Revista Cubana de Ciencias Informáticas
Vol. 12, No. 1, Enero-Marzo, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 147-157
http://rcci.uci.cu

Tipo de artículo: Artículo original
Temática: Desarrollo de aplicaciones informáticas
Recibido: 11/12/2017 | Aceptado: 22/01/2018

# Suite de componentes para la visualización de textos

# *Suite of component of text visualization*

**Jorge Antonio Gómez Colombat [*], Mónica Rubio Rojas, Yudisleidys Creagh Castillo, Pablo Puente García**

Desarrollo de Aplicaciones, Tecnologías y Sistemas, DATYS, Santiago de Cuba, Cuba

[*]**Autor para correspondencia**: jorge.gomez@datys.cu

**Resumen**

Lograr mostrar de forma clara y precisa los resultados de complejos algoritmos de Procesamiento del Lenguaje Natural (PLN) y Minería de Texto (MT) de manera que se facilite y acelere la toma de decisiones de usuarios finales no especializados en estas áreas, constituye actualmente un gran desafío desde la perspectiva académica y la comercial. En tal dirección, en este trabajo se presenta una nueva Suite de componentes para la visualización de los resultados de algoritmos de Procesamiento del Lenguaje Natural y la Minería de Texto. Estos componentes fueron diseñados con el objetivo de facilitar su reutilización, de forma transparente y sencilla, por otras aplicaciones de software. Por último, la Suite que se propone se encuentra insertada en sistemas reales para el análisis de contenido y la gestión inteligente de información textual.

**Palabras clave:** interfaces web, minería de texto, reutilización, suite de componentes, visualización de información

***Abstract***

*Nowadays, it is a great challenge from the academic and commercial perspective showing in a clear and precise way the results of complex algorithms of Text Mining (TM) and Natural Language Processing (NLP), to facilitate and accelerate the decisions making by non-specialized end users in these areas. In this direction, this work presents a new Suite of components for the visualization of the results of Natural Language Processing algorithms and Text Mining. The designing of these components aims to facilitate their reutilization, in a transparent and simple way, by other software applications. Finally, the proposed Suite is embedded in real systems for content analysis and intelligent management of textual information.*

***Keywords:*** *text mining, visualization of information, suite of components, reutilization, web interfaces*

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 1, Enero-Marzo, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 147-157
http://rcci.uci.cu

## Introducción

La cantidad de datos a los que tenemos acceso crece día a día y la mayor parte de estos aparece en forma de textos. Como resultado de ello en los últimos años ha tenido un gran auge la minería de texto, considerada un campo interdisciplinario que incluye la recuperación de información, aprendizaje automático, estadística, reconocimiento de patrones y procesamiento de lenguaje natural. En la minería de texto el objetivo es detectar información representada en los textos, generalmente desapercibida (Nualart-Vilaplana, Pérez-Montoro y Whitelaw, 2014).

En el proceso de extracción de conocimiento asociado a la minería de texto se emplean comúnmente técnicas de visualización, que son de gran utilidad para develar relaciones de interés entre los contenidos, debido a la capacidad humana de comprender imágenes más fácilmente que resultados puramente textuales. Este proceso de empleo de técnicas de visualización en la minería de texto, se ha vinculado con el término visualización de texto (VT).

El término visualización de texto se suele utilizar para técnicas de visualización de la información que en algunos casos se centran en datos de texto en bruto y en otros casos en los resultados de algoritmos de minería de texto. Estas técnicas de visualización pueden ser de uso general o muy especializadas y dedicadas a tareas analíticas o a dominios específicos de aplicación (Kucher y Kerren, 2014).

Hoy en día se cuenta con una amplia variedad de técnicas y herramientas de visualización de información; incluso algunas de estas herramientas cuentan con ejemplos de visualizaciones aplicables a tareas bastante particulares de la minería de texto, como es el caso de D3JS (D3JS, 2016); sin embargo, casi siempre tenemos la necesidad de crear o mejorar la representación visual de un conjunto específico de datos, a partir de determinados intereses.

