|  |
| --- |
| PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE |
| Memoria de Pre-Grado |
| Exploración Visual de Redes Sociales |
|  |
| **Diego A. Castro Batic** |
| **Noviembre del 2014** |

|  |
| --- |
| En esta memoria se presentan variadas visualizaciones de datos la red social Twitter -un servicio de microbloging. Se discurre por qué se planteó cada una de las visualizaciones y el valor agregado de cada una. Se espera que para visualizar ciertas características del set de datos de Twitter sean fácilmente identificables visualmente. |

# Agradecimientos

A todas aquellas personas que alguna vez se vieron interesadas en pensar sobre las herramientas que usamos y cómo éstas nos afectan. A todas las personas que, dispuestas a ser distintas, dieron espacio para la realización de proyectos únicos; como una memoria de pregrado. Finalemnte, a todas las personas que apoyaron este proyecto indirectamente, ya sea con palabras de ánimo o conversaciones nocturnas hasta entrada la madrugada.

Contents

[Agradecimientos 2](#_Toc402198265)

[Resumen 4](#_Toc402198266)

[Introducción 5](#_Toc402198267)

[Trascendencia (porqué importa Twitter) 5](#_Toc402198268)

[Contexto (las noticias en Twitter) 5](#_Toc402198269)

[Objetivos (la visualización y su análisis) 5](#_Toc402198270)

[Estructura del Documento 5](#_Toc402198271)

[Datos Analizados 6](#_Toc402198272)

[Desarrollo de la Aplicación 6](#_Toc402198273)

[Visualizaciones Desarrolladas 7](#_Toc402198274)

[Elementos Comunes 7](#_Toc402198275)

[Distancia de Texto 7](#_Toc402198276)

[Distancia Temporal 7](#_Toc402198277)

[Distancia Social 7](#_Toc402198278)

[Distancia Geográfica 7](#_Toc402198279)

[Evaluación y Análisis de las Visualizaciones 8](#_Toc402198280)

[Conclusiones 8](#_Toc402198281)

[Bibliografía 8](#_Toc402198282)

[Anexos 8](#_Toc402198283)

# Resumen

La masificación de la tecnología digital ha producido impactos significativos en distintos aspectos de nuestra sociedad. Debido a la facilidad de guardar datos en un formato digital hoy se encuentra una cantidad de datos masiva, de distinto índole, guardada digitalmente. El análisis de estos datos como conjunto plantea nuevos desafíos tecnológicos pues se necesita de nuevas herramientas y metodologías que permitan abordar esta situación. El análisis visual de grandes cantidades de datos promete ser una posible solución para este escenario. ¿Por qué es una posible solución? ¿Por qué una red?

En el presente documento se analizaron datos de la red social Twitter -un servicio de microblogging. Estos datos corresponden a tweets (mensajes de 140 caracteres). fueron procesados buscando similitudes y diferencias entre sí y teniendo en mente su posterior visualización. Se elaboraron visualizaciones de distintos atributos de los datos para luego plantear cuales proporcionan un mayor entendimiento del set de datos. Estas fueron materializadas mediante el uso del programa Gephi que permite la creación de visualización de redes.

Se probó visualizando los tweets como nodos, colocando los que fuesen más parecidos entre sí, más cerca uno de otro. Se analizó si eran parecidos en distintos aspectos para ver cómo esto afectaba a la visualización. Se logró ver la utilidad de visualizar la similitud de texto entre distintos tweets. La visualización del lugar geográfico en el cuál fueron emitidos los tweets resultó infructífera pues la información disponible no es suficiente. La visualización de la distancia de tiempo entre cual fueron emitidos los tweets resultó ineficiente, pues una línea de tiempo o un histograma la retratarían mejor. La visualización de la distancia social entre los distintos usuario que emitieron los tweets resultó inalcanzable debido a las limitaciones puestas por Twitter para acceder a información de este tipo. La visualización del contexto de los tweets basado en la semántica propia del tipo resultó intereante.

# Introducción

## Trascendencia (porqué importa Twitter)

In this way, Vygotsky argued, the tools we build to mediate these symbolic activities change the ways humans think. By building tools, people build the material basis for consciousness, transforming the environments and restructuring the functional systems in which they act and learn (Vygotsky, 1978; Wartofsky, 1983).

* Twitter es una herramienta de microblogging donde cualquier persona puede emitir mensajes públicos de 140 caracteres o menos.
* Debido a su gran aceptación, ha sido sujeto de estudio tanto su tecnología, como las implicancias sociales de esta.
* Su tecnología es interesante por la ubicuidad de sus mensajes, que pueden ser mandados rápidamente desde cualquier lado y pueden abordar cualquier tema.
* Por ende, se han visto retratadas en twitter diversas noticias de importancia mundial por personas comunes y corrientes.
* La forma en que se ve retratada la noticia es motivo de estudio, pues es un instrumento de comunicación de cada uno de sus usuarios.
* En consecuencia, entender la noticia a través de Twitter es entender el comportamiento de sus usuarios y su relación con la noticia a través de sus herramientas tecnológicas.
* En definitiva, el análisis de esta herramienta es importante porque su presencia en nuestra sociedad porque, al estar presente, afecta la forma en que la sociedad misma se comporta.
* Para entender la noticia, hemos partido analizando visualmente un set de datos dado, para ver si podemos entenderlo a grandes rasgos y en base a eso, proponer una herramienta de exploración.
* Es interesante analizar si una herramienta de exploración construida con estos criterios proporciona a un usuario de twitter un entendimiento distinto de la noticia, ya sea más rápido, más claro, o simplemente distinto.

