

Twitter Semantic Similarity (TSS)

I, me, mine

Facundo Carrillo

- [LIAA – Laboratorio de Inteligencia Artificial Aplicada, ICC, UBA/CONICET](#)
 - [Maestría de Ciencias de Datos, Udesa](#)
 - <https://www.linkedin.com/in/facuzeta/>
 - fcarrillo@dc.uba.ar
 - <https://facuzeta.blogspot.com/>
 - https://twitter.com/facu_zeta
-

Psiquiatría computacional: Qué

- Medidas objetivas poco sesgadas
- Complementarias a la información subjetiva
- Diferentes producciones de la mente:
 - Comportamiento
 - Lenguaje

Referencias:

- Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. Trends in cognitive sciences, 16(1), 72-80.
- Adams, R. A., Huys, Q. J., & Roiser, J. P. (2016). Computational psychiatry: towards a mathematically informed understanding of mental illness. Journal of Neurology, Neurosurgery & Psychiatry, 87(1), 53-63.
- Sigman, M., Slezak, D. F., Drucaroff, L., Ribeiro, S., & Carrillo, F. (2021). Artificial and Human Intelligence in Mental Health. AI Magazine, 42(1), 39-46.
- <https://elgatoylacaja.com/psiquiatriapp>
- <https://www.youtube.com/watch?v=uTL9tm7S1lo>

Psiquiatría computacional: Quienes



<https://liaa.dc.uba.ar/es/inicio/>



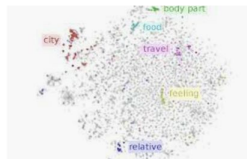
<https://www.cocucolab.org/>



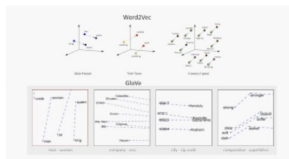
Twitter Semantic Similarity (TSS)

Embeddings

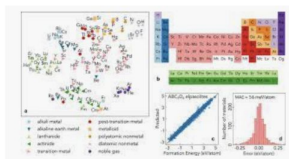
- Co-ocurrencia
- LSA
- Word2vec
- Fasttext
-



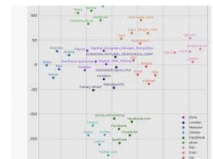
On word embeddings - Part 1
ruder.io



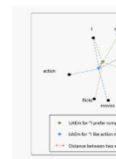
Word Embeddings for NLP: Understanding ...
towardsdatascience.com



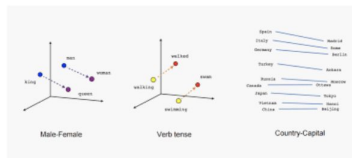
Chemistry is captured by word ...
researchgate.net



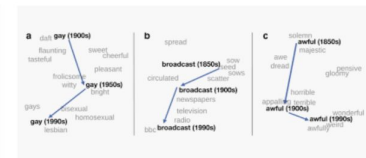
Discover The Power of Word Embeddings ...
openclassrooms.com



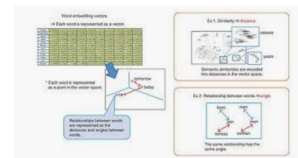
Task-Optimized Word
frontiersin.org



Creating Word Embeddings: Coding the ...
towardsdatascience.com



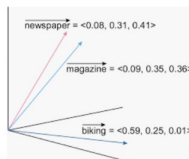
Word embeddings in 2017: Trends and ...
ruder.io



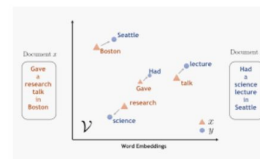
Memory-efficient Word Embedding Vectors ...
nnt-review.jp



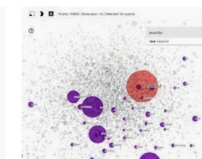
Dynamic word embeddings for ev
blog.acolyer.org



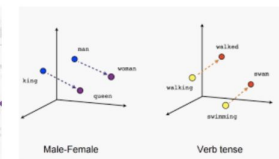
3-dimensional word embeddings ...
researchgate.net



Universal Text Embedding from Word2Vec ...
ibm.com



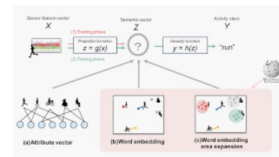
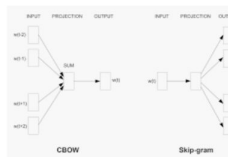
Word embeddings | Text | TensorFlow
tensorflow.org



Text Vectorization in NLP ...
medium.com



A High-Level Introduction to Word ...
predictivehacks.com



Vi
es

Embeddings

Pseudo-código:

1. Bajo un corpus
2. Entreno/fiteo
3. Queries

¡Todo muy estático! Las cosas cambian

Normalized Google Distance

Cilibrasi, Rudi L., and Paul MB Vitanyi. "The google similarity distance." *IEEE Transactions on knowledge and data engineering* 19.3 (2007): 370-383.

$$\text{NGD}(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

$f(x)$ es la cantidad de páginas indexadas por Google usando el término x como búsqueda.

