**CURSO TEXT MINING AND VISULIZATION**

¿Qué es minería de texto?

Una falencia de los usuarios de la estadística, es la creencia de que el dato está dado, el dato, en general, hay que conseguirlo. Las estructuras que llevan a la consecución de los datos, ha cambiado, en gran parte por las diferentes fuentes a las cuales podemos acceder para obtenerlos.

Todas las grandes bases de datos, particularmente las asociadas a redes sociales, datos abiertos Colombia, Banco mundial, etc., se entregan datos con wep-apps.

Una de las aplicaciones más simples y comunes de text mining es el predictor de autocompletado en los celulares, cuando se envían mensajes. Google, Google traslate, extracción de textos de audios, chat-bots (respuestas automáticas basadas en un perfilamiento de las necesidades o consulta de usuarios).

Grandes categorías dentro de Text Mining

|  |  |  |  |
| --- | --- | --- | --- |
| INFORMATION  EXTRACTION | INFORMATION  RETRIEVAL | NATULA LANGUAGE PROCESSING | DATA MINING |
| Son técnicas que permiten extraer información específica desde grandes cantidades de datos textuales.  Por ejemplo:  En aplicaciones médicas y en gestores de información legal. En lo médico se relaciona con la búsqueda de aspectos generales como enfermedades, tratamiento,  En lo legal, sentencias asociadas a un aspecto particular, | Es la actividad de obtener información buscando dentro de colecciones de documentos digitales  Motores de búsqueda que permite un ranking de relevancia dentro del tema buscado, posiblemente relacionado con frecuencias. | Interacción máquina – humano  Predicción de texto, etc. | Se trata de extraer patrones de datos de textos con la finalidad de describir o prever.  Análisis de sentimiento, encontrara palabras, cluster de documentos más o menos homogéneos. |

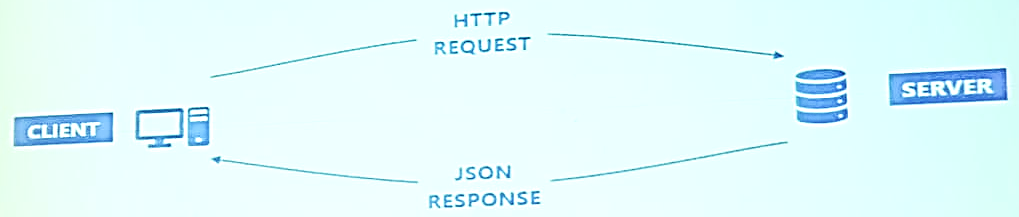
|  |  |  |  |
| --- | --- | --- | --- |
| NUEVAS FORMAS DE ACCESO | NUEVOS FORMATOS DE ARCHIVOS | BIG DATA | DATOS NO CONVENCIONALES |
| La cantidad y complejidad de los datos requieren formas de acceso lo más directas posibles, es decir, que pueden estar en nuestros programas de elaboración sin pasajes intermedios. | Las grandes cantidades de datos, la velocidad requerida para su manipulación y la multiplicidad de accesos a los mismos han generado nuevos formatos de almacenamiento muy diferentes a los clásicos. | La facilidad de almacenamiento, la masificación de experiencias, la complejidad de los fenómenos monitoreados están produciendo enormes cantidades de datos | El dato de texto es un dato cualitativo que se tiene que medir con técnicas cuantitativas. Hay que desarrollar técnicas propias para tratar este tipo de datos. |

API: Application Programing interface son un conjunto de reglas y especificaciones que las aplicaciones pueden seguir para comunicarse entre ellas. Se debe trabajar con código HTML.

WEB SCRAPING: Es un método que nos permite utilizar el código fuente de páginas web para tomar datos y organizarlos en base a datos. El registro de datos de las base digitales de periódicos, registraduría, etc.

**Características fundamentales de las Web APIs para el Data retrieval.**

Un API es una posibilidad que nos ofrece una plataforma para poder dialogar con sus servidores y obtener servicios o datos. La mayoría de plataformas de redes sociales proveen API de tipo REST (REpresentional State Transfer).



