Detección de Textos Similares a través de una Técnica de Agrupamiento Basada en Densidad

Mariano Maisonnave (mariano.maisonnave@cs.uns.edu.ar)

Instituto de Ciencias e Ingeniería de la Computación - CONICET/UNS

Objetivo

Detectar la existencia de eventos “del mundo real” a partir del texto completo de artículos periodísticos.

Objetivo

Detectar la existencia de eventos “del mundo real” a partir del texto completo de artículos periodísticos.

¿Qué es un evento?*

Objetivo

Detectar la existencia de eventos “del mundo real” a partir del texto completo de artículos periodísticos.

¿Qué es un evento?*

Un suceso del “mundo real” representado por un conjunto de noticias

Objetivo

Detectar la existencia de eventos “del mundo real” a partir del texto completo de artículos periodísticos.

¿Qué es un evento?*

Un suceso del “mundo real” representado por un conjunto de noticias donde se trata el evento dentro de una ventana temporal acotada.

Objetivo

Detectar la existencia de eventos “del mundo real” a partir del texto completo de artículos periodísticos.

¿Qué es un evento?*

Un suceso del “mundo real” representado por un conjunto de noticias donde se trata el evento dentro de una ventana temporal acotada.

* Within TDT, a topic is defined to be a set of news stories that are strongly related by some seminal real-world event. The topic begins at a set time, and is probably no longer reported in the news at some point [2002, Allan].

Objetivo

Detectar la existencia de eventos “del mundo real” a partir del texto completo de artículos periodísticos.

