UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

# IDENTIFICACIÓN DE CONTENIDO MULTIMEDIA RELEVANTE A PARTIR DE EVENTOS UTILIZANDO SU INFORMACIÓN SOCIAL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

MAURICIO DANIEL QUEZADA VEAS

PROFESORA GUÍA:
BÁRBARA POBLETE LABRA

MIEMBROS DE LA COMISIÓN:
SERGIO OCHOA DELORENZI
MAURICIO MARÍN CAIHUAN

SANTIAGO DE CHILE
ENERO 2013

usuario puede *seguir* a otros para poder ver en su *timeline* o perfil privado los mensajes de todos a quienes sigue.

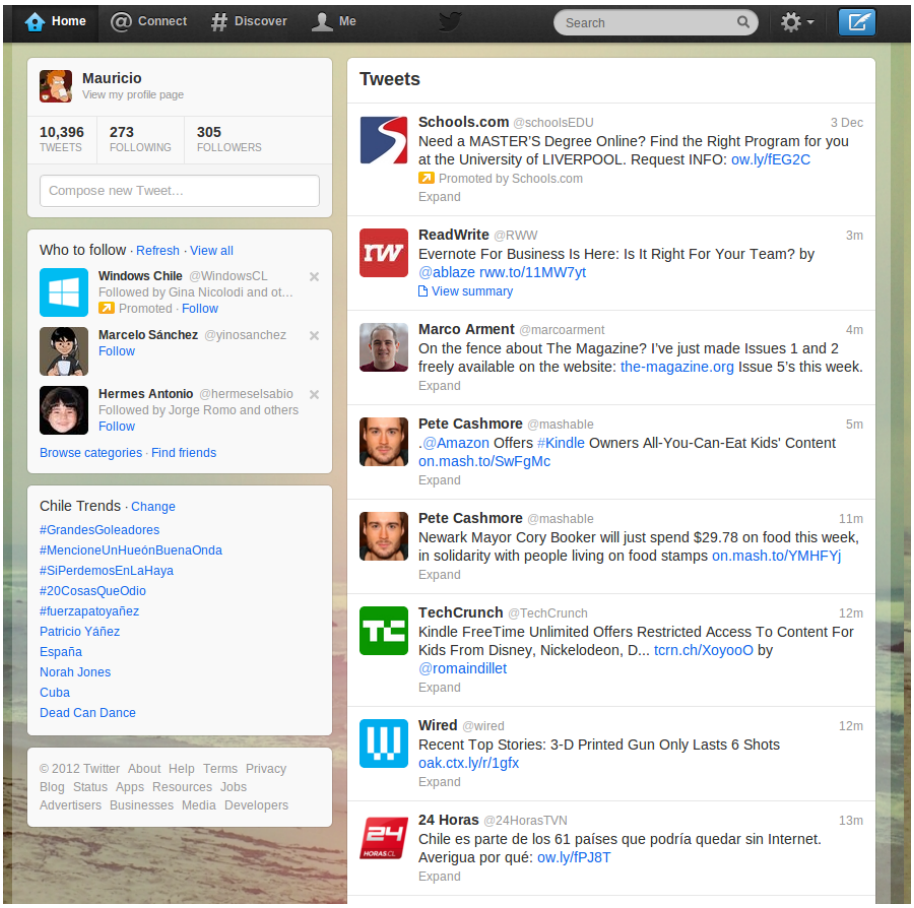


Figura 2.1: Timeline de Twitter. En éste se ve una lista en orden cronológico de los tweets generados por los usuarios que sigue el usuario actual. Además, el sitio incluye tweets promocionados por cuentas que pagan por dicho servicio, como es el caso del primer tweet en la lista.