# Contexto

## Trabajo relacionado

No se si queda claro pero acá quiero decir que Existe una herramienta (twitter), que se usa de determinada forma, resultando en un set de datos con características tales que es beneficiado cuando se analiza visualmente. Finalmente que hay consideraciones que se deben tomar visualmente al mostrar información.

Se ha estudiado mucho la difusión de los mensajes a través de twitter

* Se ha estudiando tanto el uso de twitter en la difusión de información en la comunidad científica ([websci10\_submission\_79.pdf](../Desktop/literatura2/websci10_submission_79.pdf))
* como también se ha analizado la difusión de momentos banales y su importancia para los usuarios que las comparten ([Sociology-2012-Murthy-1059-73.pdf](../Desktop/literatura2/Sociology-2012-Murthy-1059-73.pdf)).
* También, se ha analizado la comunidad de personas que "twitean" para entender cómo esta funciona ([American Behavioral Scientist-2011-Gruzd-1294-318.pdf](../Desktop/literatura2/American%20Behavioral%20Scientist-2011-Gruzd-1294-318.pdf)).
* Estos estudios son improtantes por que junto con la presencia de la herramienta surge la noción de que esta es usada de una forma específica por los usuarios que la emplean.

También es interesante conocer la semántica d elo que es difundido

* Este paper es muy bueno por que retrata detalladamente ([Tweet, Tweet, Retweet\_ Conversational Aspects of Retweeting on Twitter.pdf](../Desktop/literatura2/Tweet,%20Tweet,%20Retweet_%20Conversational%20Aspects%20of%20Retweeting%20on%20Twitter.pdf)) la forma en la que los usuarios usan twitter.
* En general los usuarios de twitter "retweetean" información con el propósito general de difundirla, siendo ésta la acción más pasiva que los ve involucrados en el tema del cual la información habla. Esto se hace de diversas maneras pero casi siempre empleando el símbolo "@".
* Por otro lado, muy activamente comentan incluyendo nuevas nociones a la idea comentada o refiriéndose a fuentes externas mediante un link (siempre tiene el símbolo "://")
* Finalmente, es bastante común que se etiqueten los mensajes compartidos usando el símbolo "#" para hacer mención bien al tema del comentarios o a un sentimiento sobre este.

Hay análisis que juntan la semántica y el contenido o la difusión.

* Cabe mencionar que esta semántica es fundamental para el buen entendimiento de la herramienta y ha sido estudiada en profundidad. Por ejemplo, hay trabajo relacionado en:
* [p643-tsur.pdf](file:///C:\Users\VAIO\Desktop\literatura2\p643-tsur.pdf), describió cómo el contenido de un mensaje afecta la propagación de un hasthag (#algo) en el tiempo, para ver que factores afectaban su ciclo de vida.
* [p1031-wang.pdf](file:///C:\Users\VAIO\Desktop\literatura2\p1031-wang.pdf), propuso análisis de sentimientos sobre las co-ocurrencias de hashtags (#algo y #algomas) para noticias dadas.
* [05590452.pdf](file:///C:\Users\VAIO\AppData\Roaming\Microsoft\Word\05590452.pdf) describe el análisis de los factores que hacen que un mensaje sea retwiteado.

Finalmente, la información de twitter tiene que ver con la semántica usada para compartirla.

* La forma en la que Twitter es usado y la cantidad de personas que lo usan crea información que tiene ciertas peculiaridades.
* Es mucha información, a veces repetida (por el hecho de RT), a veces informal (límite de caracteres y expresión libre), no estructurada.
* Estas características hacen que sea difícil analizar todos los datos en su conjunto, de una manera formal e incitan a buscar nuevas técnicas, para ver si estas son efectivas o no.
* Es por esto que el análisis visual puede ser exitoso a la hora de analizar esta información ya que uno tiende, naturalmente, a agrupar y simplificar la información cuando esta es presentada visualmente.

Sobre el análisis visual se ha escrito que

* [p45-donath.pdf](file:///C:\Users\VAIO\Desktop\Memoria\p45-donath.pdf) describió como una visualización puede evocar mucho si esta se piensa para el usuario final y sus intenciones. Es interesante ocupar todos los elementos que la visualización nos permita y asignar significado a todo (formas , colores, posición, cercanía) y no simplemente quedarse en la visualización de un grafo.
* [SocialNetworkVisualization.pdf](file:///C:\Users\VAIO\Desktop\Memoria\SocialNetworkVisualization.pdf) lleva esto más lejos y plantea que dos visualizaciones complementarias mejoran sustantivamente el entendimiento de la red por parte del usuario final.
* También se usó el libro que me prestó Valeria para guiarse en los aspectos de usabilidad y diseño.