Otro logro importante en nuestros días es el número creciente de soluciones de minería de texto, disponibles en internet, principalmente aquellas que brindan una amplia gama de servicios como MeaningCloud (MeaningCloud, 2016), IBM Watson Developer Cloud (IBMWatson, 2016), y MonkeyLearn (MonkeyLearn, 2016), por citar solo algunos ejemplos. Sin embargo, estas soluciones carecen de componentes visuales reutilizables, que expongan de una mejor manera los resultados obtenidos y puedan ser empleados en los productos clientes que las consumen; dejando en manos de los clientes o terceros la integración de los servicios de minería de texto con herramientas de visualización.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 1, Enero-Marzo, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 147-157
http://rcci.uci.cu

Una situación similar a la descrita anteriormente la tiene nuestra empresa -compañía de desarrollo de productos de software, vinculados algunos de ellos a la minería de texto-, ya que, a pesar de contar con una plataforma propia de soluciones de MT y PLN, para la reutilización de esta en el resto de los productos de software que ofrece, no cuenta con un conjunto de componentes visuales previamente definidos, generalizados y reutilizables para la representación de los resultados de MT y PLN.

A partir de esto se ha identificado como una necesidad el desarrollo de una herramienta que aglutine los componentes de visualización de textos empleados en la empresa, teniendo en cuenta las visualizaciones más apropiadas para cada tarea de MT y PLN, además de las particularidades y necesidades de cada producto. Por lo tanto, esta herramienta también debería contar con la capacidad de ser reutilizable entre los productos e integrarse de forma nativa con los servicios de la plataforma de soluciones de MT y PLN; de tal manera que los productos que emplean dicha plataforma puedan utilizar también los componentes de visualización asociados a cada solución o puedan emplearlos para representar sus propios datos.

## Materiales y métodos

Tomando como punto de partida la existencia de una plataforma común de tareas de MT y PLN, denominada Xinetica (Xinetica, 2016), se decidió que la herramienta de visualización estuviese compuesta inicialmente, por representaciones visuales asociadas a las funcionalidades brindadas por dicha plataforma, con la incorporación paulatina de otras visualizaciones necesarias en otros productos.

Por el carácter aglutinador, de generalización y homogenización de la solución propuesta, esta se concibió como una "Suite de componentes para la visualización de texto". La Suite estaría diseñada **para productos con interfaz web** y conformada por visualizaciones propias y de terceros (de uso libre), que permitiría representar apropiadamente los resultados de las funcionalidades de Xinetica, mediante la integración con sus APIs. Estas APIs se brindan en forma de servicios web y ofrecen soluciones vinculadas al procesamiento de archivos, el procesamiento de lenguaje natural y la minería de texto:

Procesamiento de archivos**:**

- Detección del formato
- Extracción de texto y metadatos

Procesamiento de Lenguaje Natural:

- Extracción de los rasgos morfológicos de las palabras

- Etiquetado morfosintáctico de las palabras
- Reconocimiento de entidades nombradas
- Corrección ortográfica
- División de palabras en sílabas

Minería de Texto**:**

- Detección y conversión de mapas de caracteres
- Modelado de textos
- Segmentación de textos
- Detección del idioma de un texto
- Detección de plagio entre documentos
- Extracción de sumario de textos
- Agrupamiento de textos
- Clasificación de textos
- Cálculo de semejanza entre textos

Para lograr una utilización óptima de los componentes de la Suite, esta debería contar con una estructura modular basada en componentes, específicamente como plugins JQuery (JQueryPlugins, 2015) en la versión inicial de la Suite, debido principalmente al amplio uso de estos en el estado del arte y la escasa madurez del estándar WebComponents (Jiménez, 2016), hacia la que debe migrar la estructura de los componentes en futuras versiones. Este tipo de estructura (basada en módulos y componentes) permitirá una apropiada organización, documentación y distribución del código. Estas características contribuyen, además, a que la Suite pueda ser empleada de forma total, o que solo se puedan seleccionar algunos componentes específicos, evitando así la sobrecarga innecesaria.

