ALGUNAS HERRAMIENTAS INFORMÁTICAS PARA ABORDAR CON ÉXITO EL ÉXITO DE LA PARTICIPACIÓN CIUDADANA

A medida que los procesos de participación democrática se van extendiendo el propio crecimiento de las propuestas ciudadanas se convierte en un reto para proponentes, votantes y gestores públicos.

Para los proponentes es difícil saber si ya existe alguna propuesta similar a la suya. Los votantes quedan abrumados ante la avalancha de propuestas para votar. Y los gestores públicos tienen dificultades para tener una visión general de qué están demandando la ciudadanía. Otro problema que surge es, por ejemplo, la dispersión del voto entre propuestas similares, lo que impide que ninguna de ellas alcance el umbral necesario para ser considerada.

Se describen a continuación algunas tecnologías (procesamiento de lenguaje natural y grafos) que pueden ayudar a superar este reto. El caso de uso es su aplicación a una descarga de 18.160 propuestas ciudadanas del Ayuntamiento de Madrid.

Estas herramientas pueden descubrir automáticamente los temas que proponen los ciudadanos, ofrecer una visión general, seguir su evolución, agregar propuestas por la similitud de su contenido, encontrar propuestas semejantes a una dada, etc.

Estas herramientas son eslabones que se pueden incorporar a la informática de apoyo a los procesos de participación ciudadana.

La materia prima son palabras

La clave de esta propuesta es que las gotas de esta cascada de propuesta ciudadanas son las palabras. Su número es tan grande que su comprensión está ya fuera de las capacidades humanas. Se necesitan herramientas automáticas, pero los ordenadores pueden comprender programas o bases de datos, pero no el lenguaje humano.

El procesamiento de lenguaje natural

El procesamiento de lenguaje natural (NLP) es un conjunto diverso de tecnologías que va jalonado el camino hacia la comprensión automática del lenguaje humano. Permiten, entre otras cosas, explotar automáticamente este volumen de información textual que ya es inabordable.

Primer paso: crear las bolsas de palabras (BoW)

Para este experimento es precisa una etapa preliminar de extracción, transformación y carga (ETL) de las fuentes textuales, en este caso una descarga de 18.160 propuestas ciudadanas del Ayuntamiento de Madrid. Transforma esta materia prima textual en la entrada adecuada para la cadena de procesamiento NLP que le sigue. El tamaño del texto es relevante, ya que si es muy corto no es estadísticamente significativo e introduce un “ruido” que falsea los resultados. Por eso se podrían añadir a las propuestas los comentarios que suscitan, que extiende el texto que habla de la propuesta.

Tecnología: Desarrollo ad hoc. En una fase más consolidada se puede emplear TALEND.

A continuación, se pasa por una cadena de procesamiento NLP que realiza estas funciones:

- Segmentación (*Tokenization*), que transforma la secuencia de ceros y unos en los elementos del texto (palabras, números, símbolos, frases, párrafos,…).
- Análisis morfosintáctico (POST o *Part-of-Speech Tagging*), en que a cada elemento del texto se le asigna una categoría sintáctica (sustantivo, adjetivo, verbo y se le añade información morfosintáctica: número, género, tiempo verbal,…).
- Lematización (*Lemmatisation*) o reducción de todas la formas de una palabra a una forma única (sin género, número, tiempo verbal, …).
- Identificación de n-gramas, que forman una unidad léxica (“hombre rana”, “caer en la cuenta”).
- Identificación de la acepción empleada (*Word Sense Disambiguation*) a cada elemento que pueda tener varias acepciones.
- Selección de sustantivos, adjetivos y verbos, eliminación de *stopwords* (palabras que se eliminan del análisis porque no aporta información, sino ruido).

