Trabajo Final de Ciclo

Análisis de categorías de noticias digitales para determinar tópicos cubiertos entre 2011-2015, mediante varios algoritmos de clustering en un dataset de tweets con temática de salud

Universidad de Cuenca
Optativa - Minería de Textos
Freddy L. Abad L., Edisson. S. Reinozo T.
{freddy.abadl, edisson.reinozo}@ucuenca.edu.ec

Contenido

Introducción

A

Contexto

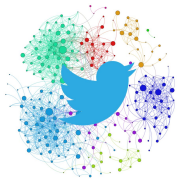
El uso de técnicas de minería de texto ha tenido un auge para el entendimiento de grandes cantidades de texto generado por redes sociales [1].

Popularización debido al acceso libre de las librerías que facilitan su adopción [3].

Necesidad

Establecer el impacto y relación, de por ejemplo una enfermedad, en la cantidad de tweets que se generan por los medios.

Medir la interacción generada por los medios de comunicación digitales y su evolución de temas en el tiempo.



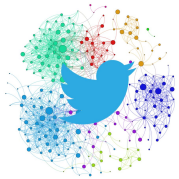
Tarea

A

1. Preprocesamiento: Obtención, Tratamiento y Unificación de los archivos fuentes.
2. Procesamiento: Tokenizado de frases, Eliminación de stopwords, Stemming y Lematización, TF-IDF. Palabras más representativas de cada documento-tweet.
3. Clustering: Entrenamiento de los datasets (3 stem, lema, stem y lema) mediante los algoritmos de clusterización de K Means, Spectral y Hierarchical. Obtención de K óptimos para cada algoritmo. Etiquetado de clusters
4. Análisis resultados.

Input: Tweets sin tratar

Output: Dataset procesado y las categorías respectivas según el tiempo y el medio de comunicación mediante el mejor modelo de clusterización



Objetivo General

Categorizar los tópicos que cubren los medios de comunicación en el área salud en la red social Twitter y determinar si existe una variación de estos en el tiempo, mediante la aplicación de algoritmos no supervisados de clusterización

Trabajos relacionados

Pohl, et. al.

Exploración del problema de identificación de subeventos en tiempo real en RRSS.

Indexación-agrupación en línea de flujos de datos para generar informes situacionales, en contextos epidemias, huracanes, etc.

Evalúa rendimiento de algoritmos no supervisados para el agrupamiento por similitud textual, Simple K Means y Spectral K Means

Xiangfeng , et. al.

Método de agrupación en incrustaciones de palabras, en text-data de **tópicos de salud pública**

Incrustación de palabras es una tendencia fuertemente usada en el NLP

Aprende los vectores óptimos de palabras circundantes y los vectores representan la información semántica de las palabras.

Técnica no supervisada, **no requiere etiquetado previo de los datos,** puede extenderse a otros **problemas de agrupamiento- enfermedades.**

B

Alsayat , et. al.

Estudia el comportamiento social humano analizando gran flujos de datos de RRSS.

Distinguir entre usuarios regulares, medios digitales y líderes de opinión

Mejorar la granularidad de las comunidades de usuarios y su comportamiento mediante marco para la detección de comunidades usando K-Means y algoritmos genético.

Hussain Shah , et. al.

Metodología para la agrupación de nombres de marcas, basada en datos de Twitter

Algoritmo BNACA: Contraste de K Means vs Hierarchical Clustering

Dendograma: Máximas similitudes entre marcas a través de los tweets.

Dataset

Dataset

“Health News in Twitter Data Set”

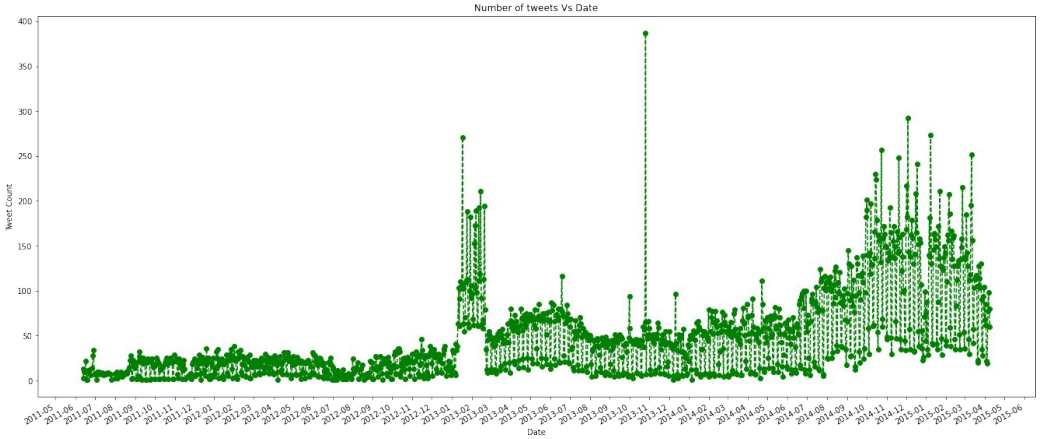
	tweet id	date and time	tweet
0	304596701757464576	Thu Feb 21 14:21:27 +0000 2013	#FastFood Makes Up 11 Percent of #Calories in ...
1	304595191329853441	Thu Feb 21 14:15:27 +0000 2013	10 snacks to help you lose weight, burn fat, a...
2	304587659018371072	Thu Feb 21 13:45:31 +0000 2013	10 foods that boost your skin AND slim your wa...
3	304580073380524032	Thu Feb 21 13:15:22 +0000 2013	What a heart attack feels like in women (it's ...
4	304572560270573569	Thu Feb 21 12:45:31 +0000 2013	#McDonalds oatmeal has almost 7 teaspoons of s...