Los componentes visuales no sólo tendrán la capacidad de integrarse con las APIs de Xinetica, sino que además se les podrá especificar un conjunto de datos propios (Figura 1). Esta segunda variante de asignación de datos permitirá que los productos que no utilicen la plataforma para una determinada tarea de MT o PLN, puedan hacer uso de la visualización correspondiente a la tarea, adaptando la estructura del resultado en caso de ser necesario.

El proceso de selección o implementación de una determinada representación visual para una tarea de MT, será el paso inicial y fundamental en la creación de cada componente. Esto se debe a que el objetivo principal es brindarle al usuario una herramienta que le permita reconocer patrones de comportamiento o información relevante en el conjunto de datos representado.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 1, Enero-Marzo, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 147-157
http://rcci.uci.cu

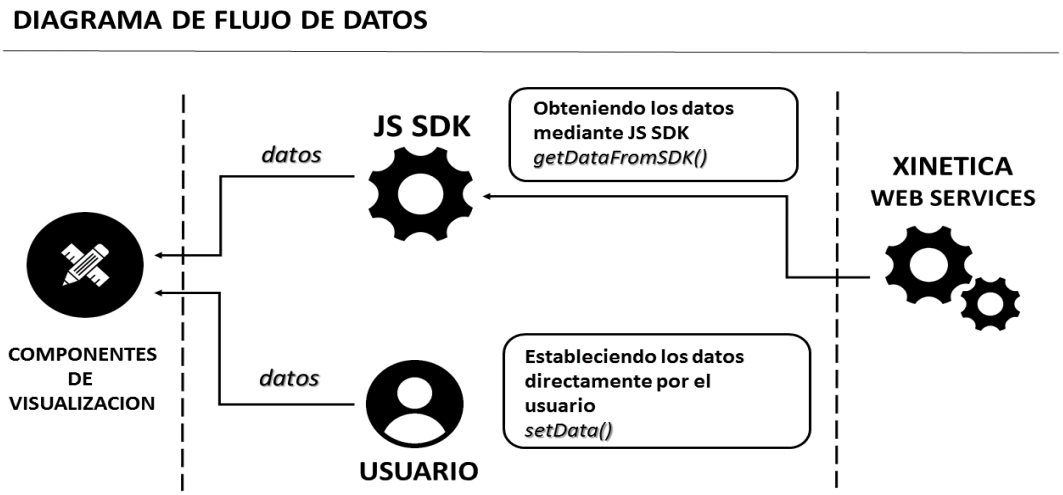


Figura 1. Diagrama de flujo de datos de la Suite

### Xinetica SDK JS

Para lograr la integración de los componentes visuales con las APIs de Xinetica, se desarrolló la herramienta Xinetica SDK JS, también como parte de la Suite. Esta herramienta tiene la capacidad de consumir los servicios de la plataforma desde JavaScript, permitiendo especificar los diferentes tipos de contenido definidos como entrada en cada servicio y la selección de estos contenidos desde diferentes orígenes, ya sea desde un formulario web, el contenido directamente o la ruta de acceso.

Xinetica SDK JS también cuenta con una arquitectura modular, garantizando una apropiada organización, distribución y documentación del código, empleando para ello herramientas como WebPack (WebPack, 2015), Jasmine (Jasmine, 2015), JSDoc (JSDoc, 2015) y la especificación CommonJS (CommonJS, 2015).