# Objetivos

* El objetivo de esta memoria es entender cómo se usa la herramienta desde una perspectiva de datos (mensaje enviado y cantidad de #, @, RT).
* Es decir, entender una noticia tomando en cuenta el uso natural de la herramienta empleada para describir esta noticia.
* Los aspectos de la noticia considerados como interesantes son:
  + temporales (cuando se dijo)
  + sociales (de quien se trata)
  + temáticos (de que se trata)
  + textuales (que se dijo/ que dice el texto)
  + geográficos (donde se dijo, o dónde ocurrió lo dicho)
* Se propone una visualización en forma de grafo (nodos y ejes), donde los nodos tienen
  + posición
  + color
  + tamaño
* Proponer una visualización que usando las propiedades del grafo, retrate los aspectos considerados como interesantes.
* Construir dicha visualización a partir de un set de datos.
* Analizar si la visualización es efectiva y clara. Analizar si la visualización permite obtener conclusiones correctas sobre la generalidad de los datos.
* Se supone que mediante la confección y el análisis visual que aborde estos temas, se llegará a un entendimiento del set de datos.
* En base a dicho entendimiento, se podrá confeccionar una herramienta de exploración de los mismos.
* Se busca ver si esta herramienta proporciona un entendimiento distinto al que proporcionaría una interfaz típica, tal como la que actualmente usa twitter.

# Estructura del Documento

* Descripción de como se consiguieron los datos
* Descripción de las visualizaciones propuestas y las razones para ellas
* Descripción de la aplicación desarrollada para su procesamiento
* Descripción del procesamiento de datos realizado
* Presentación de las visualizaciones propuestas
* Evaluación de la efectividad de éstas.

# Datos Analizados

(se puede ver la metodología desde la pg. 21 de la tesis de Mauricio)

* Facilitados por Mauricio Quezada
* Se tomó información de conocidos medios de comunicación (NYTimes, BBC, etc) y se fueron recopilando sus titulares cada cierto tiempo.
* Cuando muchos titulares coincidían en alguna palabra, esta se consideraba clave.
* Luego se buscaban en twitter mensajes que contenían estas palabras claves.
* Este proceso se repetía periódicamente.
* El resultado de dicho proceso es un set de datos que tiene un evento definido como un par de palabras claves en una fecha dada, asociadas a un gran número de tweets. Los mensajes también tienen fecha, id e id del usuario emisor.
* Cada 24 hrs se corrió un algoritmo para tratar de agrupar noticias similares y desagrupar noticias muy grandes.
* El set de datos es de carácter internacional, predominantemente en inglés
* Las noticias por lo general tienen 500 mensajes (falta chequear a gran escala)

(lo siguiente son consideraciones importantes para nosotros, sobre el set de datos que tenemos)

* Tener un set de datos rotulado con eventos definido facilita nuestra tarea. Nos podemos enfocar más en el análisis de la información que tenemos y menos en la adquisición de ésta.
* Para lo que a nosotros nos interesa, el set de datos presenta los siguientes desafíos.
  + Geográficamente: no hay suficientes datos con ubicación geográfica del tweet. Aproximadamente un 5% (falta chequear)
  + Socialmente: sólo tenemos referencia al usuario que publicó cada mensaje, pero no a la relación entre ellos ni a la relevancia de cada usuario en la red.
  + Temáticamente: no hay información acerca de que hablan los tweets (más allá de la referencia general a una noticia). Es decir, no están clasificados por tema.
  + Los otros aspectos considerados como interesantes no tienen mayores complicaciones, dada la información que tenemos.
  + Generales: ~~la api de~~ twitter propone restricciones a la cantidad de información a la que se puede acceder.[[1]](#footnote-1) Es decir, se puede completar la información de los datos a un ritmo dado, lo que obliga a pensar en un proceso eficiente para analizar los aspectos de interés. (falta explicar que es una api).

la solución a la información incompleta, es analizar la forma en la que los usuarios ocupan estas herramientas. Esta forma tiene una estructura y es por eso que mirando esta estructura (dada por #, @, etc.) se puede indagar en los aspectos donde falta información

* Para abordar esta falta de información, podemos tomar la semántica y el uso habitual de los usuarios como un elemento que nos ayuda a deducir los elementos que la información no incluye inicialmente.
* Proponemos entonces la siguiente solución:
  + Geográficamente: visualizar los lugares y usuarios más recurrentes en un grafo basado en la semántica del tweet (@alguien).
  + Socialmente: identificar los usuarios más relevantes dentro del set de datos (@alguien). Mientras más mencionados se asumen más importantes.
  + Temáticamente: Visualizar la ocurrencia de los temas más recurrentes basado en la semántica del tweet (#tema).

# Desarrollo de las Visualizaciones

Se partió construyendo una infraestructura que permitiera el procesamiento deseado y que pudiese desarrollarse rápidamente.