- Compuesto por tweets sobre salud
- Rango 2011 - 2015
- 15 medios de comunicación
- Inconsistencias en el separador de columnas “|”, en miles de líneas de varios archivos.
- Inconsistencia carácter comilla en el corpus del tweet

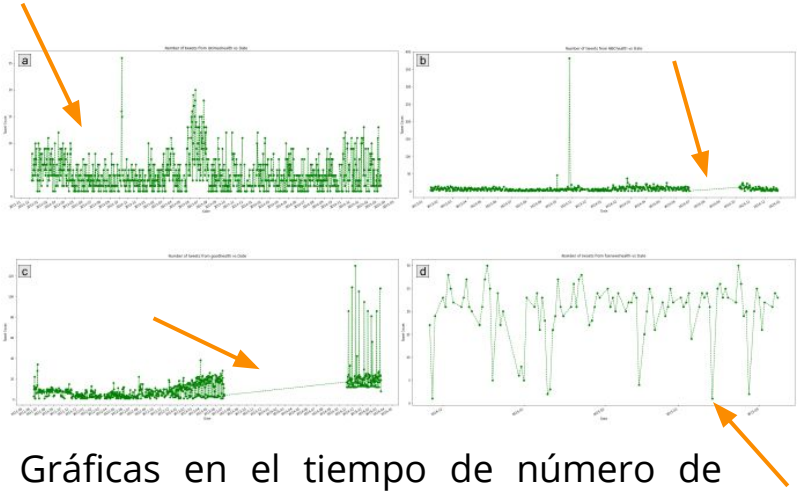
Medio Digital	Instancias
Everyday Health	3239
CBC Health	3741
CNN Health	4061
Fox News Health	2000
GDN HealthCare	2997
Good Health	7864
Kaiser Health News	3509
LA Times Health	4171
MSN Times Health	3199
NBC Health	4215
NPR Health	4837
NY Times Health	6245
Reuters Health	4719
US News Health	1400
WSI Health	3200
TOTAL	64117

Metodología

Exploración del dataset



Serie de tiempo del número de tweets global por día en el rango de 2011-2015.



Gráficas en el tiempo de número de tweets de 4 medios de comunicación del total de medios en análisis.

http://bit.ly/37r33ec

Procesamiento de datos

Tokenización

Dividir oraciones en unidades atómicas con significado propio

Eliminación de stopwords

Eliminar palabras irrelevantes “The”, “and”, “about”, “rt”, links.

Stemming y lematización

Raíz morfológica de cada palabra de un tweet.
Generación de 3 datasets.