$f(x,y)$ es lo mismo pero cuando aparecen ambos términos

N : número total de páginas indexadas en Google (?????)

Google Semantic Similarity

Ejemplo: ¿Quién está más cerca de Dios: Messi o Maradona?

$f(\text{'messi'}) = 476,000,000$

$f(\text{'maradona'}) = 33,900,000$

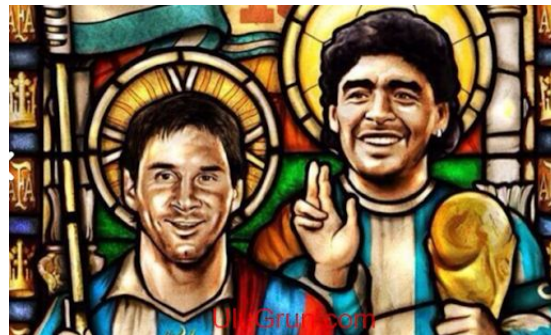
$f(\text{'dios'}) = 542,000,000$

$f(\text{'dios maradona'}) = 3,410,000$

$f(\text{'dios messi'}) = 12,600,000$

$N = 3000000000000000$ (valor grande)

$$\text{NGD}(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}}$$



Google Semantic Similarity

Ejemplo: ¿Quién está más cerca de Dios: Messi o Maradona?

$f(\text{'messi'}) = 476,000,000$

$f(\text{'maradona'}) = 33,900,000$

$f(\text{'dios'}) = 542,000,000$

$f(\text{'dios maradona'}) = 3,410,000$

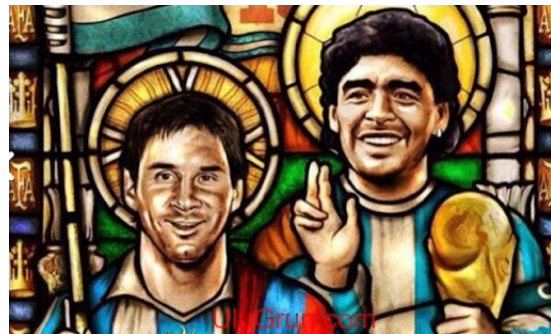
$f(\text{'dios messi'}) = 12,600,000$

$N = 3000000000000000$ (valor grande)

$$\text{NGD}(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

$\text{NGD}(\text{dios}, \text{maradona}) = 0.370$

$\text{NGD}(\text{dios}, \text{messi}) = 0.340$



Google Semantic Similarity

Ejemplo: ¿Quién está más cerca de Dios: Messi o Maradona?

$f(\text{'messi'}) = 476,000,000$

$f(\text{'maradona'}) = 33,900,000$

$f(\text{'dios'}) = 542,000,000$

$f(\text{'dios maradona'}) = 3,410,000$

$f(\text{'dios messi'}) = 12,600,000$

$N = 3000000000000000$ (valor grande)

Este slide des una clase vieja...
¿Qué dará ahora?

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

$NGD(\text{dios}, \text{maradona}) = 0.370$

$NGD(\text{dios}, \text{messi}) = 0.340$



Twitter Semantic Similarity

NGD esta buena pero tiene mucha inercia.

Basandonos en esta idea, propusimos **Twitter Semantic Similarity**

- Resolución temporal altísima
- Fácil de computar

Twitter Semantic Similarity

NGD esta buena pero tiene mucha inercia.

Basandonos en esta idea, propusimos Twitter Semantic Similarity

- Resolución temporal altísima
- Fácil de computar

Se basa en medir la co-ocurrencia de palabras tweets. Pero Twitter no nos dice cuantos tweets hay para una query en un tiempo dado. Pero si podemos buscar muchos y ver con que velocidad ocurren

Usando la **VELOCIDAD** como estimador de la cantidad.

Carrillo, F., Cecchi, G. A., Sigman, M., & Fernandez Slezak, D. (2015). Fast distributed dynamics of semantic networks via social media. *Computational intelligence and neuroscience*, 2015.