* Se utilizó código externo para ahorrar tiempo en la construcción de la comunicación con Twitter (http://sferik.github.io/twitter/)
* igual en el procesamiento de texto (https://github.com/opennorth/tf-idf-similarity).
* igual en el stemming del texto la gema 'ruby-stemmer'
* Se utilizó la librería NArray para optimizar rendimiento de rails.
* Se construyeron procesos que, dada una noticia en especial, permiten crear una BD lista para ser importada y visualizada con Gephi (http://gephi.github.io/).
* Después de investigar otros programas se consideró que este era por lejos el más completo y que entregaba más libertades para lograr las visualizaciones deseadas. (En un momento se me sugirió hacer una tabla comparativa para apoyar este punto, pero no sé si sea necesario).
* Esta BD está conformada por nodos (mensajes) que tienen atributos y están relacionados entre sí mediante ejes (que miden similitud entre mensajes).
* Se aprovechó el algoritmo Force atlas (ordenación por fuerza) que junta los nodos unidos por ejes más pesados (i.e. los nodos más similares entre sí) y separa los nodos unidos por ejes menos pesados (i.e. menos similares). Detalles del algoritmo en http://en.wikipedia.org/wiki/Force-directed\_graph\_drawing.
* En todos los casos un grafo completo se ve notoriamente más complicado por lo que se limpió la matríz de ejes, borrando los ejes que denotaban similitudes muy leves.
* Se desarrolló ad-hoc a las visualización que se esperaban lograr
* Se aumentó el programa con un plug-in de clustering creado por Thomas Efer (efer@informatik.uni-leipzig.de) que implementa el algoritmo *Chinese Whispers* según el paper publicado por Biemann (2006, http://wortschatz.uni-leipzig.de/~cbiemann/pub/2006/BiemannTextGraph06.pdf) Para ver si proponía ventajas sobre las visualizaciones propuestas (en caso de hacer segunda propuesta de soluciones). De los 6 plug-ins de clustering este parecía ser el de mejor funcionamiento y además el mejor documentado.
* acá se podría repasar que es un grafo, que representa un nodo y un eje y etc pero no lo creo necesario.

# Visualización de Similitud de Texto

* En este caso se limpió la matriz de similitud considerando que dado el set de datos inicial, donde los nodos al pertenecer a la misma noticia, coinciden por lo menos en dos palabras.
* El procesamiento visual se ve muy confuso si el grafo es completo (todos los nodos relacionados entre sí) por lo que se consideró que limpiarlo era una buena opción.
* Aquí los mensajes son más parecidos entre sí (y se ubican más cerca unos de otros en la red) mientras más parecido sea su texto.
* Resultó útil por la forma en que se utiliza twitter. Se ven claramente grupos grandes de mensajes muy parecidos que son retweets de un mensaje original.
* Por otro lado, se distinguen tweets irrelevantes pues están en la BD por casualidad pero no se parecen a ninguna idea en particular: se ven como puntos aislados fuera del general del grafo.
* Se limpiaron los mensajes para desconsiderar las expresiones más regulares (ie. conjunciones, artículos, etc).
* Se hizo un tallo (stemming) para tratar de llevar todas las palabras a su origen común (por ej. distintas conjugaciones verbales a su raíz)
* Se utilizó tf-idf y la simlitud de coseno para calcular la diferencia entre distintos mensajes.
* Se decidió que podría ser beneficioso limpiar la red, sacando ejes y agrupando los nodos que contienen el mismo mensaje en uno solo. Esto debe ser hecho después de calcular tf-idf y la similitud del coseno, pues estas metodologías están basadas en un grafo completo (relacionan todos los mensajes entre sí).
* Se descartaron distinta cantidades de ejes que denotaban menor similitud (20%, 50%, 97.5%).
* Funciona mucho mejor y es mucho más clara la solución mientras se ven menos ejes.
* Hay noticias donde el procesamiento de texto da resultados interesantes (se pueden apreciar clusters de mensajes retweeteados).
* Hay pequeños clusters (muchos) que entran en la noticia sólo por que comparten palabras comunes (ej. noticia que se atrapó a un mafioso mexicano el mismo día que jugó México en la copa del mundo) siendo esto una causa importante del ruido en la interpretación de los datos.
* Se pueden observar clusters relacionados entre sí, que comparten una noticia pero por lo general difieren en algo (ej. la misma noticia retratada por distintos diarios o sitios web).
* Si se complementa esto, con una medida de valoración social (la cantidad de veces que un mensaje fue comentado o declarado como "favorito" por algún usuario) proporcionalmente con el tamaño del nodo, esto agrega mucha claridad a cuál fue el mensaje original y se distingue mucho mejor la información visualmente. Esto es coherente ya que perdimos un poco de información al juntar todos los grupos de mensajes casi idénticos (retweets) en un solo nodo.
* Se concluye que la similitud del texto entre dos mensajes es información valiosa que podría ser utilizada en la construcción de la aplicación posterior.
* Acá no sé si poner cosas sobre el procesamiento. Por ej. la similitud de coseno entre dos mensajes, implica cruzar cada mensaje entre sí. Es decir, la complejidad de la operación es mensaje al cuadrado. Esto, tomando en cuenta la gran cantidad de mensajes, es poco conveniente.
* También, para calcular el tf-idf, debo cargar todos los tweets como documentos para crear un cuerpo que sepa cuán popular es cada palabra. De nuevo, al considerar la gran cantidad de mensajes (muchas veces repetidos) esto también es poco escalable.
* Estas últimas dos consideraciones traen consigo desafíos técnicos. Por eso traté de hacer tan eficiente como pude el código y utilizar librerías enfocadas en el desempeño.