Análisis descriptivo y analítico del dataset

TF-IDF

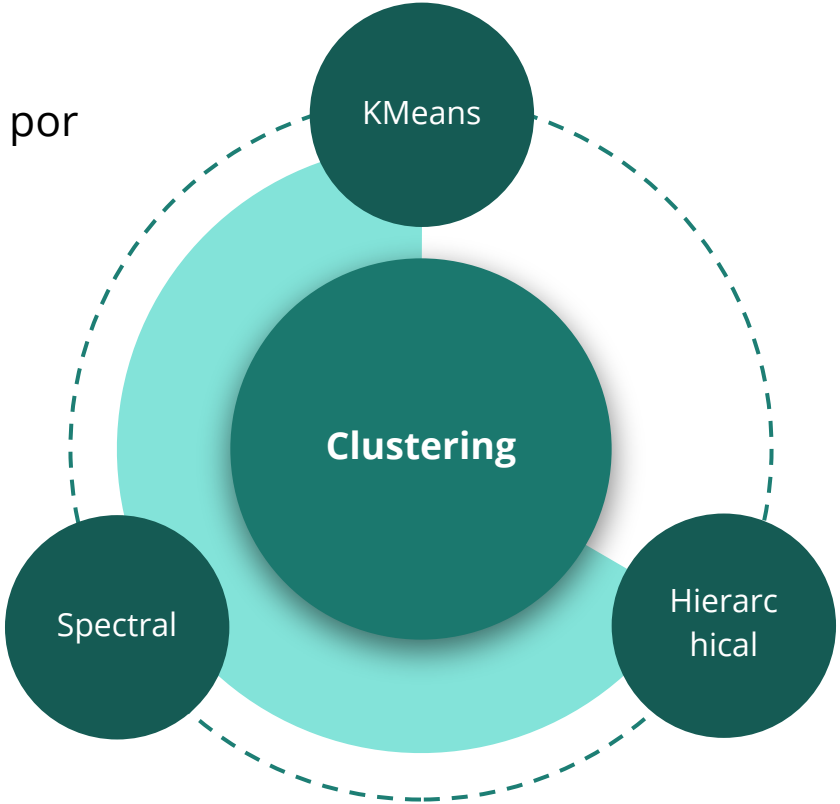
Calcula la relevancia de palabras en una oración.
Palabras clave = mayor relevancia.

Análisis descriptivo y analítico del dataset

	Word	count
0	health	5185
1	ebola	4855
2	new	3852
3	say	3550
4	studi	3466
5	may	3395
6	us	2736
7	get	2653
8	drug	2622
9	help	2223
10	cancer	2218

Algoritmos Clusterización

Aplica análisis de conglomerados evaluando principalmente por distancia.



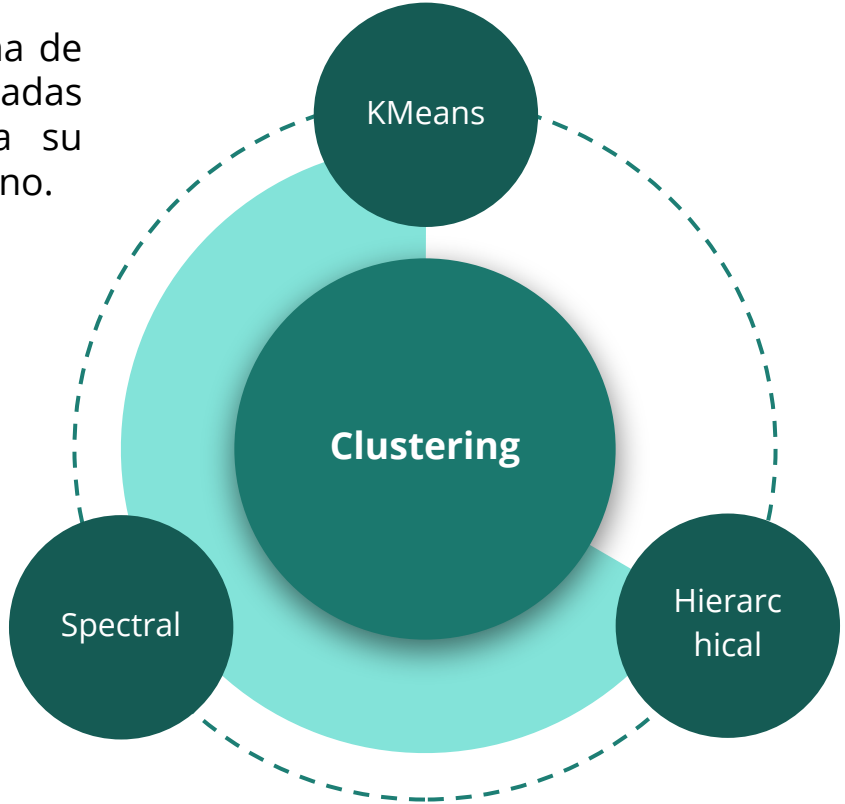
Analiza conglomerados evaluando principalmente por conectividad.

Aplica análisis de agrupación divisivos o aglomerativos para construir una jerarquía de agrupaciones.

Selección K Óptimo

Elbow curve

Métrica inertia: suma de las distancias cuadradas de las muestras a su centroide más cercano.



Heurística de Eigengap

Métrica de diferencia entre eigenvalor consecutivos

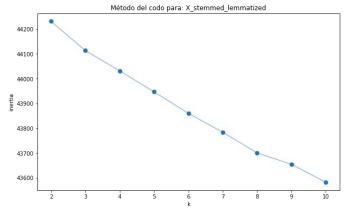
Dendograma

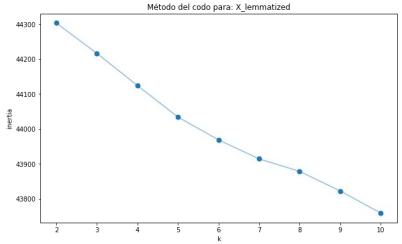
Métrica “Ward Linkage” mide distancia entre clústeres mediante la suma de diferencia de cuadrados de distancias intra-clústeres.