Estos mensajes, o *tweets*, sólo son cadenas de caracteres con metadatos que el mismo servicio asigna una vez enviado a la red social. Desde sus inicios (año 2007) se han añadido algunas capacidades adicionales a estos mensajes, como la de poner URLs, imágenes, vídeos, etc. Además, existen varias convenciones que han surgido a lo largo del tiempo. A continuación se describe una lista de tipos de mensajes que existen en Twitter, originados por estas convenciones:

1. Respuestas o *replies*: son mensajes del tipo `@usuario [texto]`, que ocurren usualmente en una conversación entre dos usuarios.
2. Menciones o *mentions*: un poco más general a una respuesta, el nombre del usuario mencionado puede estar en cualquier parte del mensaje. La diferencia semántica es que no se le habla "directamente" al usuario mencionado, como en una respuesta, sino que sólo es mencionado por si el mensaje es de su interés o no.
3. *Retweets*: son mensajes del tipo `RT @usuario: [texto]`. Ocurren cuando se quiere compartir el mensaje de otro usuario, o citarlo para mencionarlo en el mismo mensaje.
4. *Hashtags*: son palabras precedidas por el caracter #, que indican un identificador a cierto

$$E(C_r) = -\frac{1}{\log q} \sum_{\mathrm{i}=1}^{q} \frac{n_r^{\mathrm{i}}}{n_r} \log \frac{n_r^{\mathrm{i}}}{n_r}$$

donde $q$ es el número de clases en el dataset, y $n_r^{\mathrm{i}}$ es el número de documentos de la i-ésima clase que fue asignado al $r$-ésimo cluster $C_r$. La entropía del clustering se define como la suma ponderada de la entropía de cada cluster:

$$\text{Entropia} = \sum_{r=1}^{k} \frac{n_r}{n} E(C_r)$$

Un clustering perfecto tendrá clusters tal que cada cluster contenga documentos de una sola clase, en ese caso la entropía será 0. En general, conviene tener bajos valores de entropía.

- **Pureza**: mide la cantidad de documentos de la clase más grande en un cluster dividida por el tamaño del cluster. La pureza de un cluster $C_r$ se define como:

$$P(C_r) = \frac{1}{n_r} \max_{\mathrm{i}} \{n_r^{\mathrm{i}}\}$$

A mayor pureza, se considera que mejor es la solución.