# Visualización de Similitud Semántica

Se toma para la siguiente sección el término "expresión" como cualquier link, hashtag o arroba. Basados en las costumbres de los usuarios al usar twitter descritas en el trabajo relacionado, puede que este objeto de estudio sea interesante. Se deja de lado el texto que no pertenece a expresiones pues ya fue analizado en la similitud de texto

## Links

* hay noticias donde una gran porción de los mensajes se centra sobre links (ej. 80% de los mensajes contiene un link).
* En estas, cerca de un 10% de los mensajes hace referencia a un link en particular (http://t.co/7fX8V4ElBx)

## Hashtags

* cerca de un 20% contiene algún hashtag, generalmente usado como una etiqueta temática.
* el hashtag más repetido se encuentra presente en un 7% de los mensajes aproxx.

## Arroba

* tal como se mencionó antes En general, una cantidad importante de tweets menciona un usuario en particular usando la notación con arroba (10% de los tweets contienen @algun\_usuario).
* Los usuarios más mencionados se mencionan en algo cercano a un 3% de los tweets de cada noticia analizada.

## Coocurrencias Hashtag-arroba-links

* Si sumamos links, hashtags y arrobas son más que la cantidad de mensajes totales.
* en promedio cada expresion se utiliza 1.76 veces
* El texto plano es poco utilizado
* En general, los mensajes contienen una cantidad de información no despreciable en estos aspectos
* Se construyó un grafo donde cada nodo es una expresión.
* Se tomó la similitud entre nodos como la cantidad de veces que dos expresiones aparecen juntas en un mensaje cualquiera.
* igual que anteriormente, hay consideraciones sobre porque este proceso es eficiente. Lo es, porque voy contando coocurrencias al leer los mensajes y finalmente necesito leer cada mensaje una vez, sin la necesidad de hacer un cruce entre todas las expresiones, para encontrar similitudes entre ellas.
* si se usa la cantidad de repeticiones de una expresión proporcionalmente con el tamaño del nodo resulta muy claro que expresiones fueron más utilizadas.
* se ve claramente que los links compartidos giran en torno a un nodo ya sea un link o un tema.
* se nota como noticias (links) del mismo tema a veces son comentadas y a veces no.
* se podría suponer una posible conversación ya que a veces bien las etiquetas (hashtags) o links van siendo reutilizados en contextos levemente distintos.
* Se concluye que este aspecto puede ser útil en la construcción de una aplicación posterior

# Visualización de Similitud Social

* La cantidad de usuarios de los cuales se puede obtener información desde twitter es acotada.
* La cantidad de amigos de estos usuarios igual es acotada.
* Cerca de un 20% de los usuarios presenta condiciones de privacidad que hacen inaccesible su información (Unauthorized, Forbidden, Not found).
* Estamos obligados a hacer un cálculo tal que esto no sea un problema.
* No se usaron los grados de separación por que se consideró poco apropiado que dos parejas de usuarios que se conocen (una pareja sin nadie en común y la otra con muchos amigos en común) tuviesen la misma separación.[[2]](#footnote-2)
* También, hay usuarios socialmente poco relevantes, incluidos en el set de datos. Es decir, hay que tener en cuenta que un usuario puede estar muy cercano a otro y no ser necesariamente interesante relevante para la noticia a gran escala.
* Se tomaron en cuenta como socialmente relevante a los usuarios mencionados dentro de los mensajes. Estos son mencionados ya porque la noticia se trata de ellos o bien porque alguien retwittea sus mensajes haciendo mención a ellos.
* En general, una cantidad importante de tweets menciona un usuario en particular usando la notación con arroba (10% de los tweets contienen @algun\_usuario).
* Los usuarios más mencionados se mencionan en algo cercano a un 3% de los tweets de cada noticia analizada.
* Se utilizó distancia social como la cantidad amigos que ambos tienen en común, normalizada por la cantidad de amigos de ambos. Es decir: la cantidad de amigos en común dividido por la raíz de la suma de los cuadrados de la cantidad de amigos de cada uno.
* Esto proceso tiene un rendimiento mucho mejor que los grados de separación pues debo buscar sólo 1 vez por mensaje. Esto, aproximadamente, quiere decir que en una noticia promedio de 500 mensajes podría tener el cálculo listo en 8 hrs. aproximadamente.
* Es decir, con este enfoque resolvemos problemas de disponibilidad de datos y tiempo de cálculo. También, es conceptualmente coherente.
* Se distinguen al centro los usuarios centrales a la noticia.
* Si se usa el tamaño del nodo proporcionalmente a la cantidad de veces que fue mencionado cada usuario, destacan claramente los medios de comunicación que transmitieron la noticia primero(@nytimes).
* Se distinguen las fuentes de noticia desde dónde se generaron los titulares.
* Se distinguen nuevos actores relevantes que son altamente mencionados pero no corresponden a fuentes de noticias (ej. @vatican).
* El resultado no es tan claro como con la distancia semántica y de texto, pues el grafo que se forma está muy cerca de ser completo. Es decir, entre actores relevantes, casi siempre hay alguna cantidad de amigos en común.
* No se distinguen fácilmente clusters, cadenas o subgrafos.