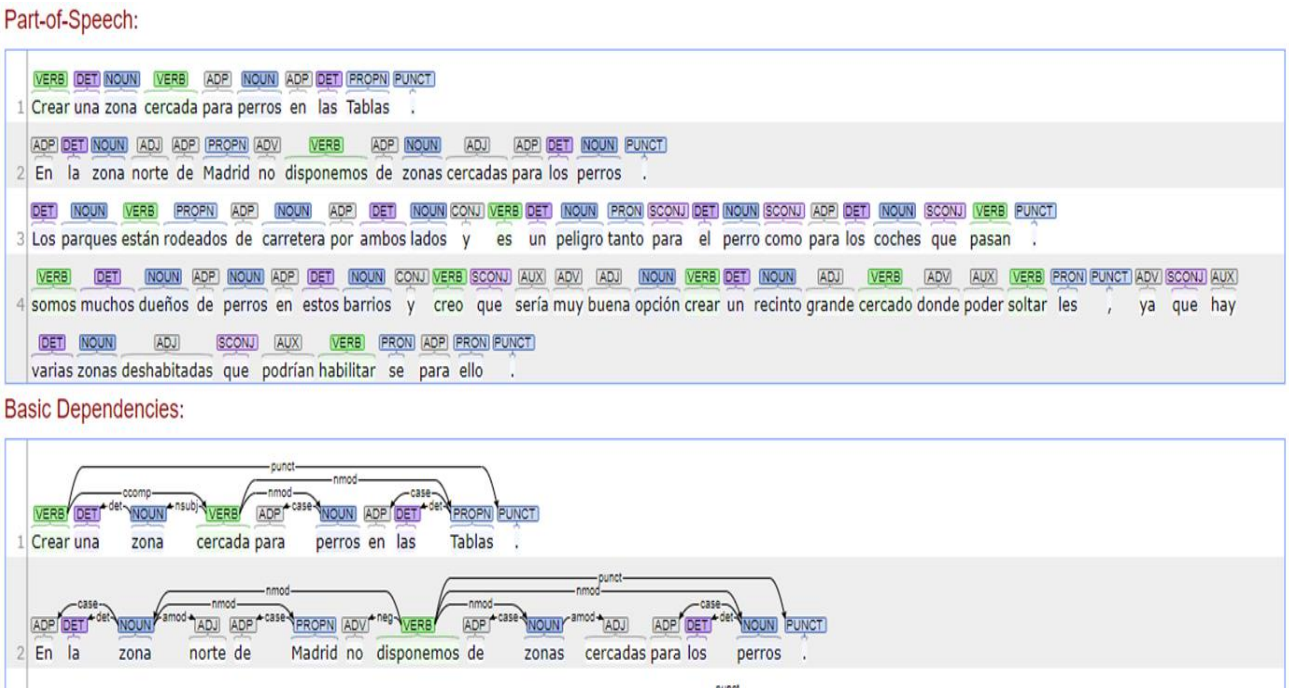

Ilustración 1: Ejemplo de análisis morfosintáctico.

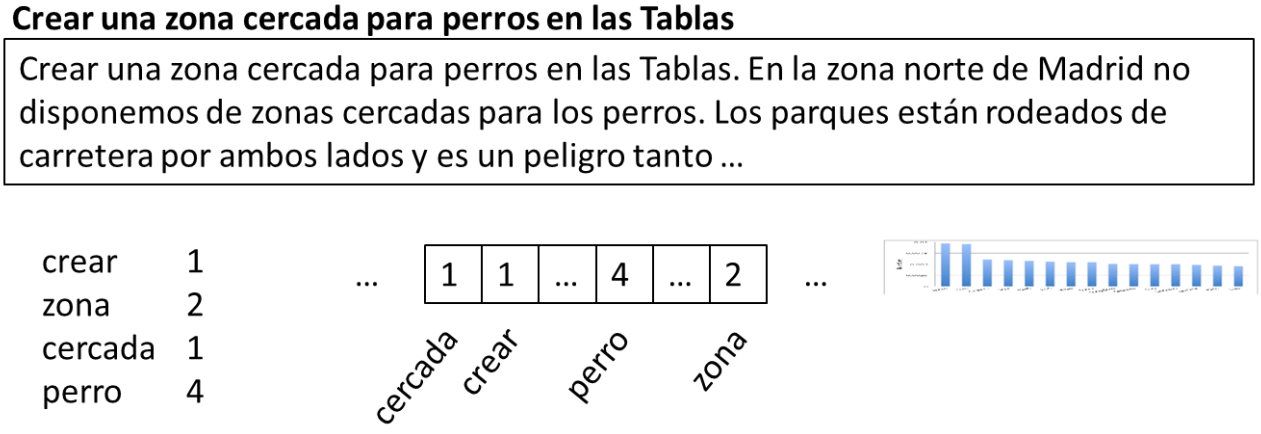

Ilustración 2: La bolsa de palabras son los vectores de frecuencias de los lemas en cada documento.