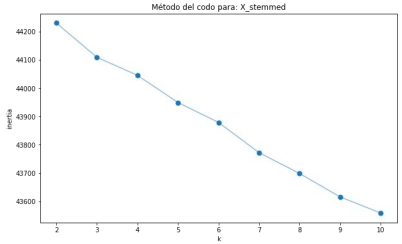
K Means

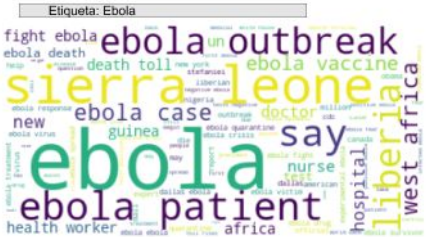
	K óptimos establecidos para los distintos datasets		
K Means	d_stem	d_lem	d_stem_lem
K Óptimo	7	7	8

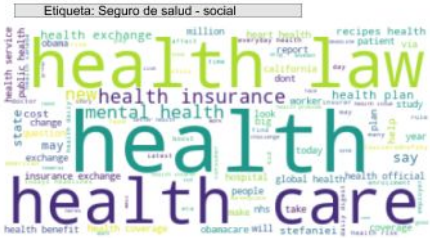
	SELECCIÓN DE ALGORITMO ÓPTIMO PARA LOS DISTINTOS DATASETS		
K Means	d_stem	d_lem	d_stem_lem
Alg. Óptimo	No Óptimo	Óptimo	No Óptimo







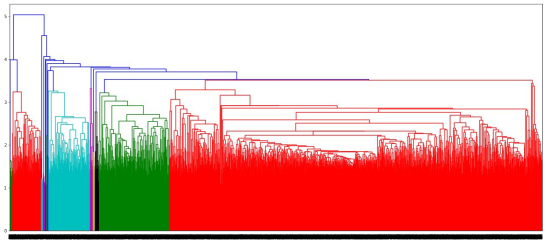


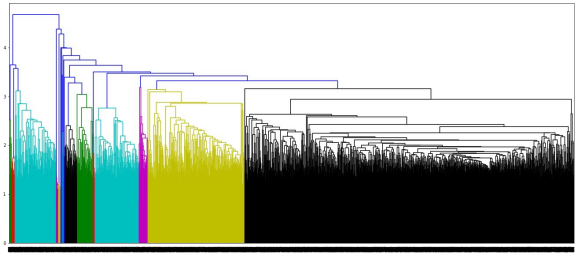


Hierarchical Clustering

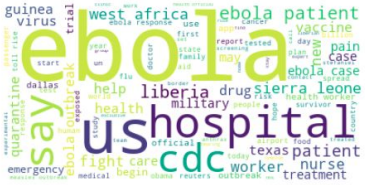
*Muestreo del 10% de instancias.
*SVD reducción 12% dimensionalidad

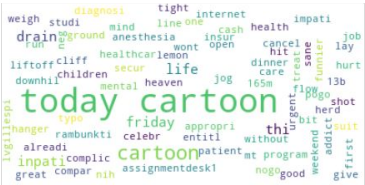
	K óptimos establecidos para los distintos datasets		
Hierarchical	d_stem	d_lem	d_stem_lem
K Óptimo	6	9	5



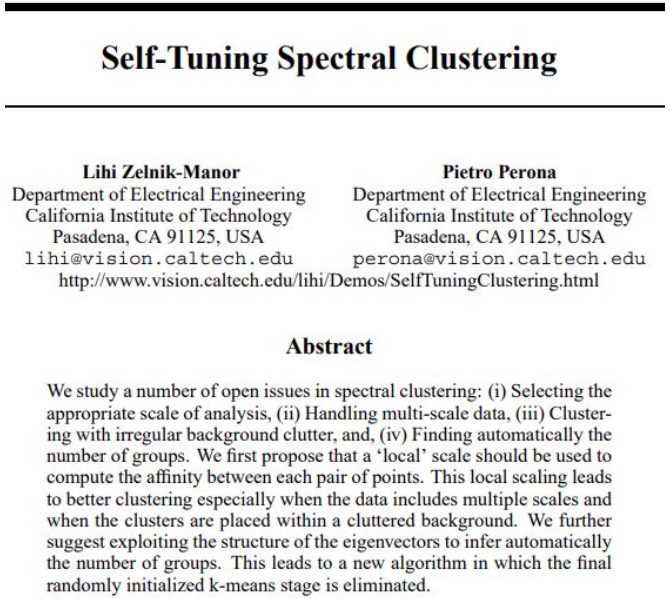


	SELECCIÓN DE ALGORITMO ÓPTIMO PARA LOS DISTINTOS DATASETS		
Hierarchical	d_stem	d_lem	d_stem_lem
Alg. Óptimo	No Óptimo	No Óptimo	Óptimo