Twitter Semantic Similarity

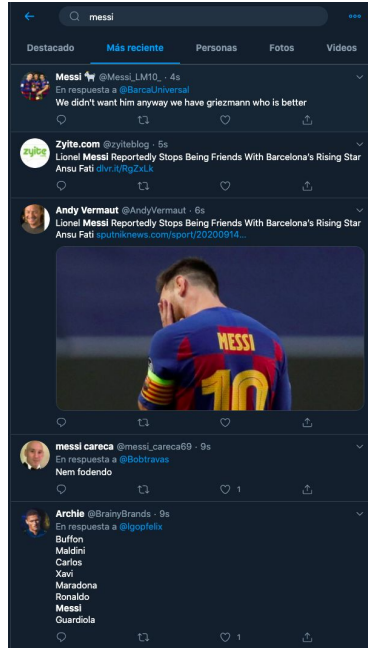
Velocidad:

$$\Phi(w) = \left(\frac{\sum_{i=1}^{N-1} (\tau_{i+1}(w) - \tau_i(w))}{N-1} \right)^{-1}$$

Buscamos los últimos N tweets y estimamos la frecuencia con que se tuitean.

Twitter Semantic Similarity

Por ejemplo, buscamos messi, los ultimos 5 tweets



Tweet created_at 12:33:14

Tweet created_at 12:33:11

Tweet created_at 12:33:09

Tweet created_at 12:33:07

Tweet created_at 12:33:03

$$((3+2+2+4) / 4)^{-1} = 0.3636$$

$$\Phi(w) = \left(\frac{\sum_{i=1}^{N-1} (\tau_{i+1}(w) - \tau_i(w))}{N-1} \right)^{-1}$$

Twitter Semantic Similarity

Con la “velocidad” resuelta, definimos TSS entre dos palabras como:

$$\text{TSS}(w_1, w_2) = \left(\frac{\Phi(w_1 \wedge w_2)}{\max(\Phi(w_1), \Phi(w_2))} \right)^\alpha$$

Twitter Semantic Similarity

Experimento control 1: ¿Cómo se compara con word embeddings ya validados?

- Armamos 100K pares de palabras
- Los medimos en LSA, Wordnet, Word2vec etc

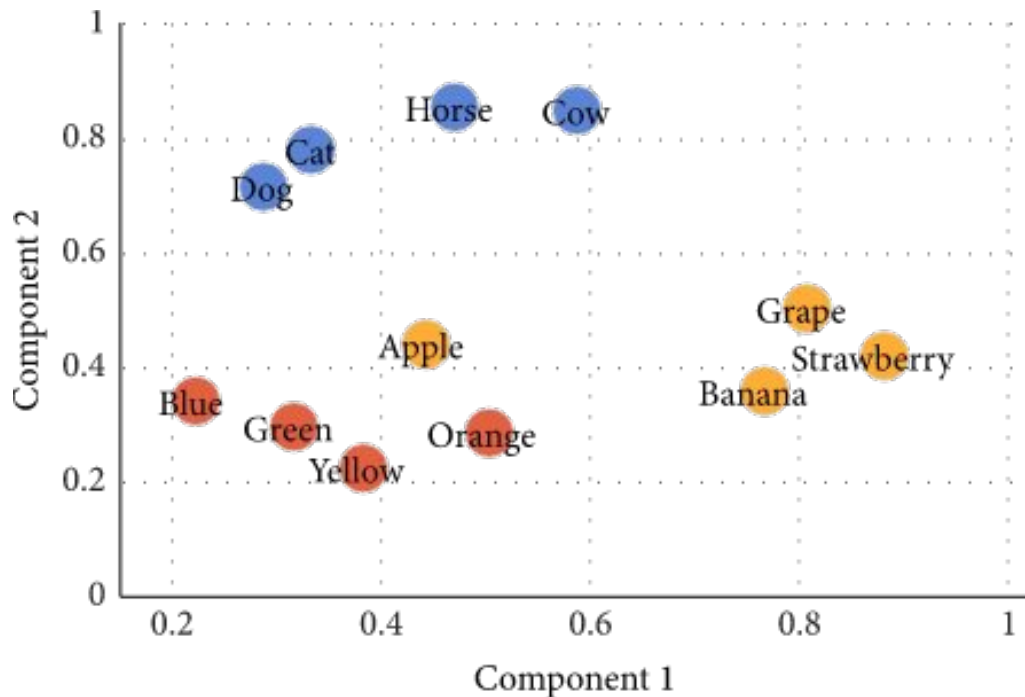
Miramos que correlacionaban bien las similitudes o distancias usando los embeddings y TSS!