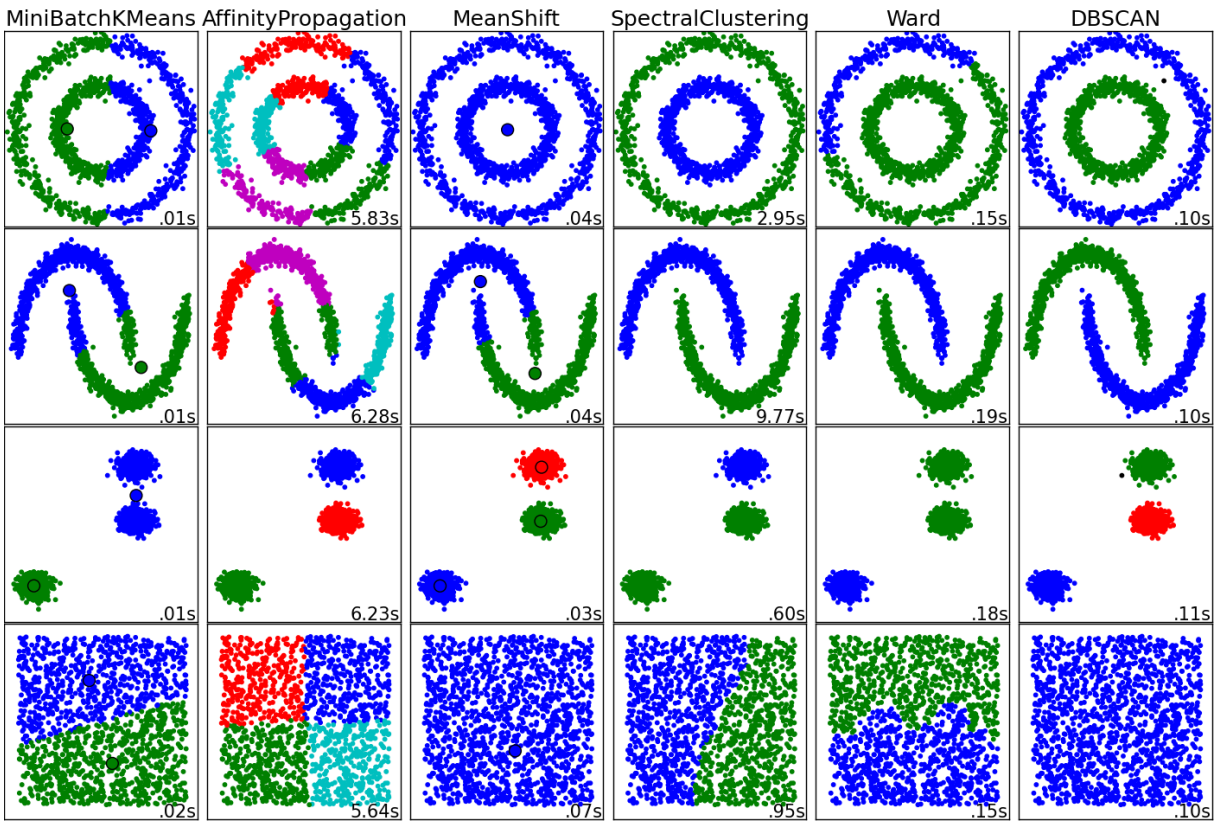


Figura 2.2: Comparación de algoritmos de clustering. En la figura se muestran las características de distintos algoritmos de clustering sobre datos en 2 dimensiones. El último caso muestra datos uniformes en el cual sólo habría 1 cluster.

```
{
''850a0f7e08e9cf2e080678679857eef9'': {
  ''domain'': ''twitter'',
  ''is_retweet'': [0, 0, 0, 0, 0, 0],
  ''num_tweets'': 6,
  ''retweets'': [''0'', ''0'', ''0'', ''0'', ''0'', ''0''],
  ''tweets_lengths'': [8, 6, 8, 6, 6, 8],
  ''url'': ''https://api.twitter.com/1/statuses/show.json?id=270299718540738560'',
  ''user_created_at'': [1309224895.0,
                        1285052229.0,
                        1309224895.0,
                        1285052229.0,
                        1285052229.0,
                        1309224895.0],
  ''user_followers'': [''45'', ''37'', ''45'', ''37'', ''37'', ''45''],
  ''user_friends'': [''2'', ''25'', ''2'', ''25'', ''25'', ''2''],
  ''user_geo_enabled'': [1, 0, 1, 0, 0, 1],
  ''user_is_verified'': [0, 0, 0, 0, 0, 0],
  ''user_lists'': [''0'', ''1'', ''0'', ''1'', ''1'', ''0''],
  ''user_statuses'': [''12484'', ''4756'', ''12484'', ''4756'', ''4756'', ''12484'']}
}
```


Código 1: Información de un documento, correspondiente al evento "Anef anuncia movilización nacional". Los campos que corresponden a listas indican los valores para cada tweet del documento, en este caso, el documento tiene 6 tweets; por ejemplo, user_followers[24]=45 indica la cantidad de seguidores que tiene el autor del tweet en la tercera posición.

- Cantidad de *retweets* hechos hasta la fecha;
- Si posee alguna URL o *hashtag* en el texto;
- Si es una *mención* a otro usuario;
- La ubicación de donde se envió el tweet;
- etc.

Además, incluye datos sobre el autor, como por ejemplo:

- Si la cuenta está *verificada*;
- La cantidad de seguidores del usuario;
- Cantidad de *amigos* (usuarios que siguen y son seguidos por el usuario);
- Cantidad de tweets;
- Su descripción, y si incluye alguna URL, etc;
- Ubicación (dada por el mismo usuario);
- Fecha de creación de la cuenta;
- etc.