# Visualización de Similitud Geográfica

* La cantidad de tweets que contiene metadata sobre dónde fue emitido cada mensaje es muy poca (1-5%) dependiendo de la noticia.
* Se pensó en usar la procedencia del usuario pero una cantidad igualmente baja de usuarios tienen información sobre su ubicación en su perfil y ésta es poco confiable. También, hay que tener en cuenta que la procedencia del usuario puede ser distinta a su ubicación actual.
* Hay muchas cuentas importantes (sitios de noticias altamente populares) que si tienen información de su procedencia pero que explicitan su ubicación como "mundial".
* Esta es información poco estructura, es decir muchas veces el usuario viene de Santiago, Santiago de Chile, stgo., SCL , santiaguino, etc.
* Se utlizó una noticia localmente importante (partido de fútbol Brazil-Chile en Cuiabá, Brazil) pero las menciones a lugares fueron poco comunes aparte de las obvias (Brazil, Chile y en algunas ocasiones Cuiabá). Cerca de un 4% de las menciones totales de lugares fueron a Brazil o Chile y luego el resto de los lugares fueron mencionados a lo más en un 2% de la totalidad de tweets con mención a un lugar.
* El hecho que la información se tan completa y poco estructurada implica que la representación visual de estas últimas tres ideas se ve muy confusa y no permite llegar a conclusiones decisivas sobre la naturaleza de los datos.
* Al limpiar los mensajes que no tienen información, quedamos con una cantidad no-representativa de datos.
* No se observan cúmulos de datos. Se ve todo como una gran nube dispersa.
* Al parecer, si tomamos una noticia donde el lugar sea importante (Chile vs Brasil en el estadio de Cuiabá), debido al carácter internacional de los datos, los lugares mencionados no son necesariamente brasileños ni chilenos.
* Se tiene en cuenta que probable que suba la cantidad de mensajes que contienen información geo-referenciada en un futuro.
* Por esto último, este análisis podría ser más interesante si se toma un lugar más pequeño donde la gente, tipicamente, use información georeferenciada. También podría serlo si se genera información a partir de otro servicio (ej. TripAdvisor, Waze, Foursquare) ya que estos, para funcionar, requieren información georeferenciada.
* El análisis visual da indicios que aún no es fructífero agrupar los mensajes según la información georeferenciada que contengan sobre su lugar de emisión, lugar de procedencia del usuario o incluso los lugares que mencionen.

# Visualización de Similitud Temporal

* La distancia temporal entre tweets es útil, pero resulta ineficiente al ser analizada en un grafo.
* El grafo, ordenado por fuerza, se ve como una línea de tiempo.
* Al apilarse los nodos uno sobre otro no se nota el ritmo al cual se fueron twitteando mensajes.
* Sería más útil analizar esto con un histograma, o bien con un grafo que se fuera modificando temporalmente.
* No obstante, la misma noticia o evento se puede ver retratada a través del tiempo en distintas ocasiones. Por ej. el evento "Israel Palestine" (considerando todas sus variaciones i.e. : israeli, israel, palestinian, palestine y las combinaciones posibles entre todos estos términos) se repite 8 veces en el set de datos proporcionado por Mauricio Quezad.
* Las visualizaciones de los otros aspectos considerados como interesantes cambian levemente entre cada una de estas distintas "versiones" del evento.
* Se puede hacer referencia aquí también a la tesis de Mauricio, que evalúa el impacto de la noticia mostrando dicho histograma. Es decir, no sólo es poco claro, sino que también varía según noticia.

# Evaluación en distintas noticias

Se que en todo los casos se mantiene pero estoy procesando todos los datos para poder corroborarlo de forma más certera. Creo que lo más interesante sería una estadística relacionada con el primer párrafo de cada una de las visualizaciones (esta escrito de forma que el primer párrafo siempre hace referencia a los datos y si en base a ellos se puede obtener algo conclusivo).

## texto

## temporal

## geografica

## social

## semantica

# Análisis de las Visualizaciones (Generalizaciones del análisis visual)

## consideraciones Generales

sobre qué elementos de las visualizaciones (color, tamaño, posición) son útiles al visualizar. También, un pequeño resumen de lo anterior.