Los API tienen dos momentos: Un ***request***: requerimientos y el servidor da una respuesta. La ventaja más grande, es que el requerimiento es un HTTP, un enlace con ciertos parámetros, y luego recibir una respuesta (response) es un archivo formato JSON.

Ejemplo:

<https://www.googleapis.com/books/v1/volumes?q=isbn:0747532699>

Resultado:

{

"kind": "books#volumes",

"totalItems": 1,

"items": [

{

"kind": "books#volume",

"id": "yZ1APgAACAAJ",

"etag": "g1Upmp1MSiE",

"selfLink": "https://www.googleapis.com/books/v1/volumes/yZ1APgAACAAJ",

"volumeInfo": {

"title": "Harry Potter 1 and the Philosopher's Stone",

"authors": [

"J. K. Rowling"

],

"publisher": "Bloomsbury Pub Limited",

"publishedDate": "1997",

"description": "Harry Potter is an ordinary boy who lives in a cupboard under the stairs at his Aunt Petunia and Uncle Vernon's house, which he thinks is normal for someone like him who's parents have been killed in a 'car crash'. He is bullied by them and his fat, spoilt cousin Dudley, and lives a very unremarkable life with only the odd hiccup (like his hair growing back overnight!) to cause him much to think about. That is until an owl turns up with a letter addressed to Harry and all hell breaks loose! He is literally rescued by a world where nothing is as it seems and magic lessons are the order of the day. Read and find out how Harry discovers his true heritage at Hogwarts School of Wizardry and Witchcraft, the reason behind his parents mysterious death, who is out to kill him, and how he uncovers the most amazing secret of all time, the fabled Philosopher's Stone! All this and muggles too. Now, what are they?",

"industryIdentifiers": [

{

"type": "ISBN\_10",

"identifier": "0747532699"

},

{

"type": "ISBN\_13",

"identifier": "9780747532699"

}

],

"readingModes": {

"text": false,

"image": false

},

"pageCount": 223,

"printType": "BOOK",

"categories": [

"Juvenile Fiction"

],

"averageRating": 4.5,

"ratingsCount": 19,

"maturityRating": "NOT\_MATURE",

"allowAnonLogging": false,

"contentVersion": "preview-1.0.0",

"imageLinks": {

"smallThumbnail": "http://books.google.com/books/content?id=yZ1APgAACAAJ&printsec=frontcover&img=1&zoom=5&source=gbs\_api",

"thumbnail": "http://books.google.com/books/content?id=yZ1APgAACAAJ&printsec=frontcover&img=1&zoom=1&source=gbs\_api"

},

"language": "en",

"previewLink": "http://books.google.com.co/books?id=yZ1APgAACAAJ&dq=isbn:0747532699&hl=&cd=1&source=gbs\_api",

"infoLink": "http://books.google.com.co/books?id=yZ1APgAACAAJ&dq=isbn:0747532699&hl=&source=gbs\_api",

"canonicalVolumeLink": "https://books.google.com/books/about/Harry\_Potter\_1\_and\_the\_Philosopher\_s\_Sto.html?hl=&id=yZ1APgAACAAJ"

},

"saleInfo": {

"country": "CO",

"saleability": "NOT\_FOR\_SALE",

"isEbook": false

},

"accessInfo": {

"country": "CO",

"viewability": "NO\_PAGES",

"embeddable": false,

"publicDomain": false,

"textToSpeechPermission": "ALLOWED",

"epub": {

"isAvailable": false

},

"pdf": {

"isAvailable": false

},

"webReaderLink": "http://play.google.com/books/reader?id=yZ1APgAACAAJ&hl=&printsec=frontcover&source=gbs\_api",

"accessViewStatus": "NONE",

"quoteSharingAllowed": false

},

"searchInfo": {

"textSnippet": "Harry Potter is an ordinary boy who lives in a cupboard under the stairs at his Aunt Petunia and Uncle Vernon&#39;s house, which he thinks is normal for someone like him who&#39;s parents have been killed in a &#39;car crash&#39;."