Un elemento importante del SDK lo constituye la capacidad de orquestar distintas funcionalidades de Xinetica para obtener un determinado resultado, por ejemplo, un usuario puede realizar la extracción del contenido textual de un documento en formato PDF mediante la solución Extracción de textos y obtener entonces las entidades nombradas mediante la solución de Reconocimiento de Entidades Nombradas, de la propia plataforma.

El SDK puede ser empleado por si sólo en otros productos que usen JavaScript y las APIs de Xinetica.

## Resultados y discusión

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 1, Enero-Marzo, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 147-157
http://rcci.uci.cu

Como resultado se obtuvo un conjunto de componentes visuales, agrupados y organizados en una estructura común denomina “Suite de componentes de visualización de textos”, con funcionalidades comunes para los compontes, y garantizando homogeneidad en cuanto a las funciones expuestas en cada uno ellos, además de la capacidad de integrarse con las APIs de Xinetica y representar datos especificados directamente.

La Suite tiene implementado al menos un componente de visualización asociado a cada API, en la Figura 2 se muestran algunos ejemplos.

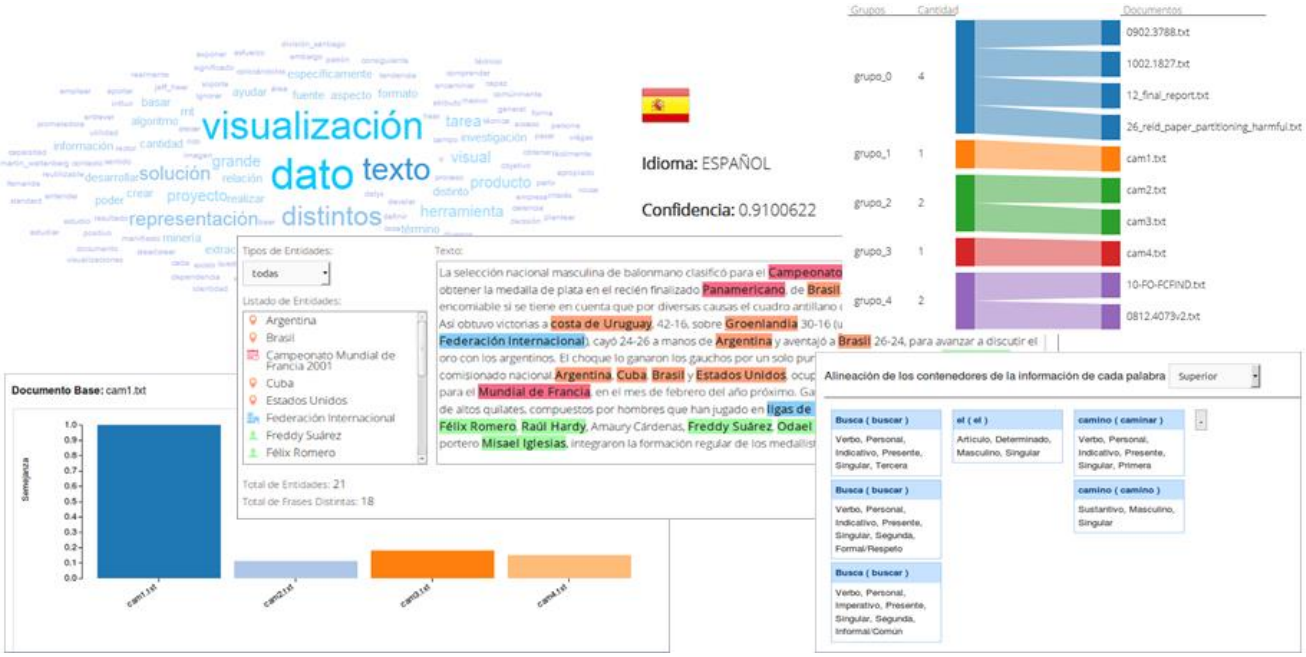


Figura 2. Algunas instantáneas de la Suite de componentes para la visualización de textos

Actualmente los componentes están siendo empleados en el portal de la plataforma común, específicamente en la sección de Xinetica Lab y en otros productos como LIDT-Noti (Sistema para la gestión de noticias).