Spectral Clustering



Self-Tuning Spectral Clustering

Lihi Zelnik-Manor
Department of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, USA
lihi@vision.caltech.edu

Pietro Perona
Department of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, USA
perona@vision.caltech.edu

http://www.vision.caltech.edu/lihi/Demos/SelfTuningClustering.html

Abstract

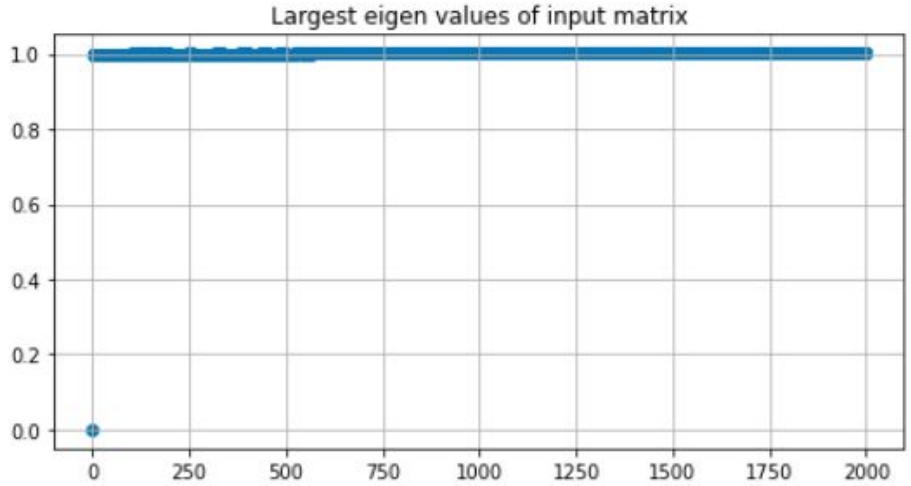
We study a number of open issues in spectral clustering: (i) Selecting the appropriate scale of analysis, (ii) Handling multi-scale data, (iii) Clustering with irregular background clutter, and, (iv) Finding automatically the number of groups. We first propose that a 'local' scale should be used to compute the affinity between each pair of points. This local scaling leads to better clustering especially when the data includes multiple scales and when the clusters are placed within a cluttered background. We further suggest exploiting the structure of the eigenvectors to infer automatically the number of groups. This leads to a new algorithm in which the final randomly initialized k-means stage is eliminated.

- Problema:
  - ¿Cómo determinar el K óptimo?
  - Eigengap heuristic sugiere que el número de conglomerados k generalmente viene dado por el valor de k que maximiza el eigengap (diferencia entre valores propios consecutivos
  - Problemas de agotamiento de recursos en máquina local y Google Colab
    - Dataset completo
    - Muestreo del 10% de instancias

Spectral Clustering

```
1 %%time
2 affinity_matrix = getAffinityMatrix(X_lemmatized_sampled[:2000].toarray(), k = 10)
3 k, _,  _ = eigenDecomposition(affinity_matrix)
4 print(f'Optimal number of clusters for X_lemmatized: {k}')
```

```
Optimal number of clusters for X_lemmatized: [  1 103 118 120 124]
CPU times: user 1min 58s, sys: 4.17 s, total: 2min 2s
Wall time: 1min 52s
```



Conclusión en el dataset existente, el método citado no es convergente

Selección Modelo Óptimo

	SELECCIÓN DE ALGORITMO ÓPTIMO PARA LOS DISTINTOS DATASETS		
Alg. Óptimo	d_stem	d_lem	d_stem_lem
K Means	No Óptimo	Óptimo	No Óptimo
Hierarchical	No Óptimo	No Óptimo	Óptimo
Spectral	-	-	-

K Means

Concluyente.
Instancias 100%
Atributos 100%
Clusters etiquetables

Siempre converge
Especializado en datasets grandes - pequeños - medianos.

Hierarchical

No concluyente.
Instancias 10%
Atributos 12%
Clusters
semi-etiquetables

Calcular y almacenar matrices de distancia n × n.
Dataset muy grandes resulta muy costoso y lento

Spectral

No concluyente
Instancias 10%
No converge

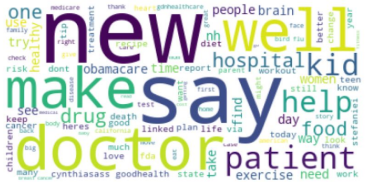
Varios Pasos para matriz de similitud, por eigenvalues.
Es semi - convergente, convergencia no garantizada.