Experimento control 2: Test de sinónimos

- Tomamos muchos pares de sinónimos de una prueba estandarizada y vimos que estaban más cerca que pares al azar

Twitter Semantic Similarity

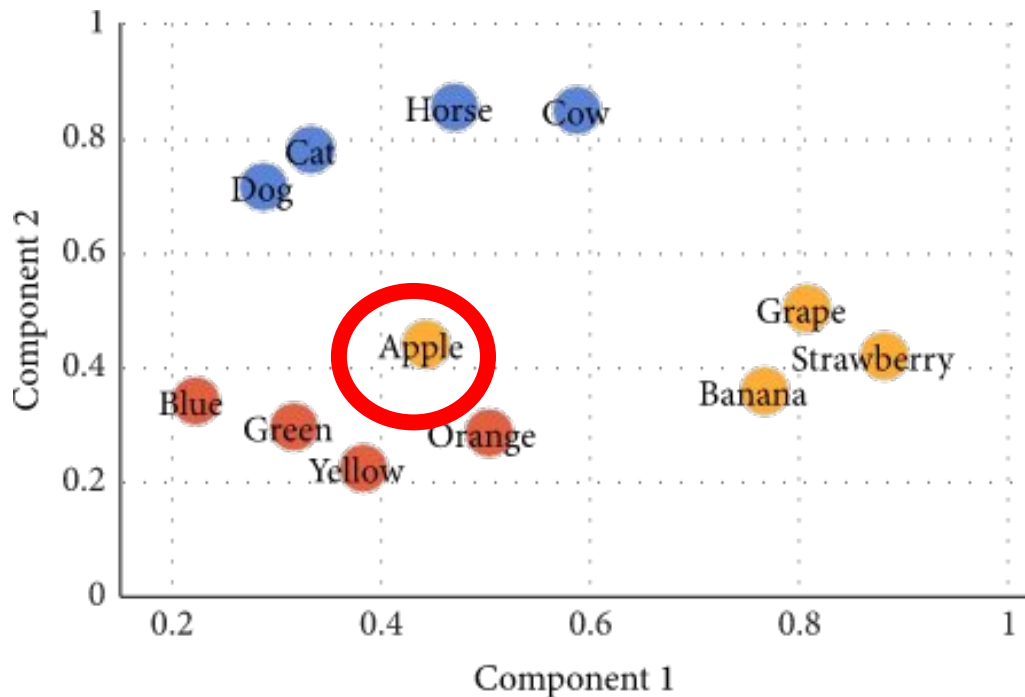
Experimento control 3: Categorías armadas a mano



(proyección en 2D a partir de la matriz de similitud)

Twitter Semantic Similarity

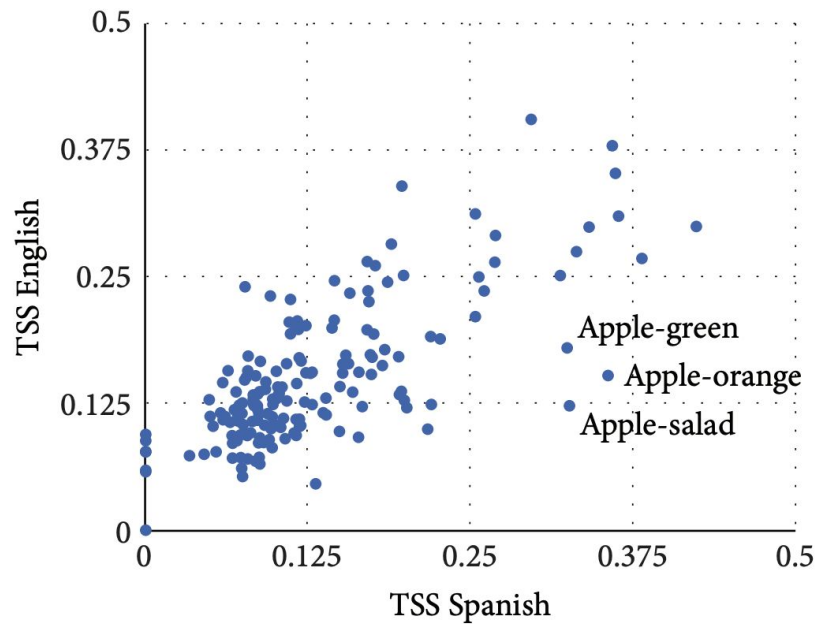
Experimento control 3: Categorías armadas a mano



(proyección en 2D a partir de la matriz de similitud)

Twitter Semantic Similarity

Polisemia!