¿Qué es un evento?*

Un suceso del “mundo r
se trata el evento dentr

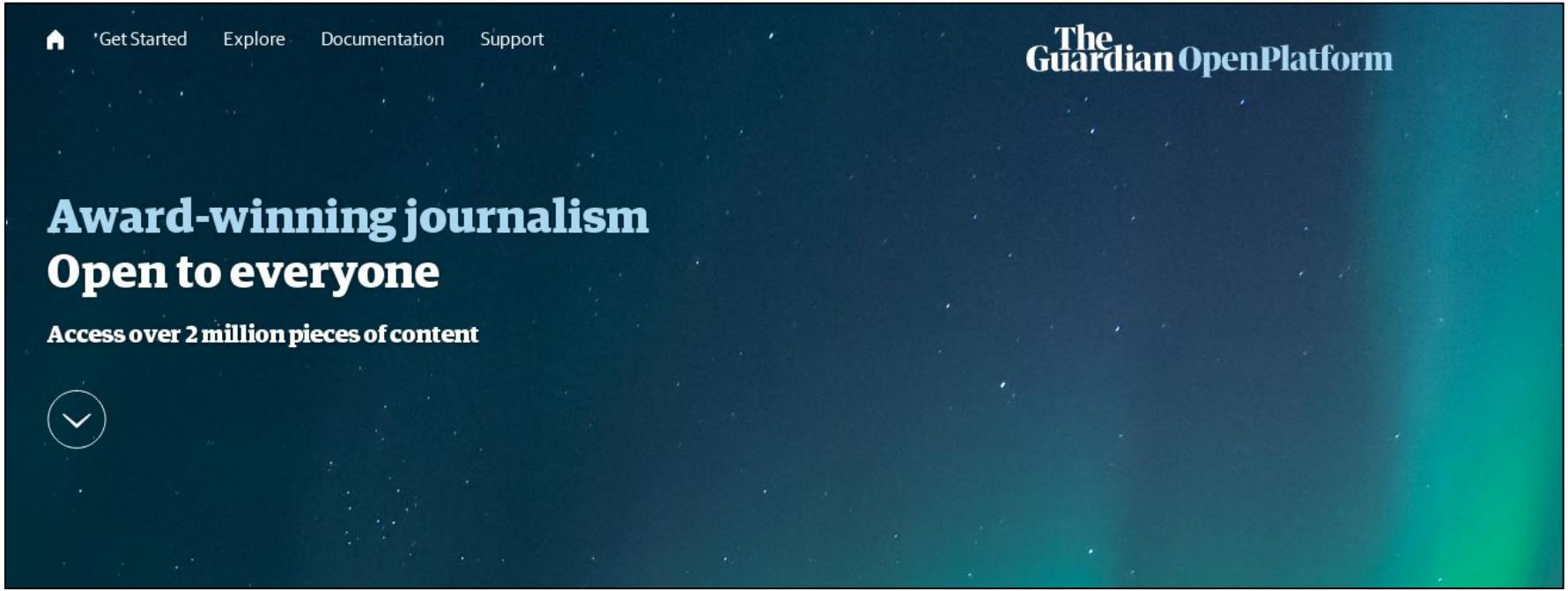

Horsemeat scandal leaves Burger King facing a whopping backlash

Burger giant forced to take out adverts in the national press apologising for error as thousands of consumers complain on Facebook and Twitter

7

* Within TDT, a topic is defined to be a set
a set time, and is probably no longer report

El Conjunto de Datos



De Textos a Matriz

1. Filtro de HTML y stopwords.
2. Stemming.
3. Eliminación Palabras Infrecuentes.
4. Escalado y Transformación TF-IDF.

```
<p>The government cannot be sure there is no
safety risk from supermarket beef products that
have been found  to contain horse DNA, the head
of the UK&apos;s leading official food control
laboratory has told the Guardian.</p>
```

Paso	Método	Palabras restantes en el dataset
1	Eliminación de stopwords	37.741 palabras
2	Stemming	30.699 palabras
3	Eliminación palabras infrecuentes	**8.291 palabras**

De Textos a Matriz

1. Filtro de HTML y stopwords.
2. Stemming.
3. Eliminación Palabras Infrecuentes.
4. Escalado y Transformación TF-IDF.

```
<p>The government cannot be sure there is no
safety risk from supermarket beef products that
have been found  to contain horse DNA, the head
of the UK&apos;s leading official food control
laboratory has told the Guardian.</p>
```

Paso	Método	Palabras restantes en el dataset
1	Eliminación de stopwords	37.741 palabras
2	Stemming	30.699 palabras
3	Eliminación palabras infrecuentes	**8.291 palabras**

De Textos a Matriz

1. Filtro de HTML y stopwords.
2. Stemming.
3. Eliminación Palabras Infrecuentes.
4. Escalado y Transformación TF-IDF.

```
The government cannot be sure there is no
safety risk from supermarket beef products that
have been found  to contain horse DNA, the head
of the UK     's leading official food control
laboratory has told the Guardian.
```

Paso	Método	Palabras restantes en el dataset
1	Eliminación de stopwords	37.741 palabras
2	Stemming	30.699 palabras
3	Eliminación palabras infrecuentes	**8.291 palabras**

De Textos a Matriz

1. Filtro de HTML y stopwords.
2. Stemming.
3. Eliminación Palabras Infrecuentes.
4. Escalado y Transformación TF-IDF.

```
   The government cannot be sure there is no
safety risk from supermarket beef products that
have been found  to contain horse DNA, the head
of the UK     's leading official food control
laboratory has told the Guardian.
```

Paso	Método	Palabras restantes en el dataset
1	Eliminación de stopwords	37.741 palabras
2	Stemming	30.699 palabras
3	Eliminación palabras infrecuentes	**8.291 palabras**

De Textos a Matriz

1. Filtro de HTML y stopwords.
2. Stemming.
3. Eliminación Palabras Infrecuentes.
4. Escalado y Transformación TF-IDF.

```
        government           sure
safety risk      supermarket beef products
           found     contain horse DNA      head
        UK     's leading official food control
laboratory     told     Guardian
```

Paso	Método	Palabras restantes en el dataset
1	Eliminación de stopwords	37.741 palabras
2	Stemming	30.699 palabras
3	Eliminación palabras infrecuentes	**8.291 palabras**