Los resultados logrados deben permitir que cualquier nuevo componente visual –asociado a la MT- a emplearse en un producto de la empresa sea incorporado en la Suite, teniendo en cuenta sus particularidades y permitiendo así que cualquier proyecto pueda reutilizarlo de una forma más cómoda.

## Visualización de las tareas de MT y PLN

Cada visualización de la Suite tiene el objetivo de facilitar la comprensión del resultado de la tarea asociada, para ello emplea recursos gráficos y técnicas de visualización apropiadas, teniendo en cuenta el estado del arte, la estructura de los datos a representar y el objetivo que se persigue con la visualización.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 1, Enero-Marzo, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 147-157
http://rcci.uci.cu

**Visualización de Reconocimiento de Entidades Nombradas.**

La tarea de Detección de Entidades Nombradas es una de las más usadas en los sistemas de análisis de contenido y como parte de tareas más complejas como el procesamiento de noticias, análisis de tendencias de mercado, políticas, entre otras. El API de Xinetica asociada a esta tarea procesa un texto y devuelve un conjunto de entidades nombradas, agrupadas en diversas categorías.

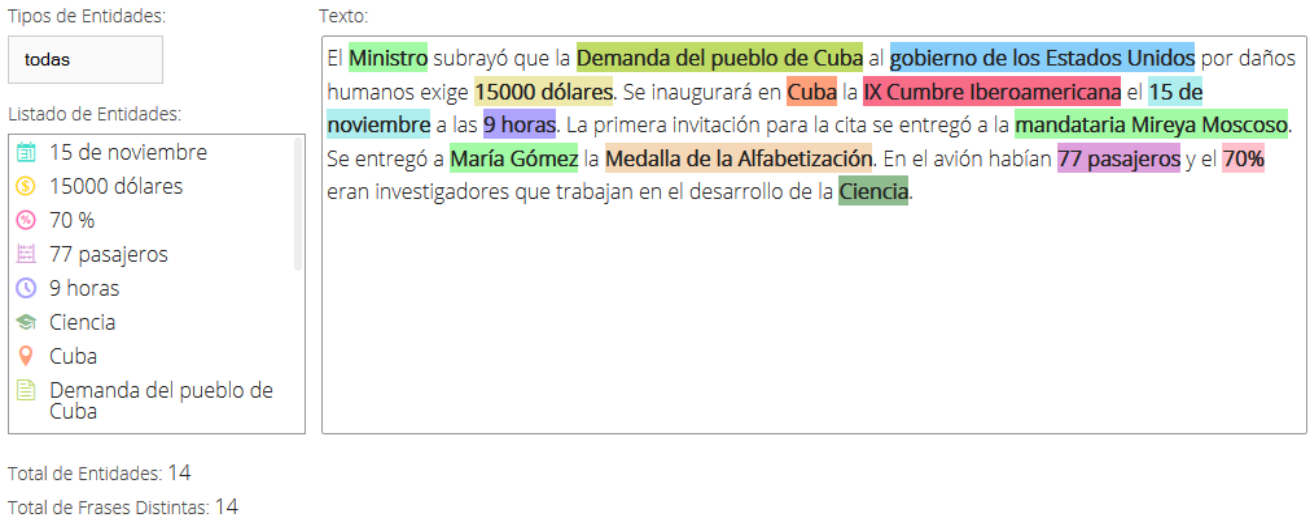


Figura 3. Visualización de la tarea de Reconocimiento de Entidades Nombradas

El componente de visualización asociado a esta tarea (Figura 3), representa las entidades en su contexto, o sea, en el texto procesado y muestra además dos listados, uno con los tipos o categorías de entidades encontradas y otro con las frases asociadas a esas entidades, además de datos estadísticos de las entidades y las frases asociadas. Para una mayor identificación de las entidades, cada tipo de entidad tiene asociado un color y un icono. Mediante el listado de las categorías se pueden filtrar los elementos presentes en el listado de frases y las frases señalizadas en el texto; y mediante la selección de elementos del listado de frases se filtran las entidades del texto.