Twitter Semantic Similarity

Queremos estudiar dinámicas:

1. Un concepto móvil
2. La red se mueve

Twitter Semantic Similarity

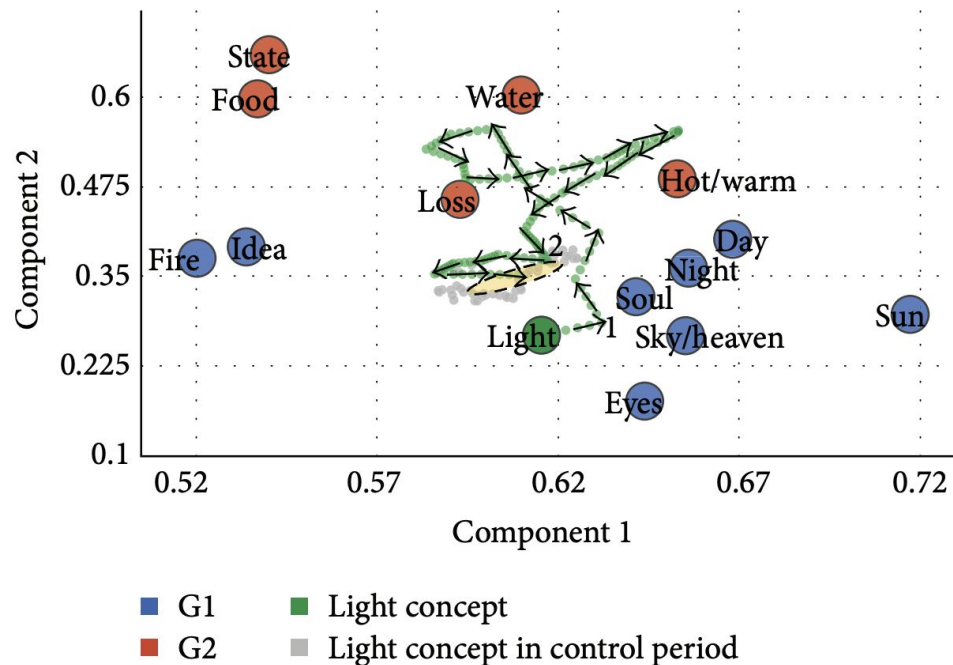
Experimento concepto móvil: Cortes de Luz Diciembre 2013



Buen escenario para capturar cambios en la semántica de cambio rápido!

Twitter Semantic Similarity

Un concepto móvil: Cortes de Luz



Propusimos dos grupos de palabras:

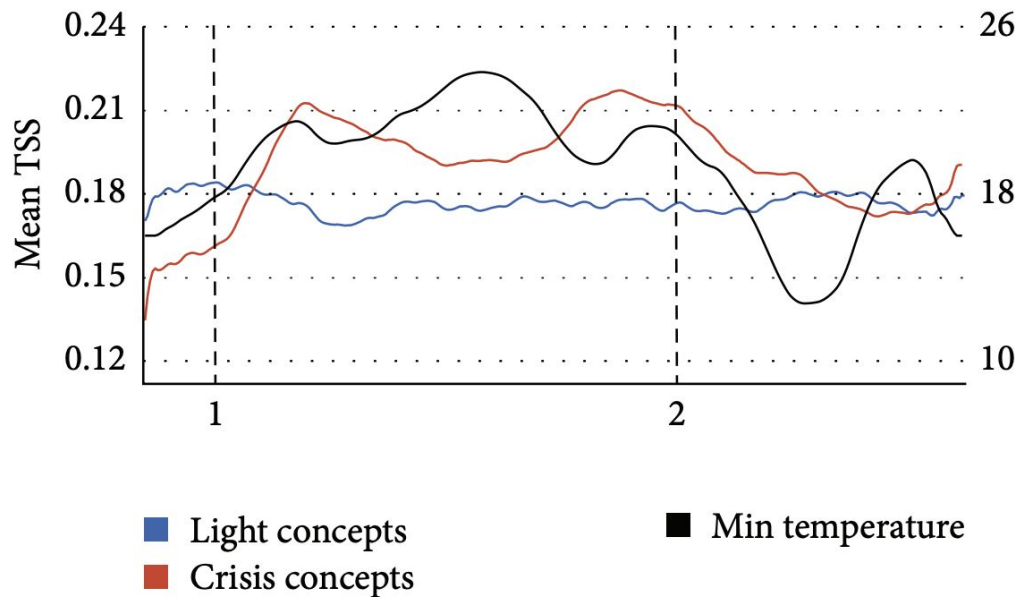
- Control
- Crisis

Estudiamos el movimiento del concepto de la palabra **luz**

Puntos grises 3 meses después

Twitter Semantic Similarity

Experimento concepto móvil: Cortes de Luz



TSS media a los conceptos según grupo

Correlación positiva con Crisis

Correlación negativa con Control

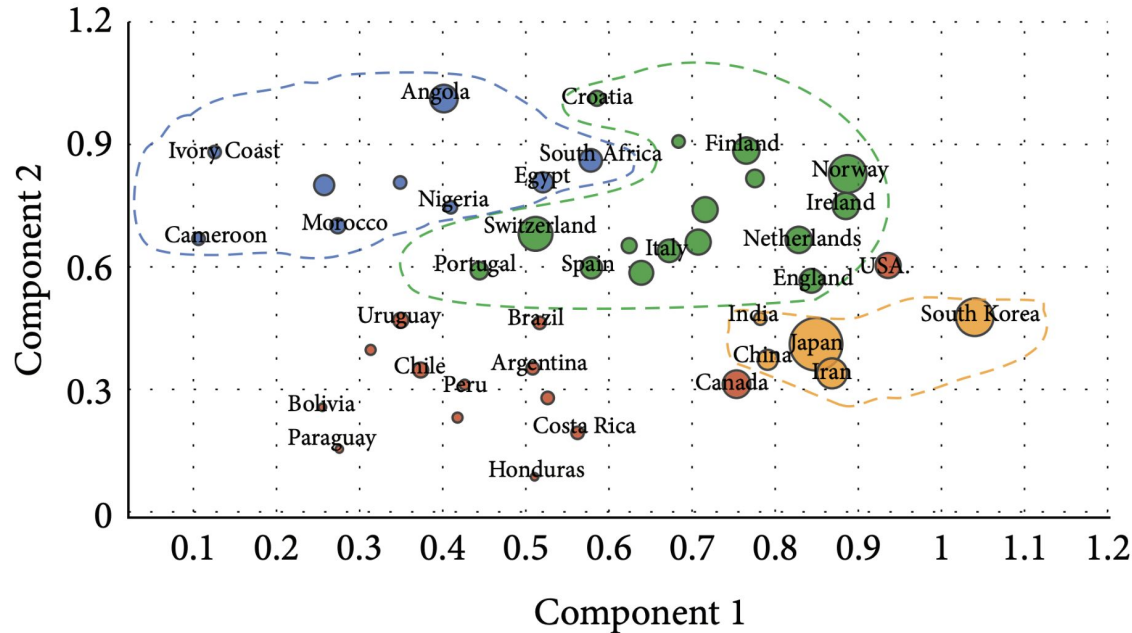
Twitter Semantic Similarity

Un evento cambiando la estructura de la red: Experimento sorteo grupos del mundial



Twitter Semantic Similarity

Un evento cambiando la estructura de la red: Experimento sorteo grupos del mundial