De Textos a Matriz

1. Filtro de HTML y stopwords.
2. Stemming.
3. Eliminación Palabras Infrecuentes.
4. Escalado y Transformación TF-IDF.

```
        government          sure
safety risk      supermarket beef products
           found     contain horse DNA      head
        UK     's leading official food control
laboratory     told     Guardian
```

Paso	Método	Palabras restantes en el dataset
1	Eliminación de stopwords	37.741 palabras
2	Stemming	30.699 palabras
3	Eliminación palabras infrecuentes	**8.291 palabras**

De Textos a Matriz

1. Filtro de HTML y stopwords.
2. Stemming.
3. Eliminación Palabras Infrecuentes.
4. Escalado y Transformación TF-IDF.

```
        government            sure
safety risk      supermarket beef product~~s~~
      ~~found~~find contain horse DNA      head
      UK     's lead~~ing~~ official food control
laboratory     ~~told~~tell  Guardian
```

Paso	Método	Palabras restantes en el dataset
1	Eliminación de stopwords	37.741 palabras
2	Stemming	30.699 palabras
3	Eliminación palabras infrecuentes	**8.291 palabras**

De Textos a Matriz

1. Filtro de HTML y stopwords.
2. Stemming.
3. Eliminación Palabras Infrecuentes.
4. Escalado y Transformación TF-IDF.

```
       government           sure
safety risk      supermarket beef product
               find contain horse DNA      head
       UK     's lead    official food control
laboratory         tell  Guardian
```

Paso	Método	Palabras restantes en el dataset
1	Eliminación de stopwords	37.741 palabras
2	Stemming	30.699 palabras
3	Eliminación palabras infrecuentes	**8.291 palabras**

Matriz Final

$$\mathbb{X}_{n\times p} = \begin{array}{c} \\ doc_1 \\ doc_2 \\ \\ doc_n \end{array} \begin{array}{c} \begin{array}{cccccc} government & dna & horse & \dots & poll & war \end{array} \\ \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \dots & x_{1,p-1} & x_{1,p} \\ x_{2,1} & x_{2,2} & x_{2,3} & \dots & x_{2,p-1} & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \dots & x_{n,p-1} & x_{n,p} \end{pmatrix} \end{array}$$

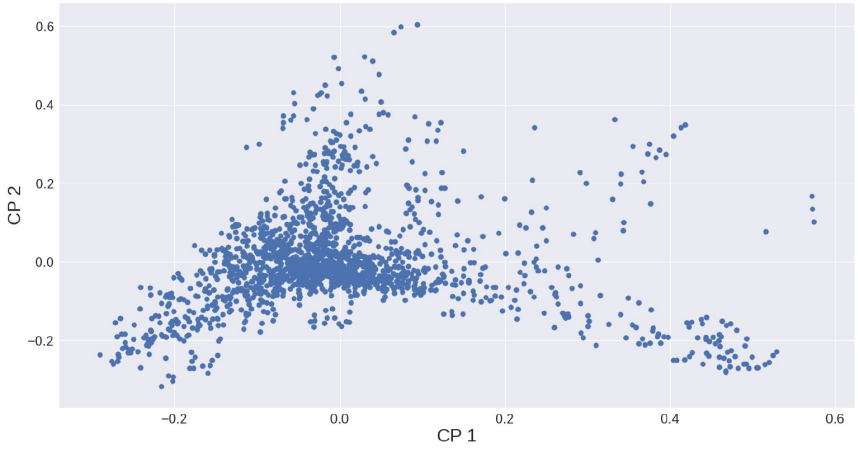
Matriz Final

$$\mathbb{X}_{n\times p} = \begin{array}{c} \\ doc_1 \\ doc_2 \\ \\ doc_n \end{array} \begin{array}{c} \begin{array}{cccccc} government & dna & horse & \dots & poll & war \end{array} \\ \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \dots & x_{1,p-1} & x_{1,p} \\ x_{2,1} & x_{2,2} & x_{2,3} & \dots & x_{2,p-1} & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \dots & x_{n,p-1} & x_{n,p} \end{pmatrix} \end{array}$$

