|  |
| --- |
| 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.
* Socialmente, ha sido estudiado el impacto de dicha tecnología al estar disponible para una gran parte de la sociedad.
* Se han visto diversas noticias de importancia mundial por personas comunes y corrientes.
* Es importante entender tanto la tecnología, como la forma en que esta se usa para evaluar el impacto de esta en nuestra sociedad.
* Entenderla y analizarla implica entender y analizar como nosotros estamos cambiando.

## Contexto (las noticias en Twitter y trabajo relacionado)

Existe una herramienta (twitter), que se usa de determinada forma, resultando en un set de datos con tales características.

* Actualmente Twitter su usa para difundir información.
* Su gran aceptación ha provocado que hoy en día se vean distintos eventos importantes a nivel mundial, retratados por usuarios comunes y corrientes, en twitter.
* En este trabajo, nos enfocamos en el análisis de una parte del conflicto Israelí-Palestino y en eventos de la Copa Mundial de Fútbol Brasil 2014.
* 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.
* 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.
* 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.
* 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.
* 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.
* Es por esto que el análisis visual puede ser exitoso a la hora de analizar esta información ya que uno tiende a agrupar y simplificar la información cuando esta es presentada visualmente.
* [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 (la visualización y su análisis)

* 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 la topología de una noticia tomando en cuenta el uso natural de la herramienta empleada para describir esta noticia.
* Los aspectos de los mensajes considerados como interesantes son:
  + temporales (cuando se dijo)
  + sociales (de quien se trata)
  + temáticos (que se dijo)
  + geográficos (donde se dijo, o dónde ocurrió lo dicho)
  + semánticos (cómo se dijo)
* 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.

## 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

* Facilitados por Mauricio Quezada
* noticias que consisten de dos palabras claves
* tweets que están relacionados a una "noticia", es decir, contienen dichas palabras claves
* tweet tiene información de su fecha de emisión y un id de la persona que lo emitió
* se tomó como foco noticias del mundial de fútbol y parte del conflicto israelí-palestino

# Análisis (Que visualizar y cómo)

(tal vez esto tiene más forma de conclusión).

* La calidad de la información presenta un problema
  + no hay suficientes datos con ubicación geográfica del tweet
  + la api de twitter propone restricciones a la cantidad de usuarios a los que se puede acceder.
* Se puede tomar en cuenta la semántica de twitter para solucionar esto
  + tomar @ como referencia a un usuario
  + tomar # como referencia temática
  + aún así sigue siendo poca la información
* Solución:
  + Geográficamente: visualizar los lugares y usuarios más recurrentes en un grafo basado en la semántica del tweet (@alguien o @donde) como color
  + Temáticamente: Visualizar la ocurrencia de los temas más recurrentes basado en la semántica del tweet (#tema) como color
  + Visualizar la distancia de texto como posicionamiento en el grafo
  + Visualizar la relevancia social en twitter (RT y Favoritoss) como tamaño
* Solución Alternativa
  + Visualizar por separado los temas, ubicaciones y personas
  + Esto implica dejar de usar color como elemento del grafo

Ver qué solución es más clara

nota: digo los @ y # más recurrentes xq por usabilidad usar más de 6 colores para distinguir no funciona según el libro.

hay otra propuesta que es ver que visualización es más clara, centralizada o descentralizada.es decir, partir de un hashtag para ver los más co-ocurrentes con ese. Y así empezar explorando.

# Desarrollo de la Aplicación

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.
* Esta BD está conformada por nodos (mensajes) que tienen atributos y están relacionados entre sí mediante ejes (que miden similitud entre mensajes).
* En todos los casos se construyó una matríz de 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.

# Visualizaciones Desarrolladas

## Elementos Comunes

acá se podría repasar que es un grafo, que representa un nodo y un eje y etc pero no lo creo necesario.

## 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.

Esta metodología se basa en la construcción de un grafo completo, por lo que 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 ocupa la valoración social (añadir el mensaje como favorito o comentarlo) proporcionalmente con el tamaño del nodo, esto agrega mucha claridad a cuál fue el mensaje original y se distingue mucho mejor la información visualmente.

## 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 temporalemente.

## 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).

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.

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.

Este proceso es muy largo pero para facilitar la visualización se indagó en la semántica del proceso que es twittear.

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.

La red social de estos usuarios es fácil de calcular (1-3 hrs, limitados por la API de twitter) e interesante.

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 (@nytimes)

## Geográfica

La cantidad de tweets que contiene metadata sobre dónde fue emitido cada mensaje es muy poca (1-5%).

Se pensó en usar la ubicación del usuario pero una cantidad igualmente baja de usuarios tienen información sobre su ubicación en su perfil y ésta es poco confiable.

Se utlizó una noticia localmente importante (partido de fútbol 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.

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.

Se tiene en cuenta que probable que suba la cantidad de mensajes que contienen información geo-referenciada..

## Semántica

### 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

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 tema 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.

# Evaluación y Análisis de las Visualizaciones

siempre es útil limpiar la red

siempre es útil usar el tamaño con relación a relevancia social

visualmente es muy fácil distinguir grupos, pero fundamental poder acercarse al grupo analizado

visualmente, si se puede acercar, es fácil ver si el grupo de tweets es relevante

es vital que el nodo tenga un nombre

Es coherente usar la repetición de información (ej. retweets) para dar relevancia.

# Conclusiones

# Bibliografía

# Anexos