Tecnología: Librería abierta IXA pipes. También se puede utilizar otra herramienta libre: Freeling, que además incorpora catalán y otras lenguas cooficiales. Se descartó NLTK.

Caracterización temática automática (*Topic modelling*)

El elemento clave y distintivo de este experimento es la implementación del algoritmo *Latent Dirichlet Allocation* (LDA)[i]. A grandes rasgos, esta técnica permite caracterizar un corpus o conjunto de textos (por ejemplo, el volcado de propuestas ciudadanas) en función de un conjunto finito de "temas" que se detectan automáticamente. Cada "tema" es un vector de números (una distribución de probabilidad) que cuantifica la probabilidad en ese tema de cada una de las palabras del léxico o conjunto de palabras del corpus. Una vez obtenidos los vectores que caracterizan cada tema, es posible, a su vez, caracterizar el corpus, cada texto e incluso un nuevo texto (que comparta léxico) con otro vector (una distribución de probabilidad) que cuantifica la probabilidad de que el corpus, el texto del corpus o el nuevo texto aborde esos temas. Lo interesante es que este vector de probabilidades (*topic vector*) se convierte, por tanto, en una especie de huella dactilar o firma del contenido temático del documento que lo caracteriza. Esta caracterización del documento, además de ser automática, es mucho más expresiva que las clasificaciones, que asignan a cada documento un único tema.

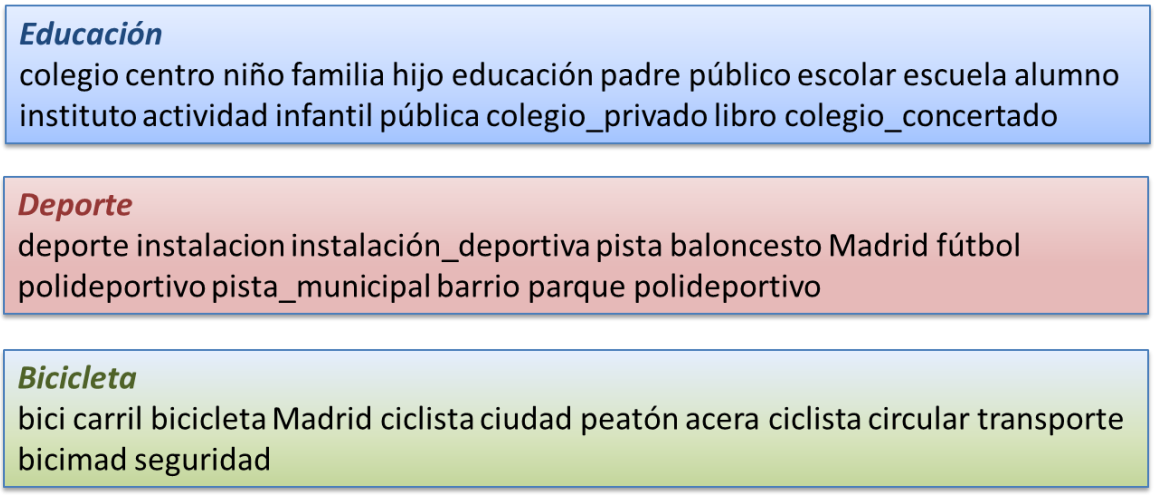

Ilustración 3: Ejemplos de "tópicos". Se representan en orden descendente las palabras más frecuentes en tres temas o "tópicos".