* siempre es útil limpiar la red, pues la visualización es mucho más efectiva con nodos que representan gran cantidad de información.
* Es coherente usar la repetición de información (ej. retweets) para dar relevancia. Si bien es mejor limpiar la red, se puede ahorrar este procesamiento y visualizarla entera, teniendo la opción de acercarse. En este caso, cuando se observa un grupo muy cohesionado también resulta evidente su relación pues por lo general comparte una característica común.
* siempre es útil usar el tamaño con relación a relevancia social, ya que da cuenta rápidamente de elementos claves.
* El uso de distintas medidas de relevancia social cambia este resultado levemente. Es decir, muchas veces el mensaje más comentado es distinto al más retwitteado y distinto al más agregado como 'favorito'
* Se necesita indagar más en el uso de distintas plataformas o en el uso de twitter para entender esto en detalle y poder llegar a alguna conclusión más certera.
* visualmente es muy fácil distinguir grupos, pero fundamental poder acercarse al grupo analizado. Sólo las herramientas que proveen zoom fueron útiles (Gephi es muy buena en esto).
* visualmente, si se puede acercar, es fácil ver si el grupo de tweets es relevante. Cuando no es relevante para la noticia el sentido común lo detecta rápidamente, pues son mensajes que por lo general tienen información que hace referencia a otras ideas, muy lejanas a la noticia inicial.
* El color es un buen elemento para distinguir distintos cúmulos de información. Hace más placentera la experiencia de buscar a través del grafo.
* No obstante, la información es poco estructurada y esto hace que hayan demasiados colores en la visualización. Esto implica que no son fácilmente distinguibles uno de otros. Finalmente, la utilización del color para distinguir grupos de datos puede ser sugerida según el usuario.
* Resumimos la sección anterior declarando como interesante para la confección de una aplicación de exploración de redes sociales los aspectos
  + textuales
  + semanticos
  + sociales

ya que, en base a nuestra investigación, su análisis

* El aspecto geográfico, por el momento, resulta infructífero.
* El aspecto temporal, inmediato, resulta infructífero.

# Aplicación de Exploración

## Funcionamiento Base

* En base a dichas conclusiones construimos una aplicación que permite explorar los mensajes del set de datos. Esta tiene un interfaz típico web (no es una aplicación de grafos).
* Se pretende que la exploración sea según lo que típicamente presenta una aplicación web. Es decir, un mensaje central y una lista de los mensajes similares a éste.
* La aplicación muestra un mensaje y a partir de este, sugiere mensajes que son similares en aspectos textuales, semánticos y sociales. Es decir, muestra 3 listas con
  + los 10 mensajes más parecidos en términos de texto
  + las 10 expresiones más relacionadas con las expresiones que el mensaje contiene
  + los 10 usuarios socialmente más cercanos al usuario mencionado en el mensaje.
* Estas opciones son posibles sólo si el mensaje ocupa expresiones. Ahora, por lo descrito anteriormente, en promedio un mensaje ocupa más de una expresión, por lo que la mayoría de estos mensajes tienen las mayoría de estas opciones de exploración disponibles.
* Se puede explorar un usuario o una expresión en particular, mostrando en una vista aparte, los usuarios y expresiones más cercanas a la analizada presentemente.

## Objetivos de la Aplicación

* Que la aplicación sea clara y su usabilidad decente
* Que permita entender la noticia central
* Que sugiera información relevante
* Que permita explorar una cantidad de datos adecuada.

## Evaluación

* Se logra un interfaz claro dónde el usuario puede observar fácilmente las sugerencias del sistema, para explorar el set de datos
* El interfaz es sencillo y no está sobrecargado.
* Se observan claramente y por separado, los distintos grupos de recomendaciones.
* Es fácil de navegar (no presenta muchas opciones tampoco)
* La exploración proporciona un entendimiento estándar. (Ni bueno ni malo, en verdad no se cuanto aporta).
* Tal vez es mejor, para entender la noticia en general, presentar información más relevante al todo general de la noticia en vez de a un mensaje en particular. Por ej. el usuario más relevante para la noticia, en vez de los usuarios más cercanos al mencionado
* La aplicación sugiere información coherente.
* Es decir, usualmente la navegación a través de la aplicación ahonda en el mismo tema y no se desvía mucho del propósito inicial
* Es interesante la exploración, o por lo menos novedosa.
* Creo que el último punto es vital pero no se bien como evaluarlo. Que es, el hecho de no quedarse pegado en una cantidad pequeña de los datos y lograr explorarlos a todos.

## Consideraciones

* Creo, también, que acá hay que hacer un par de consideraciones
  + La exploración está hecha en base a muchos tweets sobre el mismo tema, por lo que es más o menos natural que funcione bien
  + Twitter también permite estas funcionalidades pero no las sugiere tan fuertemente (se puede buscar por texto, hashtag, usuario, etc).
  + Lo que twitter no hace es presentar información desde una panorámica general. Es decir, lo que no tiene es el usuario más relevante para un grupo de tweets relacionados. Yo si lo tengo, pero sólo porque partimos acotando el grupo de tweets
* Esto se traduce en que, dado nuestro set de datos, la exploración es entretenida y el hecho de tener información sobre el set en general es un aporte (i.e.: usuario más mencionado, hashtag más usado, link más comentado, etc).

