# Prácticas laborales 2023-2023

Responsable: Lic. Carlos León González
Email: krlleonglez@gmail.com
Telegram: @krl21

---

**Tema #1:** Limpieza de los textos de Wikipedia, a través de expresiones regulares.

**Estudiantes:** Brian Ameht Inclan Quesada, Dariel Martinez Perez

**Problema:** La información consultada por Wikipedia, puede ser consumida de varias maneras; una de estas es a través de ficheros. La información visual que consumen los usuarios, están modeladas a partir de estructuras llamadas “plantilla”. Por ejemplo, la siguiente sección de texto es el formato que aparece en los ficheros, y posteriormente se muestra como los usuarios lo ven.

```
Ochenta y cinco

::''Este artículo trata sobre el número 85, para el año véase [[85|año 85]]''
{{Ficha de número
|número      = 85
|cardinal    = Ochenta y cinco
|ordinal     = Octogésimo quinto, -a
|factorización = 5 × 17
<!-- Sistemas de numeración -->
|num_romana  = LXXXV
|num_ática   = [[Archivo:Grec 50.png|12px]]ΔΔΔΠ
|num_jónica  = πε
|num_china   = 八十五
|num_china_trad = 捌 拾 伍
|num_egipcia = ∩∩∩∩∩∩∩∩IIIII
|num_armenia = ՁԵ
|num_maya    = [[Archivo:Maya 4.svg|20px]] <br> [[Archivo:Maya 5.svg|20px]]
|num_cirílica= ΠЄ
|num_india   = அய்ந
|binario     = 1010101
|octal       = 125
|hexadecimal = 55
<!-- Funciones -->
|fun_euler   = 64
|fun_divisor =
|fun_mobius  = 1
|fun_mertens = -3
<!-- Navegación -->
|anterior    = [[ochenta y cuatro]]
|siguiente   = [[ochenta y seis]]
}}

El '''ochenta y cinco''' (85) es el [[número natural]] que sigue al [[ochenta y cuatro]] y precede al [[ochenta y seis]].

== Propiedades matemáticas ==
* Es un [[número compuesto]], que tiene los siguientes factores propios: 1, 5 y 17. Como la suma de sus factores es 23 < 85, se trata de un [[número deficiente]].

* Es el número más pequeño que se pueda expresar como suma de dos cuadrados de dos maneras diferentes: 85 = 2² + 9² = 6² + 7².
```

# Ochenta y cinco

*Este artículo trata sobre el número 85. Para el año, véase año 85.*

El **ochenta y cinco** (85) es el número natural que sigue al ochenta y cuatro y precede al ochenta y seis.

## Propiedades matemáticas [ editar ]

- Es un número compuesto, que tiene los siguientes factores propios: 1, 5 y 17. Como la suma de sus factores es 23 < 85, se trata de un número deficiente.
- Es el número más pequeño que se pueda expresar como suma de dos cuadrados de dos maneras diferentes: 85 = 2² + 9² = 6² + 7².

## Características [ editar ]

- 85 es el número atómico del ástato.

| 85 | |
|---|---|
| Cardinal | Ochenta y cinco |
| Ordinal | Octogésimo quinto, -a |
| Factorización | 5 × 17 |
| Sistemas de numeración | |
| Romana | LXXXV |
| Ática | 𐅁ΔΔΔΠ |
| Jónica | πε |
| China | 八十五 |
| China financiera | 捌 拾 伍 |
| Egipcia | ∩∩∩∩∩∩∩∩IIIII |
| Armenia | ՁԵ |
| Maya | ····<br>— |
| Cirílica | ΠЄ |
| India | அய்ந |
| Sistema binario | 1010101 |
| Sistema octal | 125 |
| Sistema hexadecimal | 55 |
| Como parámetro de una función | |
| Función φ de Euler | 64 |
| Función de Möbius | 1 |
| Función de Mertens | -3 |
| Lista de números | |

[editar datos en Wikidata]

Control de autoridades **Proyectos Wikimedia** · Datos: Q427730 · Multimedia: 85 (number) / Q427730

Categoría: Números enteros

Luego, para aplicar análisis sobre los textos de Wikipedia, se hace necesario un proceso de limpieza de los textos, para eliminar las estructuras insertadas por Wikipedia en los ficheros, que puedan ofrecer ruidos a los resultados finales.

**Requerimiento:** Se necesita una función, implementada en Python, que reciba un texto con las plantillas embebidas de Wikipedia y retorne el texto correspondiente. Para esto se recomienda que se utilice expresiones regulares. Además, se recomienda tener un parámetro adicional que indique la parte a obtener del texto: resumen inicial; cuerpo; tabla de información (aparece a la derecha); u alguna combinación de las 3 posibilidades anteriores.

Por ejemplo, para el fichero 1000070.txt, ubicado en la colección de documentos, se espera que la función devuelva las siguientes posibles salidas:

Resumen inicial:
Bloque fue un grupo musical de rock urbano, rock progresivo y rock sinfónico nacido en Torrelavega, Cantabria (España), en el año 1973.