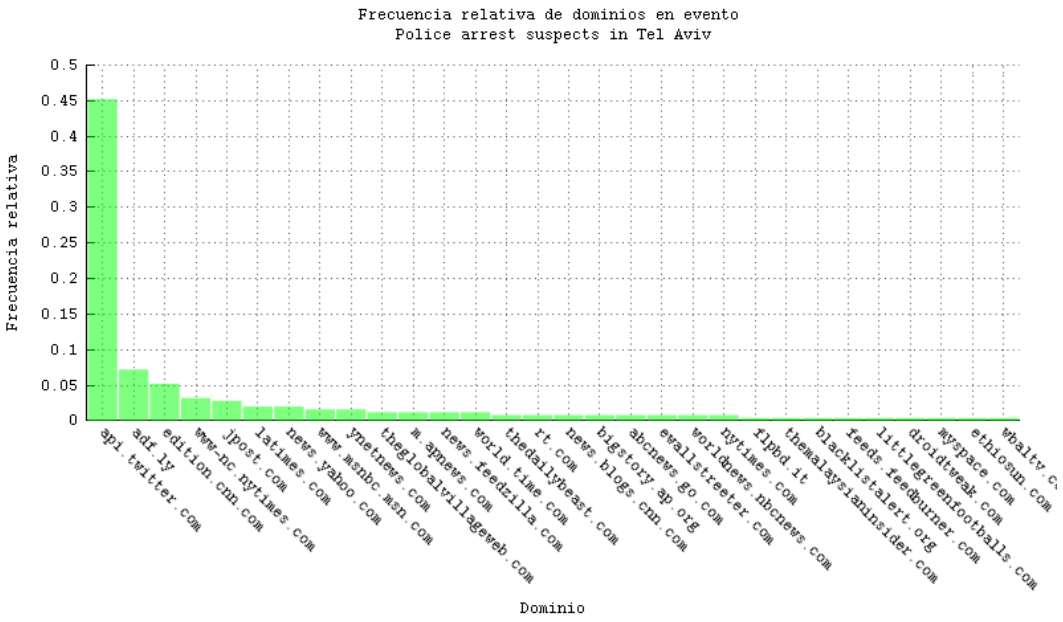


Figura 4.3: Distribución de frecuencias de dominios del evento “Police Arrest suspects in Tel Aviv”. Las frecuencias son relativas al total de documentos correspondientes al evento. Se puede apreciar rápidamente que la mayoría de los documentos tienen enlaces a Twitter.

o expandir las URLs que aparecen en los tweets. Como las URLs acortadas con `adf.ly` no tienen una redirección directa, al tratar de expandir estas URLs no fue posible recuperar destino terminal. Es posible suponer que los destinos de estas direcciones apuntan a más sitios de noticias, dadas los dominios que siguen a continuación, como CNN o NY Times.

- De la misma forma, tanto para los eventos “Clinton” y “Stockholm”, en las Figuras 4.4 y 4.5, la distribución de dominios sigue una tónica similar: la gran mayoría de documentos corresponden a un tweet de sólo texto, sin URL. Es posible que gran parte de los tweets hayan correspondido a texto debido a los términos de búsqueda que fueron utilizados para recuperarlos, tales como `clinton`, `ceasefire`, `effort`, o `stockholm`. Por una parte, los eventos “Clinton” y “Tel Aviv” son noticias, mientras que “Stockholm” es un festival de música. Los dos primeros fueron temas con amplia cobertura en el momento en que ocurrieron, y a su vez, muy comentados en Twitter, mientras que la gran cantidad de tweets obtenidos para el tercero pudo significar que muchos de estos tweets sólo hayan contenido texto.

- Un caso especial ocurre para el evento “New York” en la Figura 4.6. No hay una mayoría de documentos apuntando a Twitter: de hecho, la gran mayoría de los documentos posee una URL distinta, dada la proporción a la que se encuentran los dominios más frecuentes. Esto pudo darse debido a que al no ser un evento con mucha cobertura (comparado con los otros), los resultados obtenidos de la búsqueda de tweets fueron mucho más precisos, al no haber “ruido” de por medio.