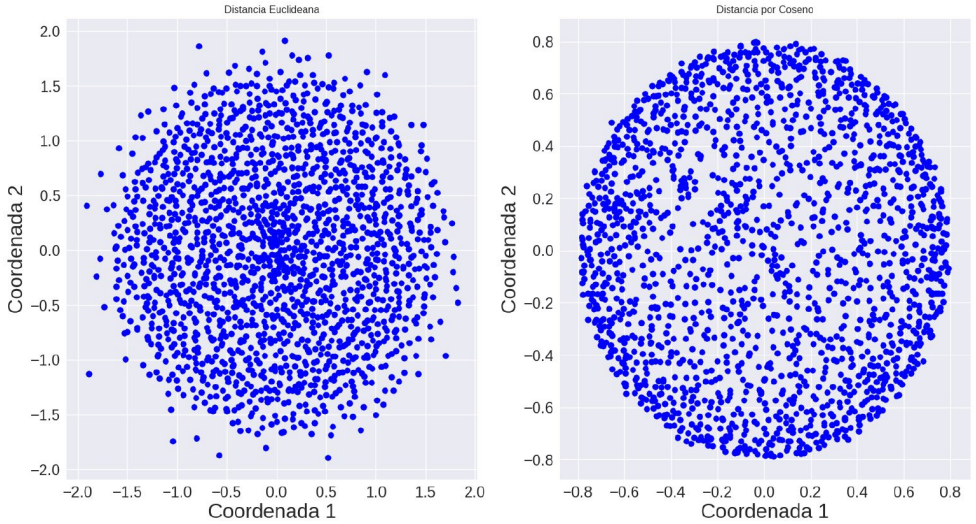
$$n = 1689, p = 8291$$

Proyección y Visualización de Instancias (lineal)

PCA



MDS



Proyección y Visualización de Instancias (t-SNE)

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

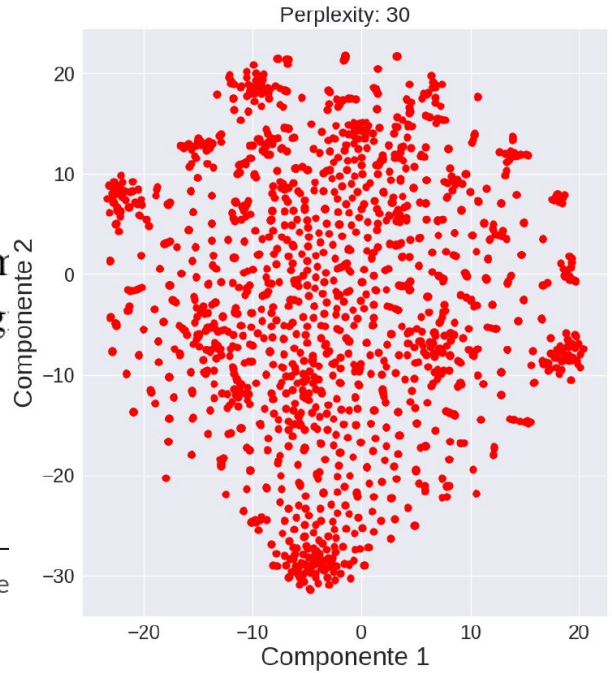
The minimization of the cost function in Equation 2 is performed using a gradient descent method. The gradient has a surprisingly simple form

