interesa el tema.

El décimo y el undécimo sí que son relevantes, ya que está diciendo que un jugador del Madrid no va a abandonar el club aún la llegada de grandes estrellas (refiriéndose al jugador Cristiano Ronaldo). El tuit doce y trece también tienen que ver, ya que comenta la presión de Cristiano Ronaldo al llegar al Madrid, y que se va a adaptar a su nueva vida muy bien en Madrid.

Los tuits 14,15 y 16 hablan sobre el Real Madrid en general, pero no comentan nada importante.

El tuit 17 también es relevante, ya que está diciendo que Cristiano Ronaldo a seguido los pasos de Kaka, un jugador actual del Real Madrid.

El tuit 18 no es relevante, el 19 habla sobre el United y Agüero (otro jugador de fútbol) y el 20 sobre Ronaldo.

Luego de todos lo tuits analizados (20) son relevantes 5, es decir el 25% de los tuits hablan acerca de Cristiano Ronaldo y su llegada al Madrid del United. Por lo que, tenemos una **precisión del 25%** en la búsqueda acerca del fichaje del jugador por el Real Madrid, uno de cada cuatro tuits es relevante.

Vamos a analizar el segundo volcado, el de términos significativos con jlh:

```
Para o chato do meu vizinho RONALDO: Fuck you, Fuck you very very much (by @danregularis).
Let me see... The Middle-East is in a full-scale crisis, but we still have a double-page spread about Paris Hilton & Ronaldo's antics?
Toronto fans should get a glimpse of Kaka, Ronaldo when Madrid visits – The Canadian Press http://ff.im/-4mg7W
score a winner with Ronaldo's tight shorts(Reuters)


Ronaldo happy with pressure http://tinyurl.com/njl8r4 (Sky Sports)
Sky: Ronaldo happy with pressure: Cristiano Ronaldo is confident he can quickly adapt to life at Real Madrid.. http://tinyurl.com/kjlme6
It was so obvious that Ronaldo would follow Kaka to Real Madrid. After a great brazilian you always get an irritating c**t.
ronaldo is a twat!!
SoccerNet: United to bid &#38;euro;60m for Argentina star Aguero?: Manchester United have been linked .. http://bit.ly/hQgJ8
Ronaldo 'too sissy' for Hilton - Ireland Online http://bit.ly/UJ827
drop ur comments =] I need time to settle at Real Madrid claims Cristiano Ronaldo: Cristiano .. http://u.mavrev.com/dpva
is back from spain!!!! got a lovely brown partly ronaldo tan but now needs to watch wimbledon and catch up on TV!!!!! coach trip here i come
Cristiano Ronaldo feeling Real Madrid pressure http://tinyurl.com/mbep3r
Cristiano Ronaldo vs Albania Close Ups 6 June 2009 WC Qualifying http://bit.ly/Njmg6
@illum_sphere acid? sleeping bag/ronaldo? take you pick WHY??????x
Ronaldo happy with pressure - Skysports.com http://bit.ly/kPloM
#manchesterunited Ronaldo: I won't crack under the pressure of being Real's £80m man: Image: ht.. http://tinyurl.com/kjt9vq
@ibeenz @raushan_p or anyone else. I am gonna keep tickets at $350 per seat, but prices are going up b/c Kaka/Ronaldo are guaranteed..holla!
uhhuhuh....so sad reading buLetin posted by Cristiano Ronaldo...

MU-mad - McClair Backs Decision To Sell Ronaldo: If he doesn't want to stay then he has to go http://tinyurl.com/mc9w61
```

En primer lugar, vamos a explicar por qué un tuit es pertinente o no, los criterios para aceptar un tuit son muy sencillos, que se hable acerca del fichaje de Cristiano Ronaldo del Manchester United al Madrid, no nos vale que hablen solo de compañeros de fútbol, ni de ambos clubes por separado sin la figura de Cristiano Ronaldo.

El primer tuit no habla de Cristiano Ronaldo directamente, habla de otro Ronaldo. El segundo es otro Ronaldo, luego irrelevante de momento.

El tercero habla sobre los fanáticos de Toronto deberían echar un vistazo a Kaká, Ronaldo cuando Madrid “The Canadian”, luego relevante, ya se está considerando a Cristiano Ronaldo como jugador del Madrid y se está asociando su persona con la de un jugador del club y el propio club.

El cuarto no es importante ya que habla de otro Ronaldo.

El tuit cinco y seis tienen que ver, ya que comenta la presión de Cristiano Ronaldo al llegar al Madrid, y que se va a adaptar a su nueva vida muy bien en Madrid. El tuit siete podría ser relevante, ya que expone que está claro que Cristiano Ronaldo va a seguir los pasos del jugador brasileño Kaka, un jugador actual del Real Madrid.

El tuit 8 está fuera de contexto, el 9 habla sobre el United y Agüero (otro jugador de fútbol) y el 10 sobre Ronaldo.

El tuit once es relevante, hay que considerarlo porque habla sobre la llegada de Cristiano Ronaldo al Real Madrid, diciendo que necesita tiempo para asentarse.

El tuit doce está fuera de contexto, porque menciona un estadio de tenis “Wimbledon”.

En el trece y dieciséis se comenta la presión de la llegada del jugador al club español, luego hay que considerarlos importantes.

En el tuit 14 se comenta un partido de Cristiano Ronaldo contra Albania, luego irrelevante. Y en el 15 habla sobre otro Ronaldo.

El tuit 17 es muy importante, quizás el más relevante, ya que habla sobre la venta de Cristiano Ronaldo al Real Madrid, el precio que costó y unas declaraciones de él acerca de la presión. También el 18 que se comenta el precio de una entrada para ver a Cristiano Ronaldo y a Kaka (un jugador del Madrid) un partido del Real Madrid.

El 19 habla de Cristiano Ronaldo, pero no comenta nada de ningún equipo ni la venta.

Y en el 20, podríamos considerar la información relevante, ya que habla sobre que fue buena decisión vender a Cristiano Ronaldo por parte del Manchester United, ya que no quería quedarse en el equipo, quería marchar.

Luego de todos lo tuits analizados (20) son relevantes 10, es decir el 50% de los tuits hablan acerca de Cristiano Ronaldo y su llegada al Madrid del United. Por lo que, tenemos una **precisión del 50%** en la búsqueda acerca del fichaje del jugador por el Real Madrid, uno de cada dos tuits es relevante.

En resumen, ¿qué métrica es mejor y por qué?

En este caso, como podemos comprobar es mejor utilizar la métrica jlh, ya que proporciona una mayor fiabilidad de resultados en la búsqueda del topic “cristiano ronaldo” [venta de Cristiano Ronaldo del Manchester United al Real Madrid]. Tenemos un 50% de efectividad contra un 25% de gnd (el doble de precisión).

¿Por qué he elegido una métrica u otra?

La métrica **gnd** selecciona los términos significativos con mayor frecuencia y evita la selección de palabras vacías; **jlh** selecciona aquellos términos de alta frecuencia si también ocurren con frecuencia en segundo plano, además que sería muy poco probable que esta métrica seleccione errores ortográficos.