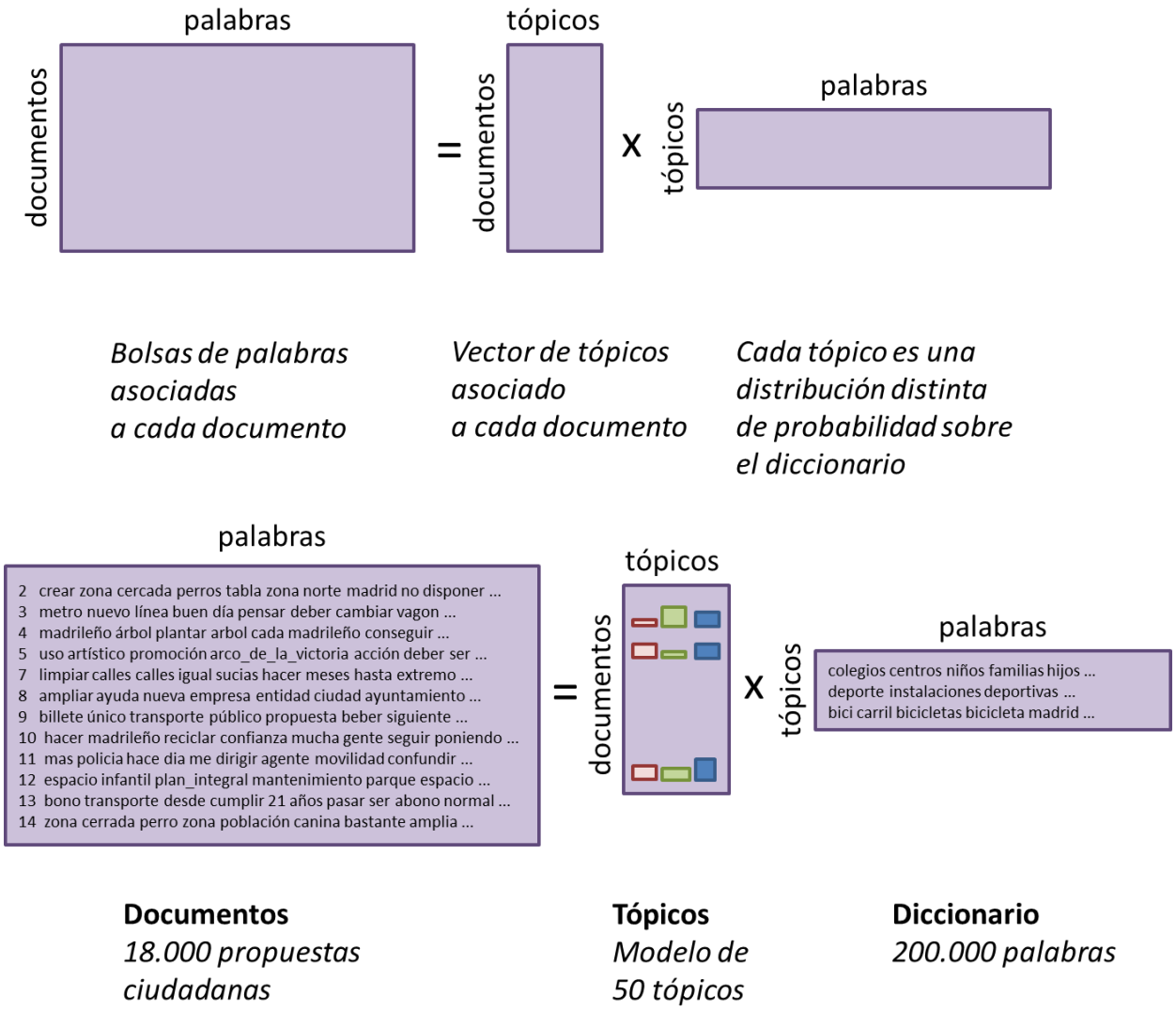

Ilustración 4: Esquema del modelo de tópicos.

Tecnología: Existe versión abierta desarrollada para el proyecto Corpus Viewer (Utiliza la librería abierta Mallet).

Cálculo de la distancia temática

Convertidos en vectores, definir una métrica temática es inmediato, lo que nos permite calcular distancias temáticas entre textos. Cuanto más cercanos, más parecido será su contenido temático.