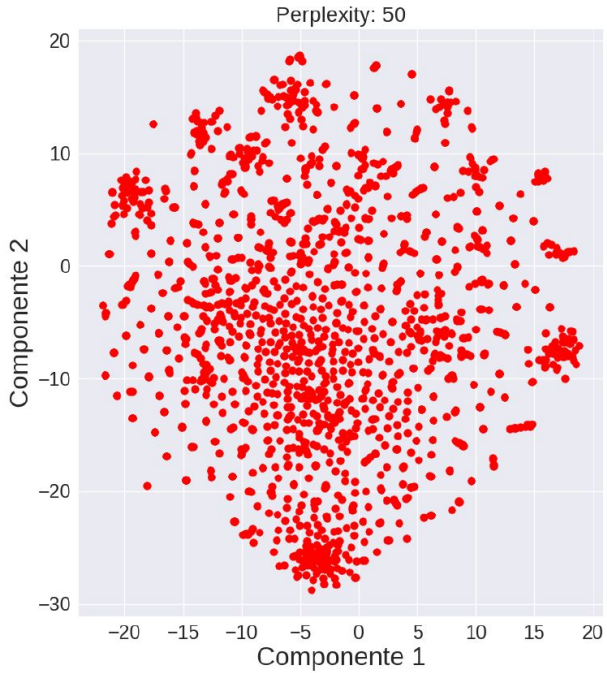
$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j).$$

Maaten, L.V.D. and Hinton, G., 2008. Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), pp.2579-2605.

Proyección y Visualización de Instancias (t-SNE)

The minim adient descent
method. The g





.2579-2605.

Maate

Density-based spatial clustering of applications with noise (DBSCAN)

¿Por qué DBSCAN?

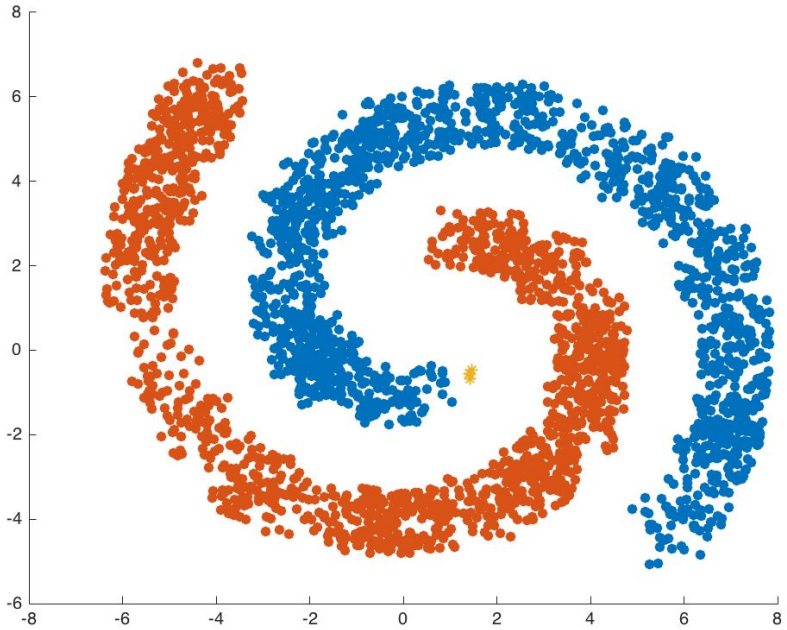
- La cantidad de clusters se define automáticamente.
- Permite la existencia de noticias que no pertenecen a ningún cluster (noticias no relacionadas a ningún evento).
- Permite que la noción de relevancia cambie en el tiempo (tópicos dinámicos).

Sander, J., Ester, M., Kriegel, H.P. and Xu, X., 1998. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data mining and knowledge discovery, 2(2), pp.169-194.

Density-based spatial clustering of applications with noise (DBSCAN)

¿Por qué DBSCAN?