## Integración de los componentes visuales de la Suite

Los componentes de la Suite están desarrollados como plugins JQuery lo que reduce la curva de aprendizaje al emplearlos, debido a que es una práctica bastante extendida de encapsular bibliotecas Javascript para el desarrollo de aplicaciones web.

**Integración de la visualización de Reconocimiento de Entidades Nombradas en una página web.**

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 1, Enero-Marzo, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 147-157
http://rcci.uci.cu

Para ilustrar mejor los pasos sencillos que debe seguir para incluir en su proyecto o página web la representación visual de las entidades de un texto, le mostramos los siguientes trozos de código.

1) Incluyendo hojas de estilos (CSS)

El componente visual *viewEntity* utiliza una fuente de iconos específica y un grupo de estilos asociados a los mensajes

```
<link rel="stylesheet" href="css/entity-icons.css" />
<link rel="stylesheet" href="css/viewMsg.css" />
```

2) Incluyendo las dependencias JavaScript

En este caso tenemos en cuenta la integración con Xinetica mediante Xinetica SDK JS.

```
<scriptsrc="js/xinSDK.js"></script>
<scriptsrc="js/jquery.viewEntity.js"></script>
<scriptsrc="js/jquery.viewMsg.js"></script>
```

3) Definiendo componente de entrada y contenedor de salida

Se define un componente HTML de entrada texto, en este caso, un `textarea`, para introducir el texto a procesar, y como contenedor de salida un `div` u otra que desee. En ambos casos debe especificar un *id* para su uso posterior.

```
<textarea id="text" placeholder="Escriba un texto" ></textarea>
<input type="button" id="process" value="Procesar" >
<div id="entity_results"></div>
```

4) Código JavaScript donde se instancian los objetos y se ejecutan las funciones de procesamiento y visualización

Se especifica la clave de acceso al API del módulo a utilizar, para obtener una clave debe contactar con el equipo Xinetica. Luego, se instancian los plugins JQuery *viewMsg*, relacionado con la gestión de mensajes, y *viewEntity*, referente a la visualización de las entidades. Posteriormente, se invoca la función `getDataFromSDK(id)`, a la cual se le pasa como parámetro el `id` correspondiente al `textarea`, para obtener y procesar el texto.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 1, Enero-Marzo, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 147-157
http://rcci.uci.cu

```
xinsdk.entity.token = "xxxxxxxxxxxxx";

var viewMsg = $('#entity_results').viewMsg();
var viewEntity = $('#entity_results').viewEntity({
  viewMsg: viewMsg,
  sdk: xinsdk
});

$('#process').on('click', function(){
  viewEntity.getDataFromSDK('text');
});
```

Después de obtenido el resultado automáticamente será visualizado.

En caso que cuente con su propia solución de Reconocimiento de Entidades Nombradas puede igual emplear esta visualización prescindiendo del SDK JS y usando el método setData del componente visual pasándole como parámetro el resultado de su solución, teniendo en cuenta que este resultado debe corresponderse con el formato esperado, lo cual se encuentra expresado en la documentación funcional de cada componente visual.

## Conclusiones

La visualización de textos es un campo prometedor y novedoso debido principalmente a los avances en la minería de texto, la mejor comprensión de imágenes que textos por los humanos y la cantidad de información textual disponible hoy día, la cual es explotada tanto con fines científicos, comerciales, como de seguridad.