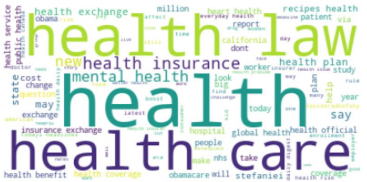
Resultados y discusión

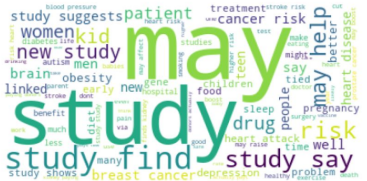
Evaluación de Modelo Óptimo

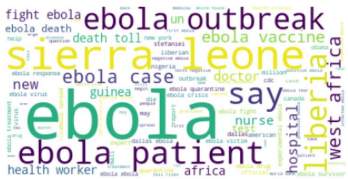
Analisis de WordClouds para etiquetar los clusters.

Clúster	Etiqueta	Cantidad
Cluster 0	Bienestar	39874
Cluster 1	Seguros de Salud-Social	4402
Cluster 2	Medicamentos	2656
Cluster 3	Fitness	3588
Cluster 4	Investigaciones médicas	6067
Cluster 5	Tercera edad	3264
Cluster 6	Ébola	4265









Evaluación clusters por medio de comunicación

Medio Digital	Tópicos Principales
Everyday Health	Bienestar, Fitness, Tercera Edad
CBC Health	Bienestar, Ébola, Seguros de Salud
CNN Health	Bienestar, Investigaciones Médicas, Fitness
Fox News Health	Bienestar, Investigaciones médicas, Ébola
GDN HealthCare	Bienestar, Seguros de salud, Fitness
Good Health	Bienestar, Fitness, Investigaciones médicas
Kaiser Health News	Bienestar, Seguros de Salud, Tercera Edad
LA Times Health	Bienestar, Investigaciones Médicas, Fitness
MSN Times Health	Bienestar, Investigaciones Médicas, Medicamentos
NBC Health	Bienestar, Investigaciones Médicas, Medicamentos
NPR Health	Bienestar, Tercera Edad, Seguros de Salud
NY Times Health	Bienestar, Ébola, Seguros de Salud
Reuters Health	Bienestar, Ébola, Seguros de Salud
US News Health	Bienestar, Fitness, Seguro de Salud, Tercera Edad
WSI Health	Bienestar, Seguros de Salud, Ébola

Caso de análisis

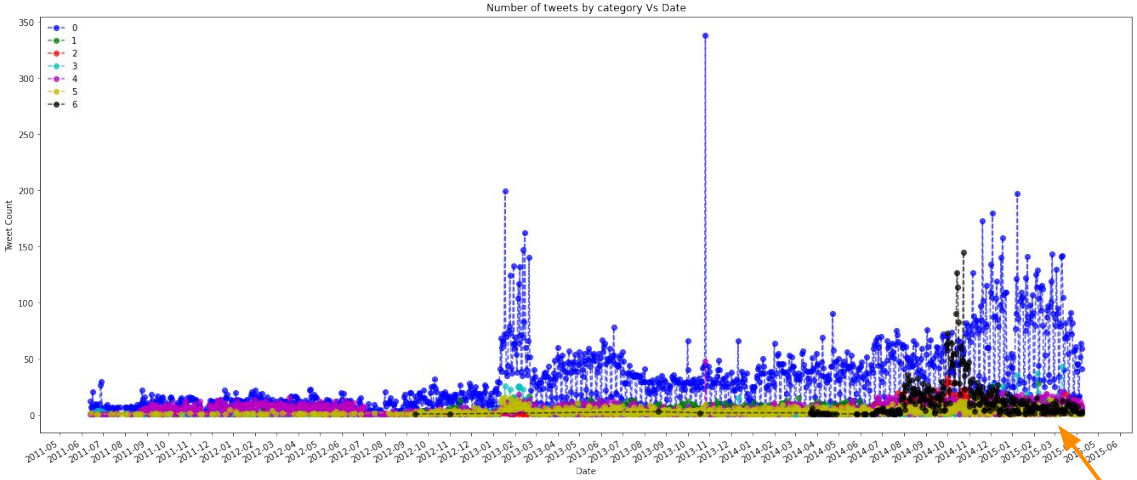
Kaiser Health News formado por:

1. Bienestar
2. Seguros de Salud
3. Tercera Edad

Pública "periodismo sobre temas de **atención médica** concerniente a personas con bajos ingresos, vulnerables al costo de la atención médica, como las personas **sin seguro**, las personas con enfermedades crónicas o Beneficiarios de **Medicaid-Medicare**"[17] (programa de cobertura de seguridad social estatal de EEUU para todas las **personas mayores de 65 años**) [18].