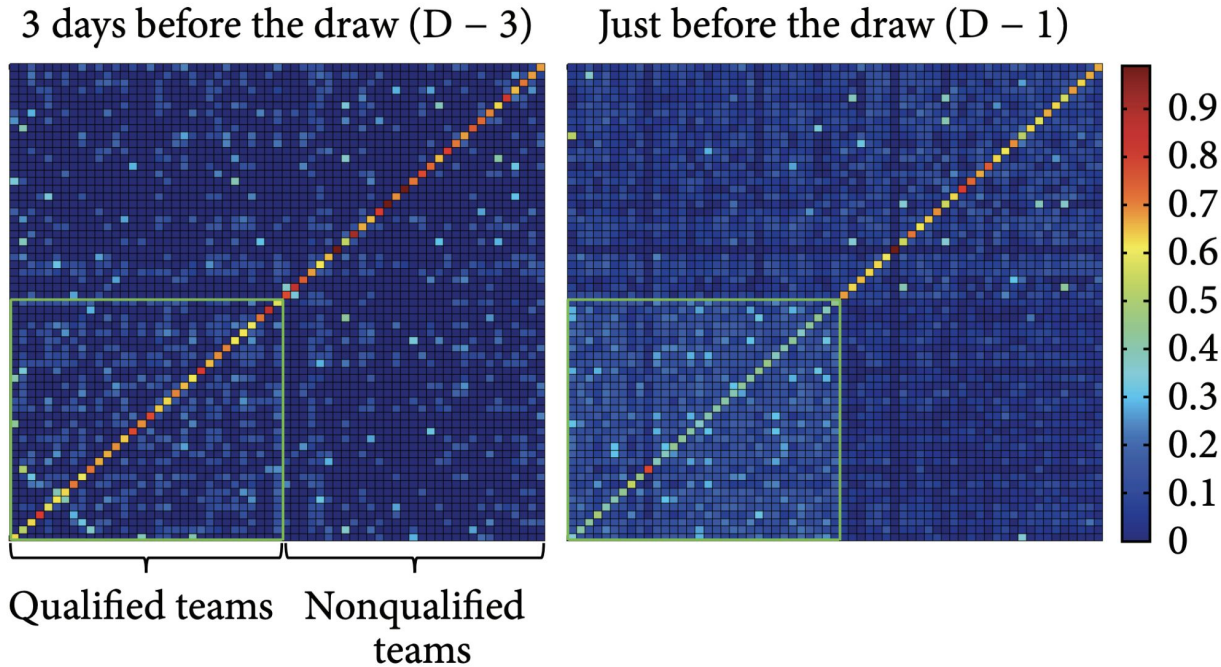


Control

- TSS capturo “bien” la geografía (K vecinos ok)
- 2do orden PBI (correlación con el centro de masa)

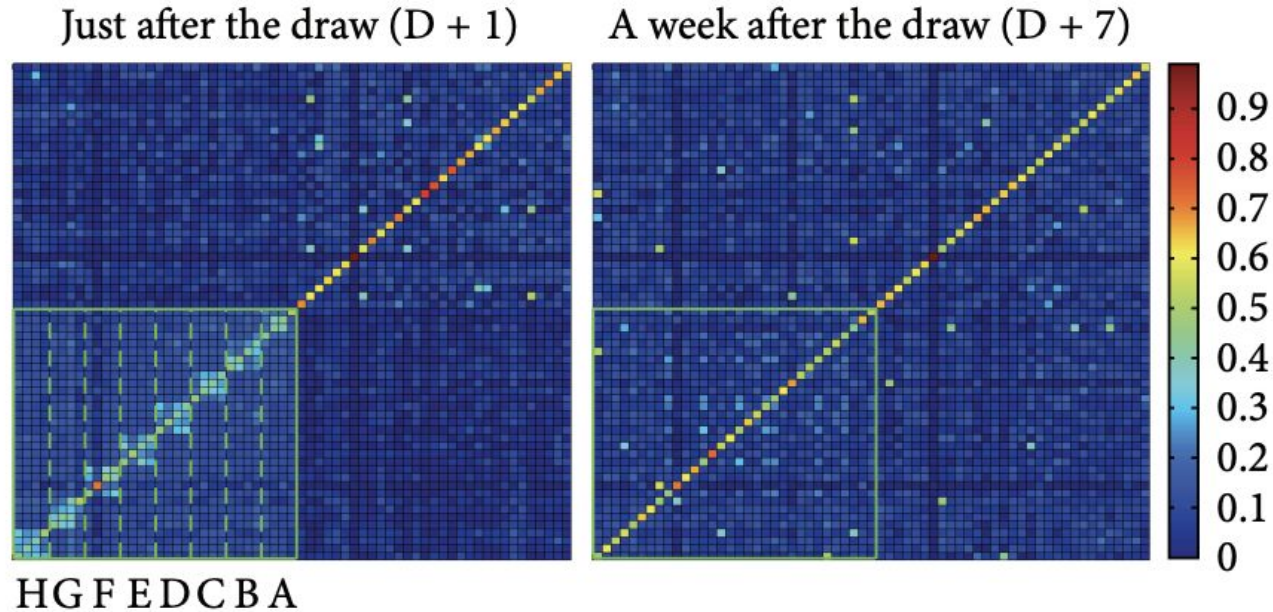
Twitter Semantic Similarity

Un evento cambiando la estructura de la red: Experimento sorteo grupos del mundial