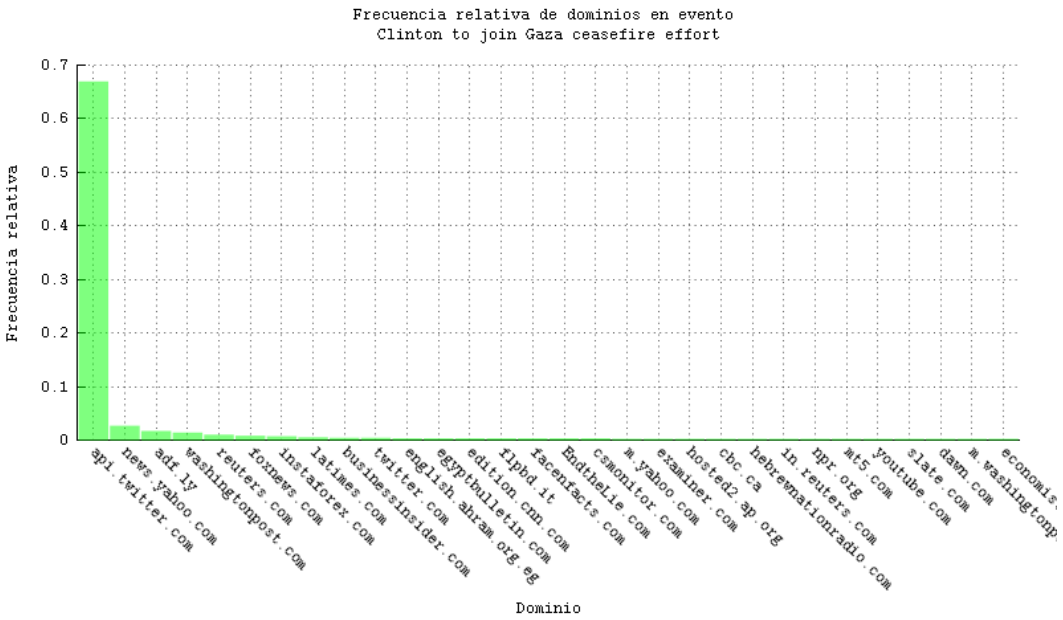


Figura 4.4: Distribución de frecuencias de dominios del evento "Clinton to join Gaza ceasefire effort". Las frecuencias son relativas al total de documentos correspondientes al evento. Se puede apreciar rápidamente que la mayoría de los documentos tienen enlaces a Twitter.

### 4.4.2. Determinación del número de clusters

Para determinar un número adecuado de subtópicos (clusters) para cada evento se utilizaron métricas de evaluación internas, dado que no se contó con las clases reales de los datos obtenidos. Para esto, se utilizó el software Cluto[7], el cual ofrece varios algoritmos de clustering con medidas de evaluación tanto internas (similitud intra-cluster y inter-cluster) como externas (pureza y entropía).

La metodología utilizada para determinar el número de clusters fue la siguiente:

- Calcular una solución de clustering particional usando $k \in \{2, \ldots, 30\}$ clusters como parámetro.
- Para cada solución obtenida, documentar la similitud intra-cluster promedio y inter-cluster promedio.
- Determinar experimentalmente la solución con mayor similitud intra-cluster y menor similitud inter-cluster, calculando el radio entre estas dos medidas.

Con esto, se puede determinar experimentalmente un número de clusters para cada evento, y analizar los resultados obtenidos. En las Figuras 4.7, 4.8, 4.9, y 4.10 se pueden apreciar los resultados obtenidos en la determinación del número de clusters:

1. Para el evento "Tel Aviv" en la Figura 4.7 se determinó como $9$ un número adecuado de clusters. Como se mencionó anteriormente, este número de determina como el número