Evaluación clusters por medio de comunicación

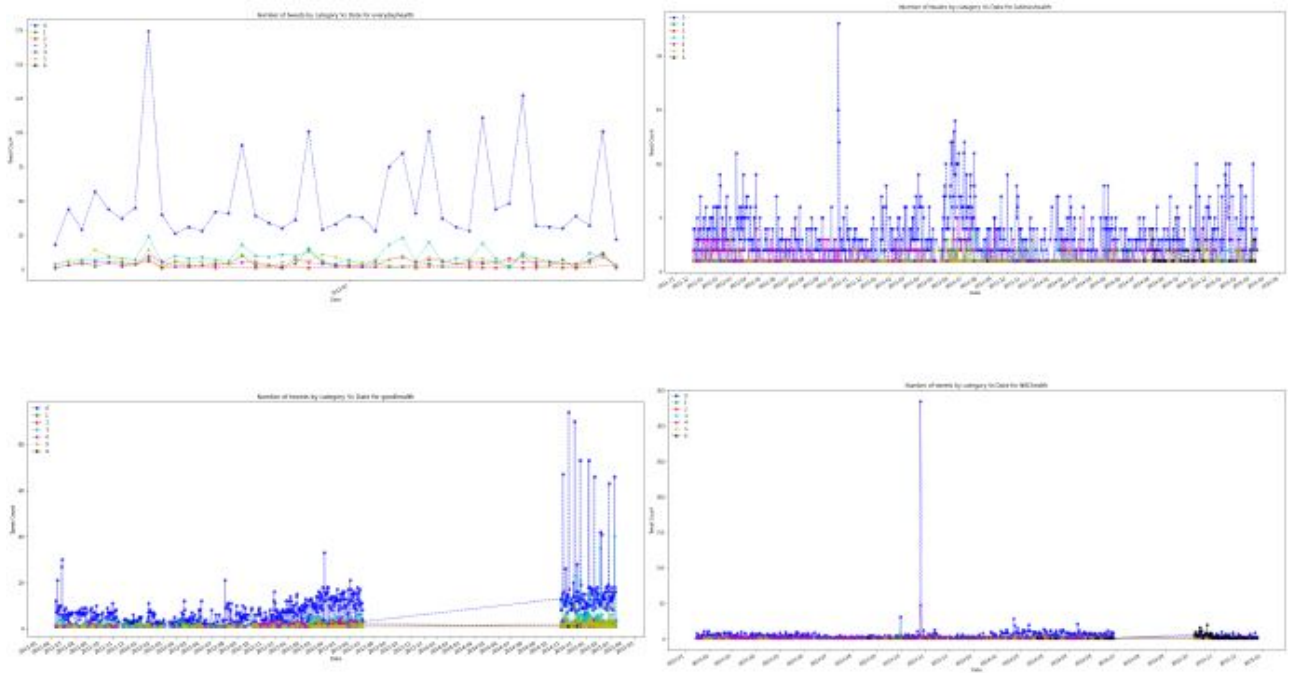
Año	Etiqueta
2011	Bienestar, Investigaciones médicas, Fitness
2012	Bienestar, Investigaciones médicas, Fitness
2013	Bienestar, Seguros de Salud, Investigaciones médicas
2014	Bienestar, Ébola, Investigaciones médicas
2015	Bienestar, Investigaciones médicas, Fitness



Caso de análisis

Tabla y figura 4 establece la variación de tópicos respecto al tiempo (por año y por día).
En 2014 incremento del tópico ébola debido a los brotes en Guinea [19].

Variación de tópicos en el tiempo



http://bit.ly/3s5G9kq

Conclusiones

Proceso metódico para categorizar los tópicos que cubren los medios de comunicación en el área salud en la red social Twitter, además de determinar la variación de estos en el tiempo, mediante la aplicación de algoritmos no supervisados de clusterización y su posterior análisis.

Método de procesamiento para la obtención de las raíces morfológicas con mejor rendimiento fue lematización. Algoritmo de clustering con mejor rendimiento es K Means.

Etapa de exploración de datos evidenció que los medios de comunicación deben establecer mayor continuidad y periodicidad en la publicación de tweets o existió un error en la obtención del dataset.

Hierarchical en dataset muy grandes resulta muy costoso y lento, no converge.

Spectral resulta costoso y lento, k óptimo difícil de obtener, convergencia no garantizada.