}

}

]

}

Métodos de request.

|  |  |  |
| --- | --- | --- |
| POST | Crear / Publicar | Al fin y al cabo trabajar con API siempre termina siendo un trabajo en el cual hacemos operaciones sobre bases de datos remotas. |
| GET | Leer (es el más usual) |
| PUT | Reemplazar |
| PATH | Modificar |
| DELETE | Borrar |

ENDPOINTS

Son URL que contestan a unos requerimientos hechos por un usuario o una aplicación. En cada **endpoint** se usa un método específico (GET, POST, etc.) declinándolo a una tarea específica. Estos endpoints permiten acceder a los datos abiertos, sin intermediación, solo estableciendo los respectivos requerimientos a la misma. En más del 90% de los casos los wepapis requieren de autenticación.

|  |  |  |
| --- | --- | --- |
| **HTTP BASIC AUTHENTICATION** | **API KEYS** | **OAUTH** |
| Es el método de autenticación clásico username/contraseña. Solamente cuando es transmitido por un http se vuelve inseguro. Cuando se hace este proceso con username y passwd estos datos viajan en un archivo http. | A cada usuario nuevo se le asigna un valor único que se les pide todas las veces. Esto usualmente implica registrarse con una cuenta dentro del portal o página sobre la cual deseamos hacer un request. | Es un método de autenticación y autorización al tiempo. El usuario inicia sesión en el sistema y el sistema le pide una autenticación en forma de token. Este token está vinculado a unos servicios específicos y está normalmente limitado en el tiempo.  Autenticarse: Identificarse  Autorizarse: Que puedes hacer |

**Limitaciones del servicio.**

Algunos servicios de API son públicos, es decir, no se paga nada para su acceso. En muchos casos y más que todo cuando el flujo informativo es muy grande (muchos usuarios y/o muchos datos), los proveedores suelen limitar la cantidad de request o contener su dimensión.

|  |  |  |
| --- | --- | --- |
| **PAGINATION** | **RATES LIMITS** | **THROTTLING** |
| En algunos casos los datos restiruidos por el endpoint vienen partidos por grupos. Por ejemplo si el request produce un resultado de 1000 records, el endpoint puede obligar a hacer 10 request en cada uno con 100 records. Para hacer esto se usa un parámetro que se llama paginación el cual indica el subconjunto de records a descargar. | Son límites de request que pone un endpoint. Pueden ser límites de tiempo (descargas limitadas a un momento en el tiempo mes, semana, día un historial), límites de cantidad (no descargar más de 15000 records por request), límites de queries ( no se pueden hacer más de cierto número de request). Son también comunes rates limits que combinan límites de tiempo y de queries/cantidad. | Es la práctica de limitar la banda de la descarga cuando se presentan picos picos de requerimeintos hacia un endpoint, por ejemplo cuando varos usuarios se conectan desde el mismo IP. |

**Los APIs y R**

R puede dialogar con Web-APIs de datos y servicios por medio de algunos paquetes. La ventaja de esta operación es que podemos formular request desde R e importar datos directamente en R, sin pasar por archivos externos. Esto nos permite ahorro de tiempo y de espacio. Hay dos formas con las cuales podemos acceder a Web APIs desde R con y sin paquetes client.

|  |  |
| --- | --- |
|  |  |
|  |  |

Fases de Gestión de un API por medio de R.