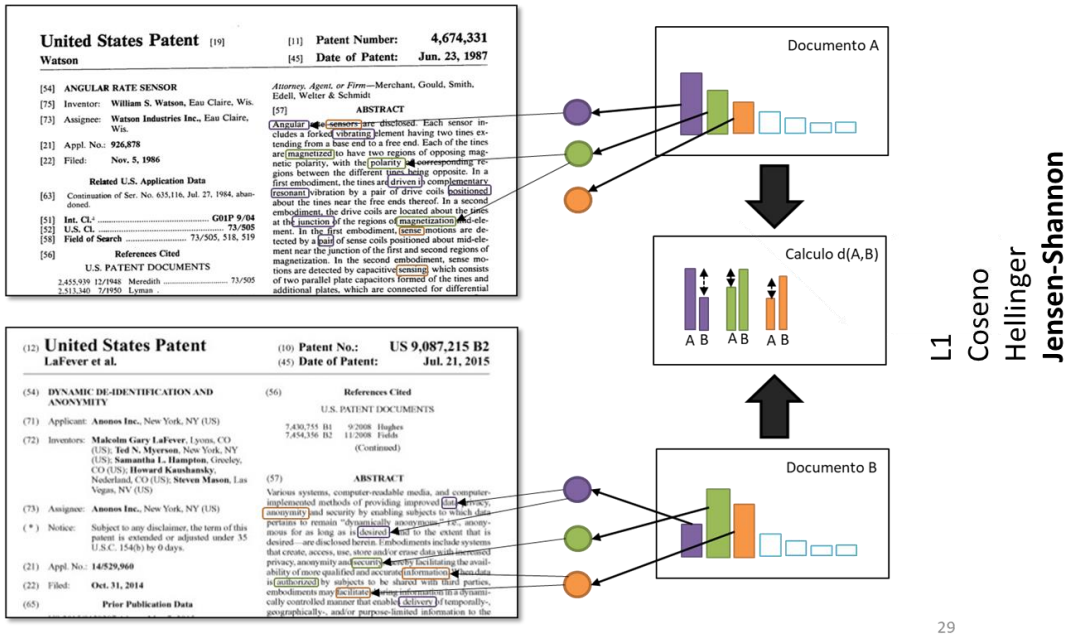

Ilustración 5: Cálculo de la distancia temática entre documentos.

Tecnología: Desarrollo propio abierto. Emplea la distancia de Jensen-Shanon.

¿Para qué puede servir todo esto?

Algunas utilidades:

- Buscar automáticamente las propuestas más cercanas por su contenido temático a una determinada en el conjunto de propuestas (se trata de buscar vectores similares, algo que un ordenador puede hacer muy rápidamente). Esto puede ayudar al proponente a comprobar rápidamente si ya existen propuestas similares a su propia propuesta.
- Ayudar a los votantes a buscar propuestas por temas de forma semánticamente más ajustada que por palabras o palabras clave.
- Ayudar a los gestores públicos y a los ciudadanos a tener una visión de conjunto de los temas planteados por la ciudadanía en sus propuestas.
- Ayudar a los gestores públicos y a los ciudadanos a ver la evolución temporal de los temas que preocupan a la ciudadanía.
- Ayudar a los gestores públicos y a los ciudadanos a buscar solapamientos y sinergias entre las propuestas con el fin de propiciar la agregación de propuestas.

Un paso más allá: visualizar las distancias temáticas

Por tanto, utilizando modelos de temas LDA es posible calcular automáticamente la distancia temática entre las propuestas. A pesar de la cifra (18.160 propuestas), los ordenadores lo hacen con bastante rapidez. Pero surge un nuevo reto: hay 164.883.720 pares de distancias. Es fácil encontrar las propuestas más parecidas (ordenar las distancias y seleccionar las más pequeñas), pero ¿Cómo podemos tener una idea de cómo se relacionan las propuestas entre sí por su contenido temático? ¿Cómo navegar en esta matriz de distancias? ¿Cómo ver grupos de propuestas similares?

Existe una larga tradición de métodos de *clustering* (k-means, ISODATA, etc.) que nos permiten clasificar las propuestas en conjuntos de propuestas similares (*clusters*). Estos métodos tienen algunos problemas bien conocidos, cómo estimar de antemano el número

de clusters que hay, el ruido introducido por los valores atípicos, la complejidad computacional, etc. Sin embargo, el problema de visualizar esas distancias temáticas sigue existiendo. En este espacio métrico, las propuestas son puntos en un espacio multidimensional cuya dimensión es el número de temas descubiertos con el algoritmo LDA. Este número es mayor (mucho mayor) que 2. Así que es imposible representarlo en 2 dimensiones sin distorsión. Existen algoritmos para minimizar esta distorsión (PCA, Multi-dimensional Scaling,...) pero dado el gran número de propuestas y el número de temas la distorsión es tan grande como para que resulte inútil.