- La cantidad de
- Permite la exis
  ningún eventc
- Permite que la

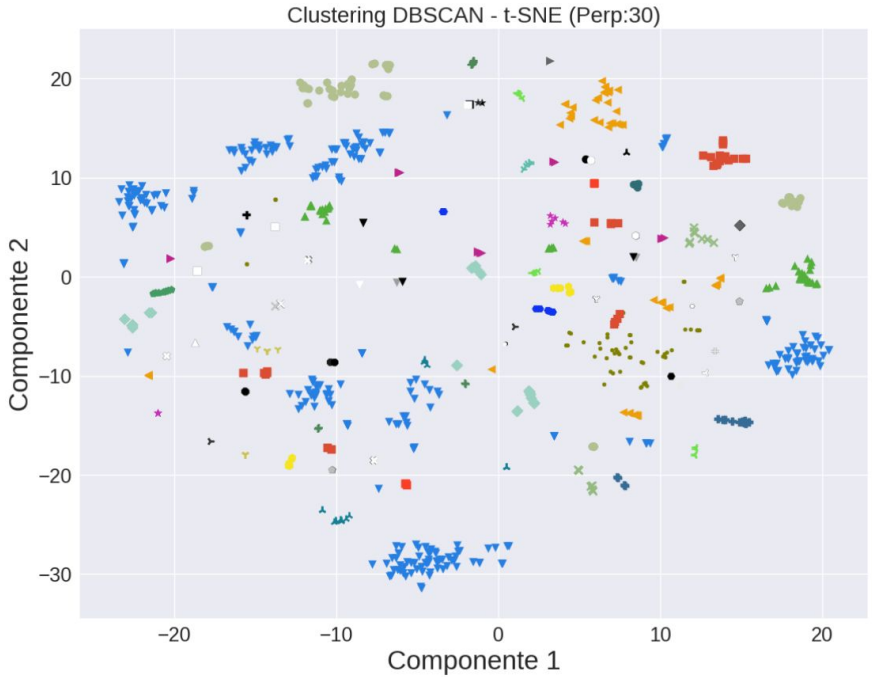


o relacionadas a

Sander, J., Ester, M
applications. Data

gorithm gdbscan and its

Agrupamiento de Noticias Resultante



Algunos clusters

Cuadro 5: Los cinco términos más importantes de cada cluster en orden decreciente.

Cluster	tam	$term_1$	$term_2$	$term_3$	$term_4$	$term_5$
$cluster_0$	3	pm	rate	bring	poll	lift
$cluster_1$	8	anger	unit	trust	hospital	lewisham
$cluster_2$	17	scandal	hors	supermarket	horsemeat	burger
$cluster_3$	84	algeria	hostage	algerian	eurozone	crisis
$cluster_4$	5	russia	syria	strike	israeli	air
$cluster_5$	3	right	begin	fresh	stay	film