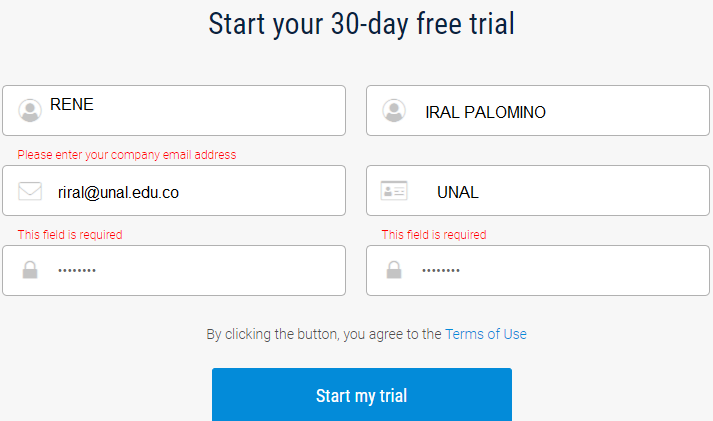
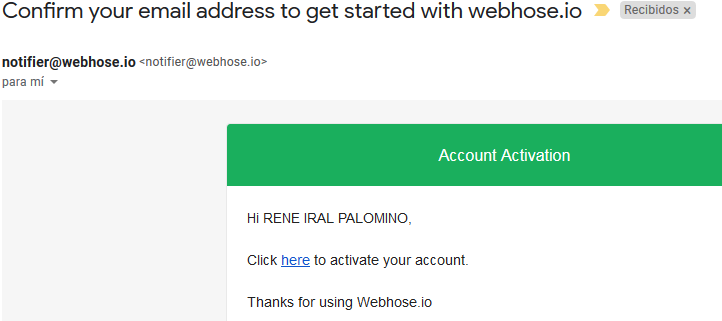
|  |  |  |
| --- | --- | --- |
| CREDENCIALES | REQUEST | RESPONSE |
| Se crean credenciales de acceso, normalmente en la página web del servicio de API que se va a usar. Esta fase se gestiona parcialmente en R, pero depende del tipo de autenticación requerida por el API. | Se formula un request por medio de uan serie de parámetros (método, endpoint, path, query, autentication). Esta fase se gestiona por completo con R o bien con un wrapper o por medio de httr. | Si el request es exitoso recibiremos una respuesta en formato json, que se debe organizar en un formato que se pueda analizar fácilmente. Este proceso se llama parsing y se hace por medio del paquete **jsonlite**. |

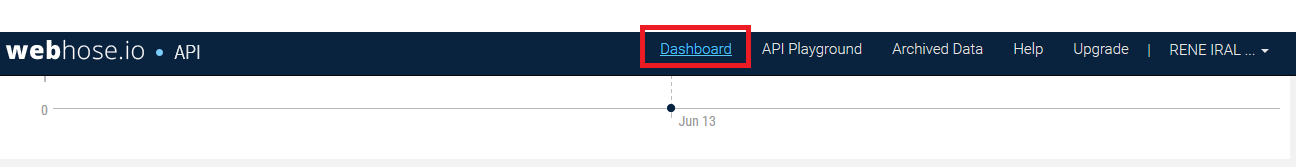
Práctica.

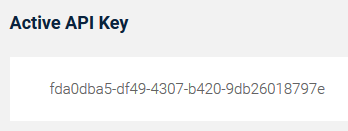
<https://webhose.io/>

login: [riral@unal.edu.co](mailto:riral@unal.edu.co)

passwd: renato12



 Códigos para la búsqueda, UTF8, no tiene en cuenta tildes

Código LATIN1, español Latinoamérica con acentos y tildes.

Status 200 dice que esta bien



EJERCICIO 1. Se realizará un request sobre el tema Digital Colombia

Ver archivo de códigos en R: “Text mining Ejercicio 1.txt”

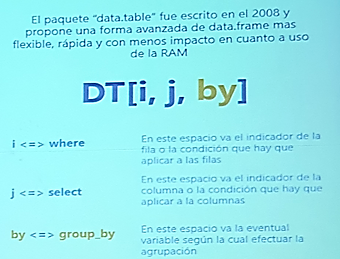
Formas de lidiar con BIGDATA usando R.

El formato data.frame y las funciones básicas no son aptas para poder trabajar con bases de datos grandes. Existen dos tipos de soluciones.