# Conclusiones

El análisis visual como paso inicial para la creación de una aplicación de exploración de redes sociales es viable. El análisis visual puede ser ocupado en un set de datos para entenderlo rústicamente. Este entendimiento puede ser usado para proponer aplicaciones de exploración y maneras de navegar el set de datos. El resultado es coherente, adecuado y novedoso. Es decir, se puede. La falta de información, desafíos técnicos y la capacidad de procesamiento son aspectos a mirar en detalle.

Es difícil solucionar la falta de información principalmente por las limitaciones impuestas por Twitter y porque la gente no ocupa algunos elementos a su disposición (como geo-referenciales, por ejemplo). Esto se puede solucionar mirando en detalle cómo la gente ocupa twitter y deduciendo formas de analizar la información en base a esto. Es apropiado, poner atención en las prácticas ampliamente utilizadas que surgieron espontáneamente (@my\_house) como una opción válida para reemplazar en el análisis a las prácticas no utilizadas comúnmente(geo-referenciación). Sin embargo, no necesariamente un actor relevante será mencionado. Suponer eso es coherente, pero necesita ser estudiado más en detalle.

Los desafíos técnicos consisten en idear nuevas formas de comparar grandes cantidades de datos. Acá el análisis visual puede aportar mucho, pero el pre-procesamiento sigue siendo complicado. Básicamente, cualquier algoritmo que cruce todos los mensajes entre sí, se vuelve demasiado costoso. En este caso, sólo la distancia de texto sufre de este alto costo, pero se deben proponer nuevas formas de comparar sets de datos, sin la necesidad de comparar todos los datos entre sí. La coocurrencia entre expresiones, creo, es un ejemplo de una forma inteligente de abordar este problema. Al calcularla, sólo se debe pasar una vez por mensaje contando coocurrencias, lo que ahorra mucho tiempo de procesamiento y provee información valiosa.

Finalmente, la capacidad de procesamiento es un tomar en cuenta. Dependiendo del modelo de datos, se pueden usar bases de datos no relacionales para agilizar el manejo de datos en las visualizaciones. No obstante, para el tamaño promedio de las noticias que tratamos, no fue necesario hacer tales ajustes. El procesamiento más arduo se dio en el cálculo de la similitud de texto, donde es necesario comparar todos los mensajes entre sí. Este proceso no se da en conjunto con la base de datos y se trató de agilizar usando librerías para tratar matrices y listas cuyo fuerte fuese el desempeño. No obstante, los tiempos de ejecución siguen dejando mucho que desear. En este punto, ayudó plantear los problemas desde una forma técnicamente inteligente, como fue descrito anteriormente. Tal vez, sería conveniente empezar el trabajo con miras a esta dificultad y usar un lenguaje cuyo fuerte también sea, el desempeño. La capacidad de procesamiento requerida para visualizar simultáneamente todos estos datos también es considerable. Acá recurrimos al programa Gephi para ahorrar tiempo de desarrollo. Este programa funciona muy bien y provee de todas las funcionalidades necesarias para analizar visualmente un set de datos.

nota: necesito un sinónimo de inteligente que no suene tan arrogante.

En cuanto a otros aspectos,

Sería interesante tener mediciones vs otros sistemas de recomendación. Ver si recomendamos más o menos datos, en cuáles coincidimos y en cuáles no. Según tengo entendido, tienen problemas por que recomiendan una cantidad muy limitada de datos.

Sería interesante ahondar más en contextos tipo. Por ej., tal vez una noticia típica deportiva es muy distinta a una política. Como el trabajo está basado en un entendimiento estándar de cómo la gente usa twitter sería interesante poner este supuesto a prueba en distintos tipos de notica. Hemos visto, acá, el resultado de aplicarlo a distintas noticias, pero no hemos tipificado (que tipo de noticia son y cómo los distintos tipos son diferentes entre sí).

# Bibliografía

los papers usados y la tesis de Mauricio

# Anexos

las fotos supongo

1. Por ejemplo, socialmente en https://dev.twitter.com/rest/reference/get/friends/ids. Se pueden pedir los amigos de 15 usuarios cada 15 minutos con un máximo de 5000 amigos Ojo hay otro que es preguntar por friendships/show pero este es solo entre dos usuarios. Hay otro que es friendships/list, pero tiene un máx de 200 en vez de 5000. [↑](#footnote-ref-1)
2. También, este procedimiento se vuelve demasiado pesado muy rápidamente y entre en conflicto con el detalle anterior. Por ej. una persona con 2 grados de separación me obliga a buscar (asumiendo un promedio de 200 amigos pp) los 200 amigos de cada uno de sus 200 amigos 40000 en total. Serían aproximadamente 600 hrs de cálculo [↑](#footnote-ref-2)