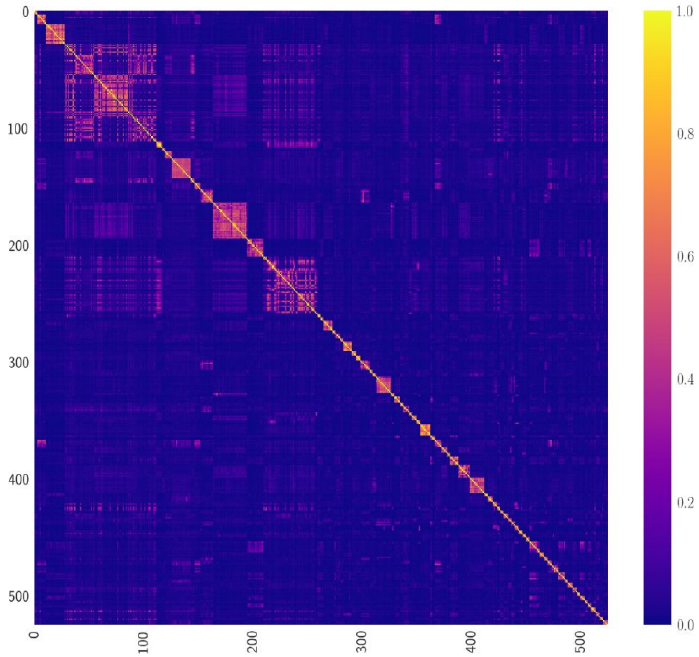
Algunos clusters

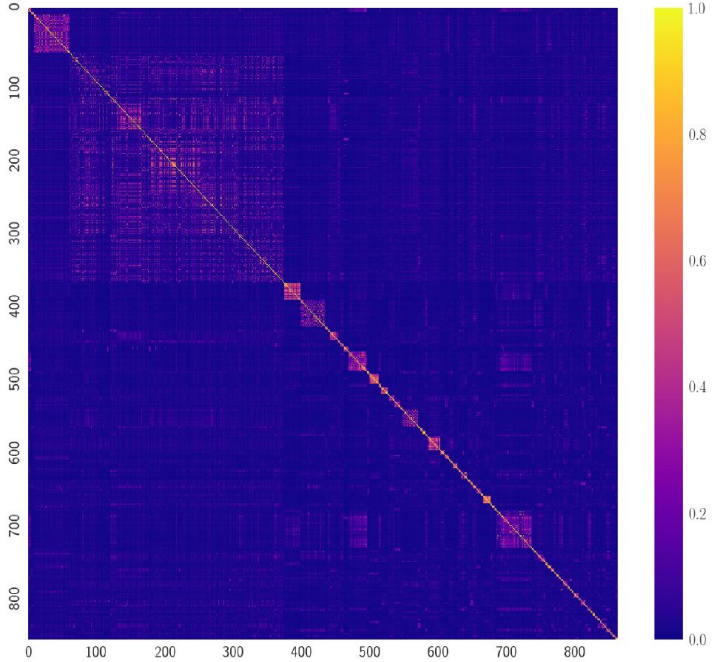
Cuadro 5: Los cinco términos más importantes de cada cluster en

Cuadro 6: Los títulos de la primer y última noticia de cada cluster.

Cluster	Fecha	Titulo
$cluster_0$	31-01	Tories tell PM: lift poll ratings or face revolt
	27-01	Adam Afriyie denies plot to bring down David Cameron
$cluster_1$	31-01	Hunt 'risking future of smaller hospitals' with Lewisham ruling
	08-01	Indebted NHS hospital trust should be dissolved and replaced, says report
$cluster_2$	31-01	Burger King reveals its burgers were contaminated in horsemeat scandal
	16-01	Horse DNA found in beefburgers from four major supermarkets
$cluster_3$	31-01	Eurozone crisis live: Spanish PM accused of secret payments
	02-01	Fiscal cliff deal: Markets soar as compromise agreed
$cluster_4$	31-01	Israel faces repercussions of air strike on Syria
	06-01	Israel to build border fence between Golan Heights and Syria
$cluster_5$	31-01	North Dakota battle over reproductive rights begins with anti-abortion hearing
	24-01	Mississippi's sole abortion clinic faces fresh legal fight to stay open

Validación Visual de los Clusters





Conclusiones

Conclusiones

- La similitud entre textos usando bag-of-word puede utilizarse para detectar eventos
- Características adicionales como la componente temporal deben ser incorporadas para diferenciar de técnicas de clustering por tópico.
- Se necesita de una técnica de clustering que tenga en cuenta un posible cambio de relevancia en el tiempo.

Trabajo futuro

- Incorporar componente de decaimiento temporal en la técnica de clustering.

Muchas Gracias!
Preguntas?

Mariano Maisonnnave

mariano.maisonnave@cs.uns.edu.ar

Referencias (I)

- Maaten, L.V.D. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), pp.2579-2605.
- Sander, J., Ester, M., Kriegel, H.P. and Xu, X., 1998. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, *2*(2), pp.169-194.
- Porter, M.F., 1980. An algorithm for suffix stripping. *Program*, *14*(3), pp.130-137.
- Sparck Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), pp.11-21.
- Allan, J. ed., 2012. *Topic detection and tracking: event-based information organization* (Vol. 12). Springer Science & Business Media.

Referencias (II) - Datasets de Eventos

- Pustejovsky, James, et al. TimeBank 1.2 LDC2006T08. Web Download. Philadelphia: Linguistic Data Consortium, 2006.
- Allan, James, et al. TDT Pilot Study Corpus LDC98T25. Web Download. Philadelphia: Linguistic Data Consortium, 1998.
- Walker, Christopher, et al. ACE 2005 Multilingual Training Corpus LDC2006T06. Web Download. Philadelphia: Linguistic Data Consortium, 2006.