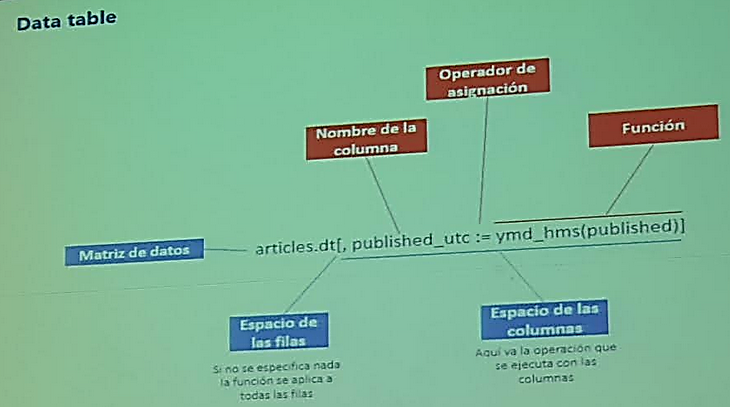
|  |  |
| --- | --- |
| **SOLUCIONES INTERNAS** | **SOLUCIONES EXTERNAS** |
| Son librerías que permiten trabajar con bases de datos medianamente grandes directamente desde nuestra computadora. Estos son: **data**.**table** y el conjunto de librerías **tidyverse.** | Son soluciones en la nube como por ejemplo Servidores Cloud o servicios distribuidos como hadoop/Spark. Permiten trabajar con bases de datos enormes. Son servicios a los cuales se puede acceder desde R por medio de paquetes que dialogan con estos servicios Web |

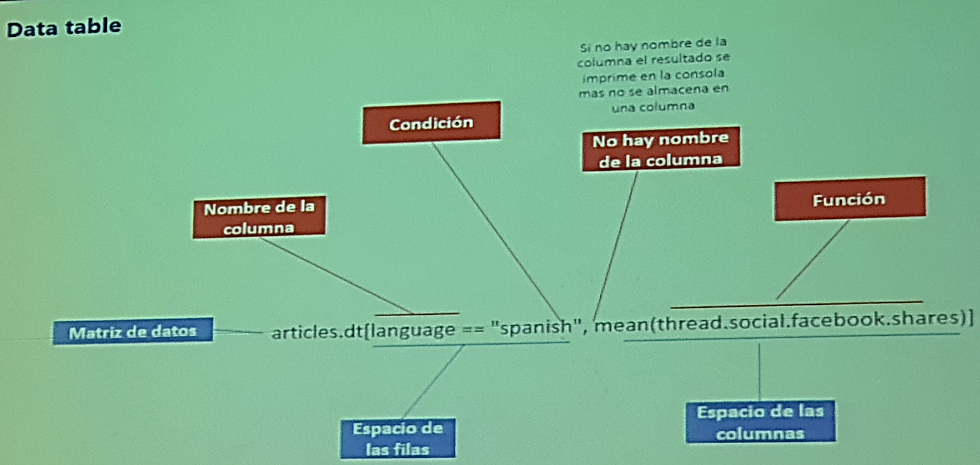
|  |  |
| --- | --- |
| **DATA.TABLE** | **TIDYVERSE** |
| La librería data.frame intenta solucionar el problema de los Bigdata por medio de un formato de datos nuevos, más moldeable y flexible que los data.frames. Inspirado en las query SQL. | Es un conjunto de librerías con una misma filosofía. A pesar de que promueven también un formato nuevo de datos (tibble), se basan en trabajar con Big Data por medio de funciones ágiles y de fácil comprensión. Otra característica es que utilizan operadores pipe( %>%) |

Formato del data.table



**Data.table**





Ejercicio 2: Explorar el archivo descargado por medio de data.table

Responder:

* En qué idioma están los artículos
* De qué países son
* ¿Cuántos artículos de prensa, de blog y publicaciones de foros hay?

Los operadores que permiten búsqueda más inteligente:

AND OR NOT name$ (solo busca el nombre$)

Servicios de APIs gratuitos

* webhose ([https://webhose.io](https://webhose.io/))
* News API (<https://newsapi.org/>)