Una vez explorados esos caminos, finalmente tomamos un enfoque diferente: los grafos.

Grafos

Los grafos son estructuras matemáticas compuestas por nodos y líneas que unen pares de nodos. La idea es que por debajo de algún umbral (un parámetro que podemos elegir) de similitud temática (o, inversamente, por encima de algún umbral de distancia temática) podemos considerar que dos propuestas no están relacionadas, y por encima de este umbral que están relacionadas. Así que tenemos un grafo de propuesta (nodos) donde se conectan aquellos pares de propuestas que están relacionadas temáticamente.

De esta manera descartamos una gran cantidad de información irrelevante y ruidosa, pero sigue siendo equivalente a sustituir en la matriz de similitudes los valores de similitud por debajo un umbral por cero (o distancias sobre un umbral por infinito).

Pero los grafos van más allá y son útiles al menos para las siguientes funciones:

- Es posible tener una representación visual de las relaciones temáticas de las propuestas ya que se levanta la condición de preservación de distancia y sólo se necesita representar la relación. Existen varios algoritmos para visualizar grafos de manera significativa.
- Esta representación también ayuda a la navegación en las constelaciones de propuestas.
- El método Louvain[ii] para la detección de comunidades ofrece una alternativa prometedora a los métodos de clustering para identificar automáticamente conjuntos de propuestas similares (Gephi lo incluye).