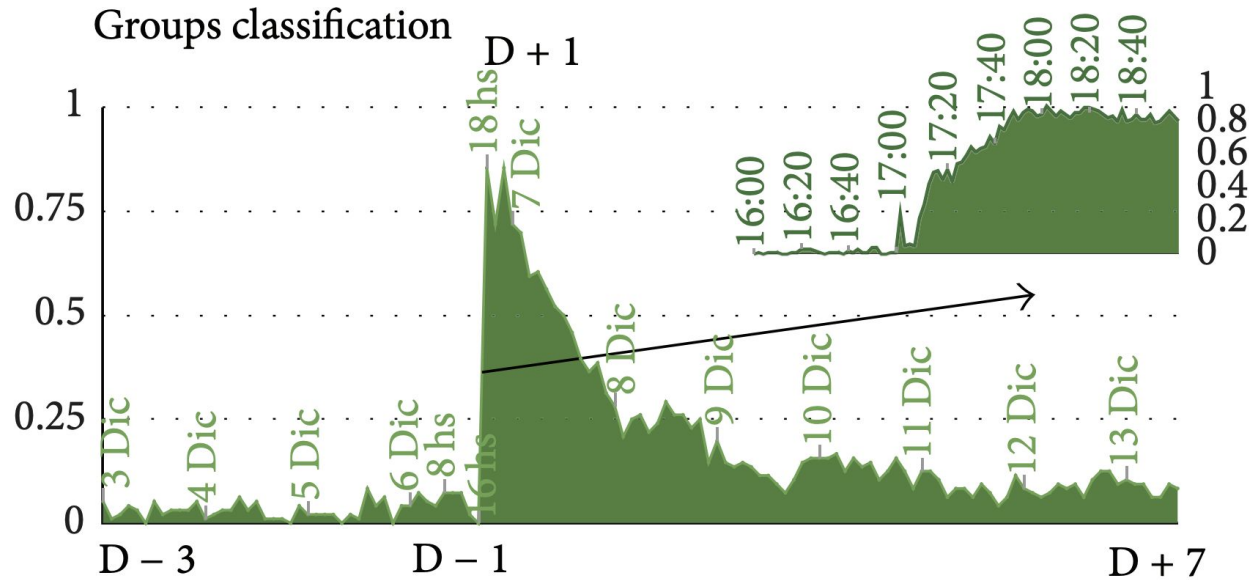
Twitter Semantic Similarity

Un evento cambiando la estructura de la red: Experimento sorteo grupos del mundial



Twitter Semantic Similarity

Un evento cambiando la estructura de la red: Experimento sorteo grupos del mundial



Performance clasificador
grupos

Twitter Semantic Similarity

Conclusión

- TSS nueva medida de similaridad semántica con alta temporalidad
- Barata (consumiendo la api de Twitter)
- Conserva estructura en general

Bonus track: Blog

<http://facuzeta.blogspot.com/>

- Datos FCEN de cursada, aprobación materias y como se respeta o no el plan
- Infracciones de tránsito PBA
- Distribución de tiempos de trámites en GCBA
- Concursos docentes DC

Sitio d
denun

En <https://>
hacer dife

En particular, yo lo use para denunciar un auto al hay que subir una foto, completar unos datos del siguiente de hacer la denuncia ya estaba impaciente, quería tener trámite.


La web ofrece, a medida que el tramite se mueve un estado que s

 Fecha y hora de ingreso:
01/06/2018 - 09:53 h.

 Fecha y hora de cierre:
04/06/2018 - 18:45 h.

Estado: Cumplido

Descripción de estado:
Ya resolvimos la solicitud.

 Fecha y hora de:
22/04/2018 - 13.

Estado: En proceso

Descripción de estado:
Estamos llevando a cabo i
resolver la solicitud.

Infracciones de tránsito de la Provincia de Buenos Aires

Quería averiguar si el auto de un amigo tenía infracciones y como está radicado en la provincia de Buenos Aires me fijé en la web de *INFRACCIONESBA*. Encontré un sistema para *evitar* pedidos masivos que implementa una especie de captcha muy particular...

Búsqueda por Dominio

Búsqueda por Documento

Dominio

Ej.: FLY888

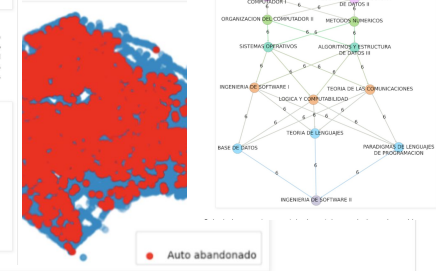
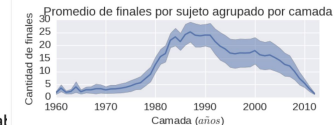
Código Captcha

PLMXC

Buscar D



La distribución de alumnos por el año de su libreta muestra valores coherentes para cierto rango de años. Se ve notoriamente un pico en 1985 debido al inicio de la puesta en marcha del ciclo básico común (CBC) en reemplazo del examen de ingreso. A su vez se ve un pico positivo para el año 1973 (desconocemos el motivo, tal vez el contexto argentino). Los datos de los alumnos más antiguos probablemente no estén completos en el sistema por eso también vemos muy pocos datos previos a 1960.



En fin, suponiendo entonces que el barrio no modificaba lo que tarda un pedido en ser resuelto, miré las distribuciones de tiempos para trámites de "Remoción de vehículo", separando en los que terminan favorablemente para él que hizo el pedido y los que no.

Distribución de tiempos separando cómo terminan los trámites para *Remoción de Vehículo abandonado* en la web de Gobierno de la Ciudad de Buenos Aires