Teniendo en cuenta esta realidad y en aras de obtener mejores resultados en nuestra compañía se desarrolló una herramienta web distribuida en plugins JQuery para visualizar resultados de soluciones de MT y PLN, con la capacidad de integrarse automáticamente a una plataforma que ofrece este tipo de soluciones (Xinetica) o visualizar datos directamente obtenidos desde otro procesamiento. También como parte de esta solución se creó Xinetica SDK JS, herramienta que permite consumir las APIs de Xinetica desde JavaScript. Ambas herramientas tienen como valor adicional ser reutilizables en productos con interfaz web vinculados a la MT.

También podemos concluir que, con la Suite como nuevo producto, la plataforma común se posiciona un paso delante de otros productos similares del mercado internacional, al contar con una solución reutilizable capaz de visualizar sus resultados, constituyendo una facilidad carente aún en muchos productos similares.

Como trabajo futuro estamos proyectando la evaluación y migración de los componentes visuales a estándares web modernos como los WebComponents y las tecnologías que los implementan. Lograr mayor parametrización de los componentes en cuanto a estilos, estructura y orientación visual. Continuar el estudio de metáforas visuales apropiadas para datos y soluciones específicas.

## Agradecimientos

Los autores desean agradecer de manera especial al Lic. Roberto Carlos Toledano Gómez, al Msc. Daniel Castro Castro, Mirlayne Campuzano Álvarez y al equipo de Xinetica por sus valiosos aportes, sugerencias y comentarios.

## Referencias

COMMONJS [En línea]. [Fecha de consulta: 26 de noviembre de 2015]. Disponible en: http://en.wikipedia.org/wiki/CommonJS

D3.JS – Data Driven Documents [En línea]. [Fecha de consulta: 5 de febrero de 2016]. Disponible en: http://d3js.org/

IBMWATSON [En línea]. [Fecha de consulta: 5 de febrero de 2016]. Disponible en: https://www.ibm.com/es-es/marketplace/cognitive-application-development

JASMINE [En línea]. [Fecha de consulta: 26 de noviembre de 2015]. Disponible en: https://github.com/jasmine/jasmine

JIMÉNEZ, J.A., et al. WebComponents: un vistazo rápido [En línea]. [Fecha de consulta: 5 de diciembre de 2016]. Disponible en: https://www.adictosaltrabajo.com/tutoriales/webcomponents-un-vistazo-rapido/

JQUERYPLUGINS [En línea]. [Fecha de consulta: 26 de noviembre de 2015]. Disponible en: https://learn.jquery.com/plugins/

JSDOC3 [En línea]. [Fecha de consulta: 26 de noviembre de 2015]. Disponible en:

http://usejsdoc.org/about-getting-started.html

KUCHER, K.; KERREN, A., et al. Text Visualization Browser: A Visual Survey of Text Visualization Techniques. [En línea] IEEE Information Visualization (InfoVis ’14), Paris, France, 2014. Disponible en: http://cs.lnu.se/isovis/pubs/docs/kucher-infovis14.pdf

MEANINGCLOUD [En línea]. [Fecha de consulta: 5 de febrero de 2016]. Disponible en: http://www.meaningcloud.com/

MONKEYLEARN [En línea]. [Fecha de consulta: 5 de febrero de 2016]. Disponible en: http://www.monkeylearn.com/

NUALART-VILAPLANA, J.; PÉREZ-MONTORO, M.; WHITELAW, M., et al. Cómo dibujamos textos. Revisión de propuestas de visualización y exploración textual. [En línea]. El profesional de la información. 2014, vol. 23, n. 3, [Consultado el: 12 de enero de 2016] 221-235 p. Disponible en: http://www.elprofesionaldelainformacion.com/contenidos/2014/may/02_esp.pdf

WEBPACK [En línea]. [Fecha de consulta: 26 de noviembre de 2015]. Disponible en: http://webpack.github.io/docs/what-is-webpack.html

XINETICA [En línea]. [Fecha de consulta: 5 de febrero de 2016]. Disponible en: http://www.datys.cu/spa/site/product/17