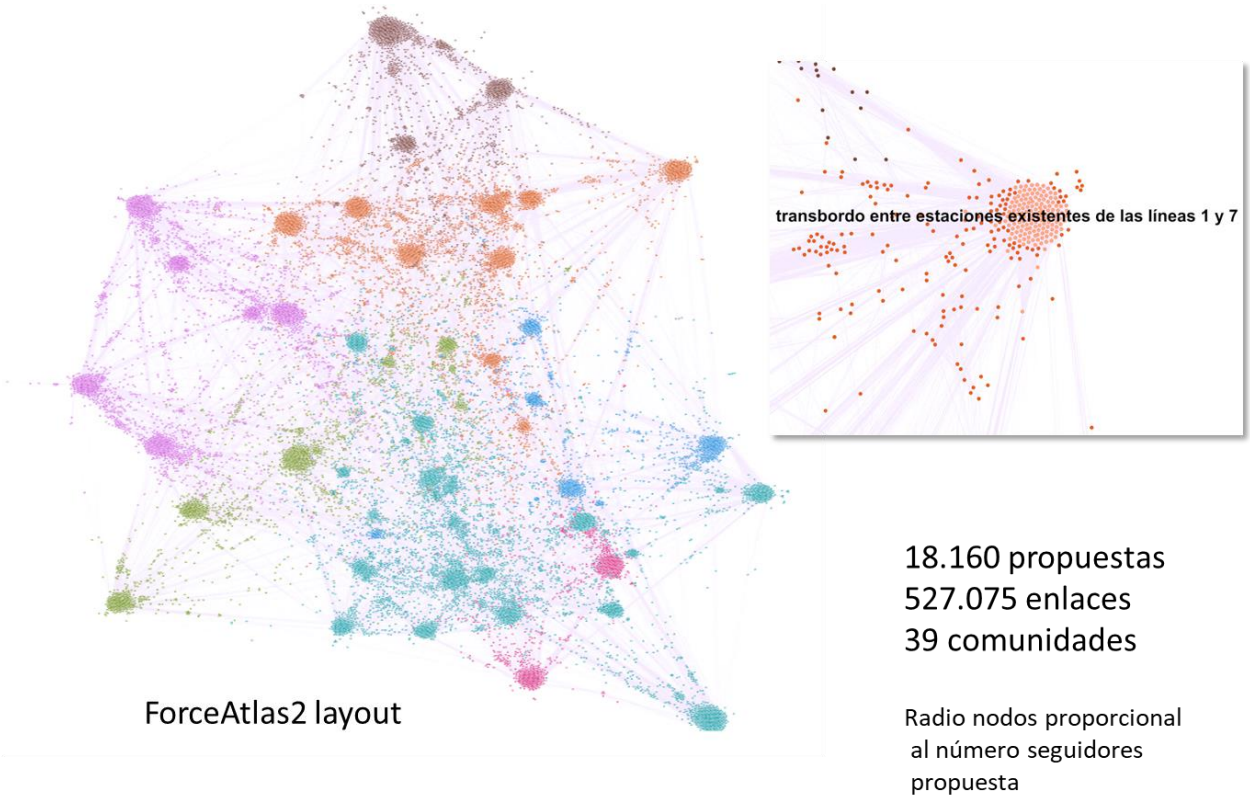

Ilustración 6: Grafo de propuestas ciudadanas.

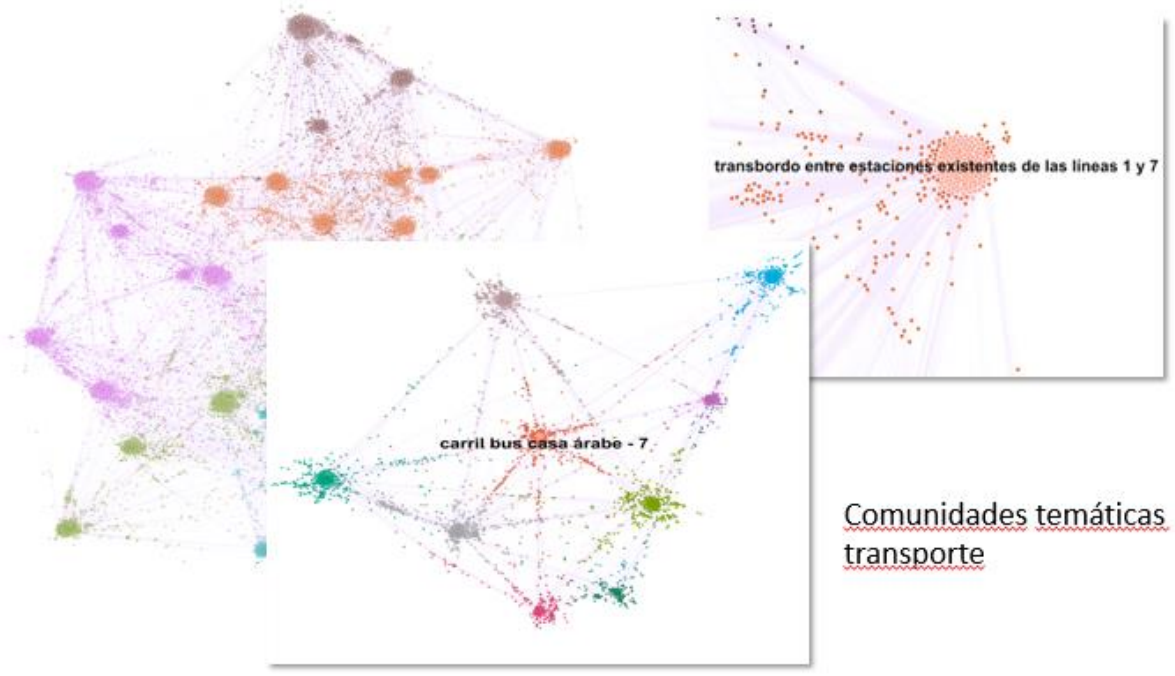

Ilustración 7: Ejemplo de comunidad y ampliación de la misma.

Tecnología: Gephi (abierto).

Recapitulando

No hay duda de que un cerebro humano entiende un texto mucho mejor que un ordenador, pero un ordenador puede manejar miles de textos en mucho menos tiempo.

Cuando el número de propuestas ciudadanas desborda la capacidad humana para leerlas y comprenderlas, podemos contar con algunas herramientas automáticas para ayudarnos, ya seamos proponentes, votantes o gestores públicos.

Las tecnologías del lenguaje son el eslabón para convertir el texto en datos sobre los que ya se podrán aplicar otras técnicas de análisis y visualización.