[7] http://glaros.dtc.umn.edu/gkhome/views/cluto

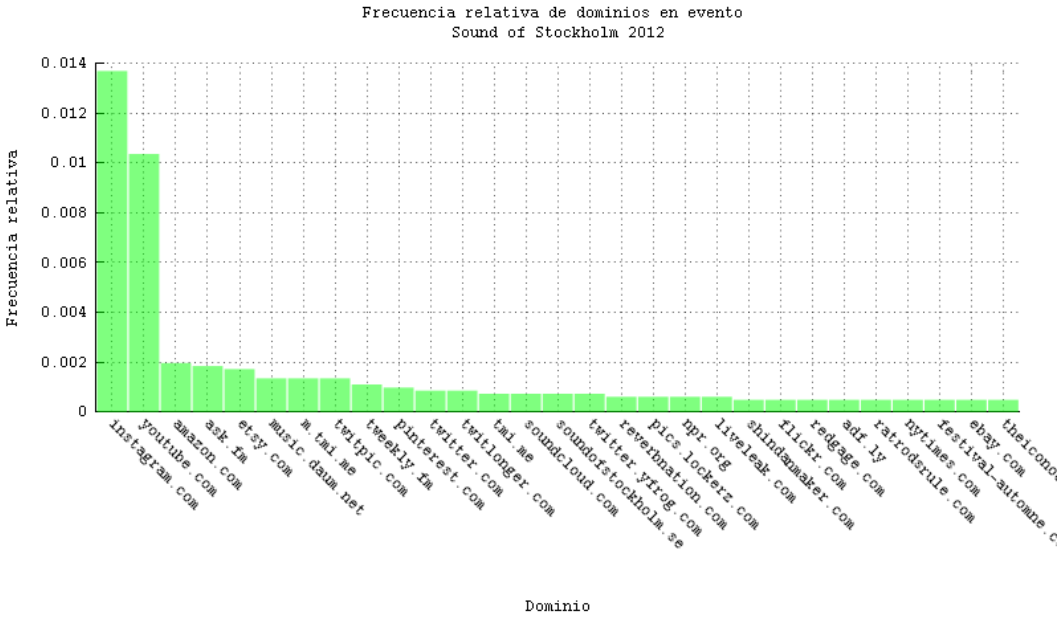


Figura 4.5: Distribución de frecuencias de dominios del evento “Sound of Stockholm 2012”. Las frecuencias son relativas al total de documentos correspondientes al evento. En este caso se quitó del gráfico lo que corresponde a Twitter, dejando los datos restantes. Se puede apreciar que los siguientes dominios representan no más del 10 % del total.

de clusters que entrega mayor similitud intra-cluster, y menor similitud inter-cluster. Éste corresponde a un máximo local, puesto que a mayor número de clusters (en el extremo, 1 cluster por documento) estas medidas pueden dar un óptimo global.

2. Para “Clinton” en la Figura 4.8 se determinó como 27 el número de clusters a utilizar. Como el radio siguió aumentando a medida que aumenta el número de clusters, posiblemente se trate de un máximo local.

3. Para “Stockholm” el máximo se encontró con 11 clusters en la Figura 4.9.

4. Finalmente, en la Figura 4.10, el evento “New York” presenta un máximo con 25 y 28 clusters. Fueron utilizados 25 clusters para el análisis.

Estas medidas indican que una cantidad adecuada de clusters es la obtenida, de acuerdo al criterio del radio de las similitudes. Sin embargo, esto no quiere decir necesariamente que la solución sea óptima incluso si se trata de un máximo global para cada caso, dado que esta medida sólo se basa en la similitud de los documentos, y no en el contenido de éstos. Por ejemplo, un evento puede tener todos sus documentos con tweets distintos, siendo que sólo hablen de dos subtópicos distintos; en ese caso el óptimo se encontraría a un número muy alto de clusters, siendo que bastaría con sólo dos. Por esto se realizó un análisis de los eventos caso por caso, sin suponer que el número de clusters determinado entregará una solución óptima.

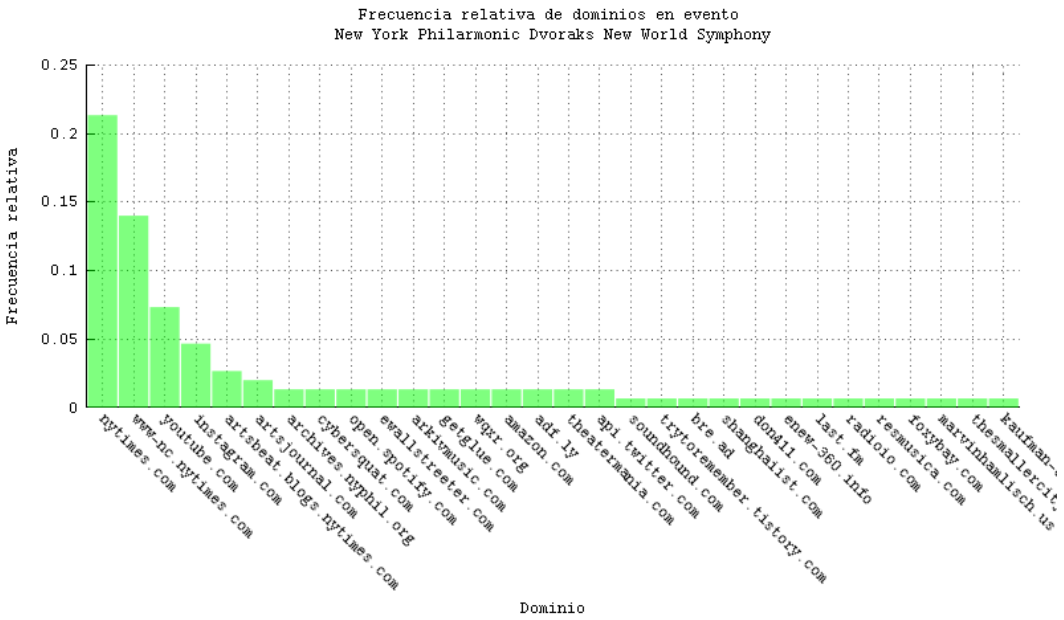


Figura 4.6: Distribución de frecuencias de dominios del evento "New York Philarmonic Dvorak's New World Symphony". Las frecuencias son relativas al total de documentos correspondientes al evento.