Hierarchical requiere un equipo con buenas prestaciones (CPU, GPU, RAM) - Google Colab pago (12GB RAM resulta poco).

Naturaleza del Dataset presenta mejores resultados con K Means.

K Means etiquetado de clusters resulta concluyente.

Dataset presentó ciertas anomalías, tweets repetidos, los medios posteaban el mismo tweet por varios días.

Contribución

Expone el rendimiento de los algoritmos no supervisados de clusterización K Means, Spectral, Jerárquico en datos de tipo texto, en dataset masivos.

Metodología para el análisis de categorías tratadas para la identificación de tópicos principales en cada año y un posible patrón de comportamiento en el uso de los medios de comunicación en redes sociales.

Análisis útil a un gobierno para sustentar la necesidad de atender los brotes en la población, cada cierto tiempo, o a su vez, mejorar los procesos de publicación de noticias en un medio digital, al proveer variedad de temáticas.

Replicable, en dominios relacionados.

Referencias

[1] C. C. Aggarwal and H. Wang, “Text Mining in Social Networks,” in Social Network Data Analytics, Springer US, 2011, pp. 353–378.
[2] C. H. Lee, “Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams,” Expert Syst. Appl., vol. 39, no. 18, pp. 13338–13356, Dec. 2012, doi: 10.1016/j.eswa.2012.05.068.
[3] H. Sheikha, “Text mining Twitter social media for Covid-19 Comparing latent semantic analysis and latent Dirichlet allocation,” 2020. Accessed: Feb. 18, 2021. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:hig:diva-32567.
[4] D. Pohl, A. Bouchachia, and H. Hellwagner, “Online indexing and clustering of social media data for emergency management,” Neurocomputing, vol. 172, pp. 168–179, Jan. 2016, doi: 10.1016/j.neucom.2015.01.084.
[5] K. Singh, H. K. Shakya, and B. Biswas, “Clustering of people in social network based on textual similarity,” Perspect. Sci., vol. 8, pp. 570–573, Sep. 2016, doi: 10.1016/j.pisc.2016.06.023.
[6] X. Dai, M. Bikdash, and B. Meyer, “From social media to public health surveillance: Word embedding based clustering method for twitter classification,” May 2017, doi: 10.1109/SECON.2017.7925400.
[7] A. Alsayat and H. El-Sayed, “Social media analysis using optimized K-Means clustering,” in 2016 IEEE/ACIS 14th International Conference on Software Engineering Research, Management and Applications, SERA 2016, Jul. 2016, pp. 61–66, doi: 10.1109/SERA.2016.7516129.
[8] T. Hussain Shah, N. Naveed, and Z. Rauf, “A Methodology for Brand Name Hierarchical Clustering Based on Social Media Data,” Dec. 2018. doi: 10.36785/JAES.V8I1.238.
[9] Amir Karami, “UCI Machine Learning Repository: Health News in Twitter Data Set.” http://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter (accessed Feb. 18, 2021).
[10] PyData Editors, “pandas - Python Data Analysis Library.” https://pandas.pydata.org/ (accessed Feb. 18, 2021).
[11] Juan González Villa, “TF IDF: herramientas para mejorar la relevancia de tus contenidos - USEO.” https://useo.es/tf-idf-relevancia/ (accessed Feb. 18, 2021).
[12] P. Editors, “nltk · PyPI.” https://pypi.org/project/nltk/ (accessed Feb. 18, 2021).
[13] P. Editors, “wordcloud · PyPI.” https://pypi.org/project/wordcloud/ (accessed Feb. 18, 2021).
[14] Maklin Cory, “Hierarchical Agglomerative Clustering Algorithm Example In Python | by Cory Maklin | Towards Data Science.” https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019 (accessed Feb. 18, 2021).
[15] Abhishek Gupta, “Difference between K means and Hierarchical Clustering - GeeksforGeeks.” https://www.geeksforgeeks.org/difference-between-k-means-and-hierarchical-clustering/ (accessed Feb. 18, 2021).
[16] L. Zelnik-Manor and P. Perona, “Self-Tuning Spectral Clustering.” Accessed: Feb. 18, 2021. [Online]. Available: http://www.vision.caltech.edu/lihi/Demos/SelfTuningClustering.html.

Aportes

Aportes

Abad F. propuesta dataset, metodología del proyecto, preprocesamiento de datos, unificación de datos, búsqueda bibliográfica, clustering KMeans.

Reinozo E. procesamiento de dataset, métodos de evaluación de k óptimos, clustering Hierarchical.

Trabajo en grupo: Análisis exploratorio de dataset, Clustering Spectral, filtración y análisis de resultados, establecimiento de gráficas relevantes para resultados y conclusiones, informe y presentación.

Preguntas