Estas herramientas (NLP y grafos) pueden hacer cosas como:

- Descubrir automáticamente los temas que plantea la ciudadanía.
- Tener una visión global de los temas.
- Seguir la evolución de estos temas.
- Agregar propuestas por su contenido temático.
- Encontrar propuestas similares por su contenido a una dada.
- Comparar barrios u otras particiones de la ciudadanía (género, edad, …, según los metadatos disponibles) según sus preocupaciones, así como la evolución temporal.

Y estas funcionalidades pueden ayudar a hacer fácilmente cosas como:

- Proponentes, para comprobar si ya existen propuestas similares a las suyas.
- Votantes, para encontrar propuestas interesantes para ellos.
- Gestores públicos y ciudadanos, para tener una visión global de los temas que plantea la ciudadanía.
- Encontrar grupos de propuestas similares para iniciar un proceso de negociación que las agregue en una sola propuesta y así evitar la dispersión del apoyo.

TABLA RESUMEN DE TECNOLOGÍAS

Función	Categoría	Tecnología
Extracción, transformación y carga (ETL) de las fuentes textuales.	ETL	Ad hoc / TALEND
Segmentación (*Tokenization*)	NLP: crear bolsa de palabras	IXA pipes
Lematización (*Lemmatisation*)	NLP: crear bolsa de palabras	IXA pipes
Análisis morfosintáctico (POST o *Part-of-Speech Tagging*)	NLP: crear bolsa de palabras	IXA pipes
Identificación de n-gramas	NLP: crear bolsa de palabras	IXA pipes
Identificación de la acepción empleada (*Word Sense Disambiguation*)	NLP: crear bolsa de palabras	IXA pipes
Caracterización automática de temáticas (*Topic analysis*)	NLP: topic analysis	MALLET
Cálculo de distancias temáticas	NLP: topic analysis	Desarrollo propio
Almacenamiento y base de datos	Grafos	Código Java
Representación gráfica	Grafos	Gephi
Navegación en la constelación de propuestas	Grafos	Gephi
Detección de comunidades (propuestas relacionadas)	Grafos	Gephi

EJEMPLO DE COMUNIDADES DETECTADAS

Se reproducen a continuación las agrupaciones de las primeras 20 propuestas ciudadanas. Comparten color de fondo las propuestas que se han identificado automáticamente como pertenecientes a la misma comunidad.

id	title	cached_votes_up	created_at	retired_at	geozone_id	modularity_class
0	crear una zona cercada para perros en las tablas	134	15-09-2015		0	40
1	metro nuevo en la línea 1	93	15-09-2015		0	14
2	un madrileño un Árbol.	2849	15-09-2015		0	0
3	uso artístico y de promoción del arco de la victoria	49	15-09-2015		0	64
4	limpiar las calles	6560	15-09-2015		0	47
5	ampliar ayudas para nuevas empresas a entidades ciudadanas	104	15-09-2015		0	64
6	billete Único para el transporte público	34722	15-09-2015		0	25
7	hagamos que los madrileños reciclen con confianza	317	15-09-2015		0	47
8	más policías	64	15-09-2015		0	61
9	espacios infantiles	1699	15-09-2015		0	40
10	bono transporte	195	15-09-2015		0	63
11	zona cerrada para perros	165	15-09-2015		0	40
12	peatonalizar la calle mayor, carretas y duque de alba	387	15-09-2015		0	55
13	horario nocturno transporte público	9032	15-09-2015		0	63
14	arreglar las calles	576	15-09-2015		0	73
15	limpieza parque calle Antonio Leyva	55	15-09-2015		0	75
16	unir el centro con Madrid río de forma peatonal.	906	15-09-2015		0	38
17	pasos de peatones en ciudad de los poetas	49	15-09-2015		0	73
18	pasarela peatonal las tablas-estación de Fuencarral	503	15-09-2015		0	67
19	seguridad y limpieza de Madrid	445	15-09-2015		0	47
20	cerrar el tráfico en d. centro a vehículos no residentes	451	15-09-2015		0	21

REFERENCIAS:

[i] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5): pp. 993–1022.

[ii] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte And Etienne Lefebvre, "Fast unfolding of communities in large networks". Journal of Statistical Mechanics: Theory and Experiment, 9 October 2008.