### 4.4.3. Análisis de resúmenes

Para analizar los resultados obtenidos, se generaron nuevamente clusterings utilizando el número estimado en la sección anterior, de forma de determinar una solución de mejor calidad. Una vez generados los clusterings, se procedió a hacer un análisis caso a caso de los resultados obtenidos.

Por ejemplo, para el evento "Tel Aviv", se determinó como 9 el número óptimo de clusters, dando como resultado una solución de clustering que particionaba los tweets de tal forma que cada cluster contenía principalmente sólo tweets de cierto tipo. A continuación se muestra el tweet más frecuente por cluster, o bien una descripción de los tweets del cluster, al no haber un tweet más repetido:

1. `Previous bomb attacks in Tel Aviv...`
2. `RT @BreakingNews: Israel's army spokesman says Israeli Arab arrested for Wednesday's bus bombing in Tel Aviv`
3. Este cluster contiene tweets de distintos temas, tanto de Tel Aviv como otros relacionados a Israel, Egipto o Estados Unidos, sin haber ninguno repetido en particular.
4. `RT @chaimlevinson: #breakingnews : after massive police hunt, 2 suspects in the tel aviv bus bombing arrested in 443 road`
5. `Police Arrest Suspects in Tel Aviv Bus Blast, Including Israeli Citizen:`
6. `Israel arrests suspects in Tel Aviv bus bombing: Israeli authorities arrested an Israeli Arab on suspicion of planting a bomb in a Te...`
7. `Arrest announced in Tel Aviv bus bombing: An arrest has been made in`

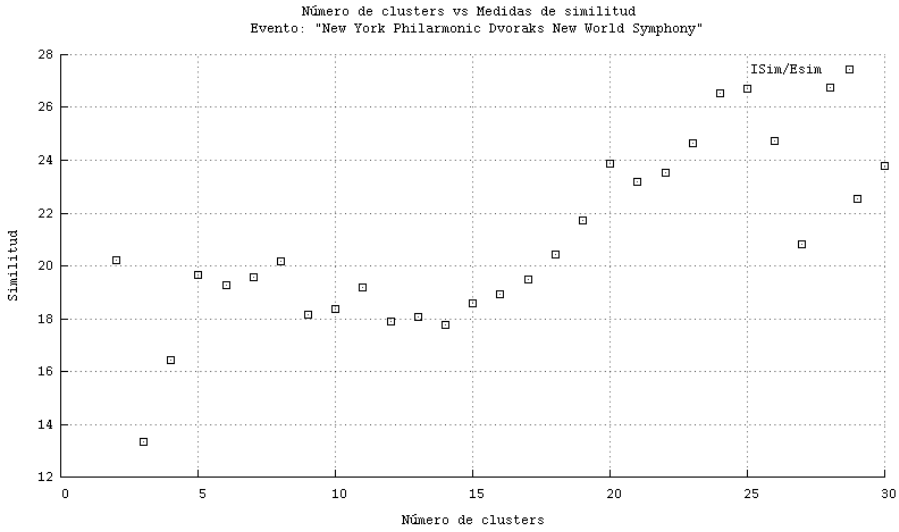


Figura 4.10: Radio ISim/ESim del evento "New York". Se aprecian dos máximos, con 25 y 28 clusters, respectivamente. Al ser máximos locales, es posible que a mayor número de clusters las soluciones puedan tener mejores medidas de similitud.


Figura 4.11: Documento (tweet) con más relevancia dentro del evento "Tel Aviv". Su relevancia dentro de los resultados se debe principalmente a que el autor posee una cuenta verificada en Twitter, además del número de retweets.


Figura 4.12: Documentos con alta relevancia dentro del evento “New York”. El primero corresponde a una imagen en Instagram evaluada con alto puntaje, y el segundo a un tweet de un medio de noticias hablando sobre el evento.