Tabla de información:
Bloque
Origen: Torrelavega, Cantabria (España)
Estado: Disueltos
Género(s): Rock progresivo
Período de actividad: 1973 - 1983
Artistas relacionados: Asfalto, Triana, Leño, Coz, Ñu, Smash, Storm
Miembros: Luis Pastor, Juan José Respuesta, Sixto Ruiz, paco Baños, Juan Carlos Gutiérrez

Cuerpo:
Historia
Bloque se formó en Torrelavega y Santander (Cantabria) en 1973, cuando Luis Pastor, bajo, Sixto Ruiz, guitarra, Ito Luna, batería y Juan Carlos Gutiérrez, voz, inspirándose en grupos como The Allman Brothers Band, Yes o King Crimson, decidieron dar rienda suelta a su imaginación musical y crear esta banda. Con la llegada del guitarrista Juanjo Respuela en 1976, encontraron el sonido que les caracterizaría, enérgico y envolvente, en el que destacan los solos a dos guitarras de Respuela y Ruiz, al modo de The Allman Brothers Band.

A esa formación original de Bloque, se añadieron en los años siguientes otros músicos: Manolo Quinzaños, teclista, con el cual actuaron en el Festival de Burgos en 1975, y al cual sustituyó Mario Gómez Calderón a los teclados, que junto a la incorporación del guitarrista Dioni Sobrado fueron la formación que actuó en el Festival de León de 1976. Ito Luna, Dioni Sobrado y Mario Gómez abandonan el grupo ese mismo año, creando con la incorporación de Lili Alegría (bajo y voz) el grupo Ibio, también de rock sinfónico. Tras esa escisión de 1976, se incorporan Juan José Respuela, guitarra y Paco Baños a la batería, pasando Juan Carlos Gutiérrez a los teclados y voz, formación con la que graban su primer disco en 1978. Al año siguiente, 1979, grabaron su segundo disco, con Carlos Terán a la batería tras la marcha de Paco Baños. Tras la grabación de su segundo disco, sale Carlos Terán y entra Tivo M.Salmón a la batería.

Sus primeros conciertos importantes tuvieron lugar en los festivales de León y Burgos, así como en Cataluña, en el "Nadal Rock". Más tarde actuaron en el famoso local de rock "M&M", en Madrid, y en el programa de televisión de TVE Voces a 45.

A raíz de esta serie de eventos, Chapa Discos, un sello perteneciente a la discográfica Zafiro, les contrata y editan su primer LP en 1978, titulado como el grupo y producido por Vicente Romero y Luis Soler. Del disco, grabado en solo cinco días, se extraen dos singles, "La Libre Creación/ Nostalgia" y "Undécimo Poder/Abelardo y Eloísa".

Al año siguiente, 1979, editan Hombre, tierra y alma, que produjeron ellos mismos.

En 1980 sacan El hijo del alba, grabado con mucha más calidad, pero que no tuvo el éxito esperado, debido, sobre todo, al cambio en los gustos musicales que se estaba viviendo con la aparición de la movida madrileña. Se extraen de él dos singles, "El hijo del alba/La razón natural" y "Quimérica laxitud/Danza del Agua" (fragmento).

Por último, editaron un último disco, al que titularon Música para la libertad (1981). De él se publicaron dos sencillos, "Detenidos en la Materia/Mágico y salvaje" y "Solo sentimiento/ Detenidos en la materia".

La banda se disuelve en el año 1983.

En 1993 dos de los miembros originales, Juan José Respuela y Juan Carlos Gutiérrez, consiguen unir a la banda y hacer una serie de conciertos, acompañados por varios colaboradores.

En 1999 se edita el disco En directo, una grabación realizada en 1994 en la sala Revólver de Madrid, con la participación de Iván Velasco (guitarras), Luis Escalada (batería), Pepe Masides (bajo) y Marcos Gómez (teclados).

En 2008, reaparecen en el "Festival del Lago" en la localidad gaditana de Bornos junto a Imán Califato Independiente y Gwendal, consiguiendo un éxito rotundo ante un público entregado que abarrotaba el patio del Convento Corpus Christi. Tras esta actuación se reúnen para celebrar un concierto extraordinario en la plaza porticada de Santander junto a grupos como Danza Invisible, Soil & Pimp Sessions y Achtung Babies.

Miembros
Luis Pastor - Bajo.
Juan José Respuela - Guitarra acústica y eléctrica, voz.
Sixto Ruiz - Guitarra acústica y eléctrica, voz.
Paco Baños - Batería.
Juan Carlos Gutiérrez - Voz y teclados.

Discografía
Álbumes
Bloque - (1978)
Hombre, tierra y alma - (1979)
El hijo del alba - (1980)
Música para la libertad - (1981)
En directo - (1999)

**Nota:** Lo anterior expuesto es la tarea definida para aprobar la asignatura. En cuanto se termine y el tiempo lo permita, se añadirán otros requerimientos que enriquecerán los resultados y el conocimiento adquirido por parte de los estudiantes.

**Tema #2:** Confección de una estructura jerárquica, a partir de las categorías predefinidas en los artículos de Wikipedia.

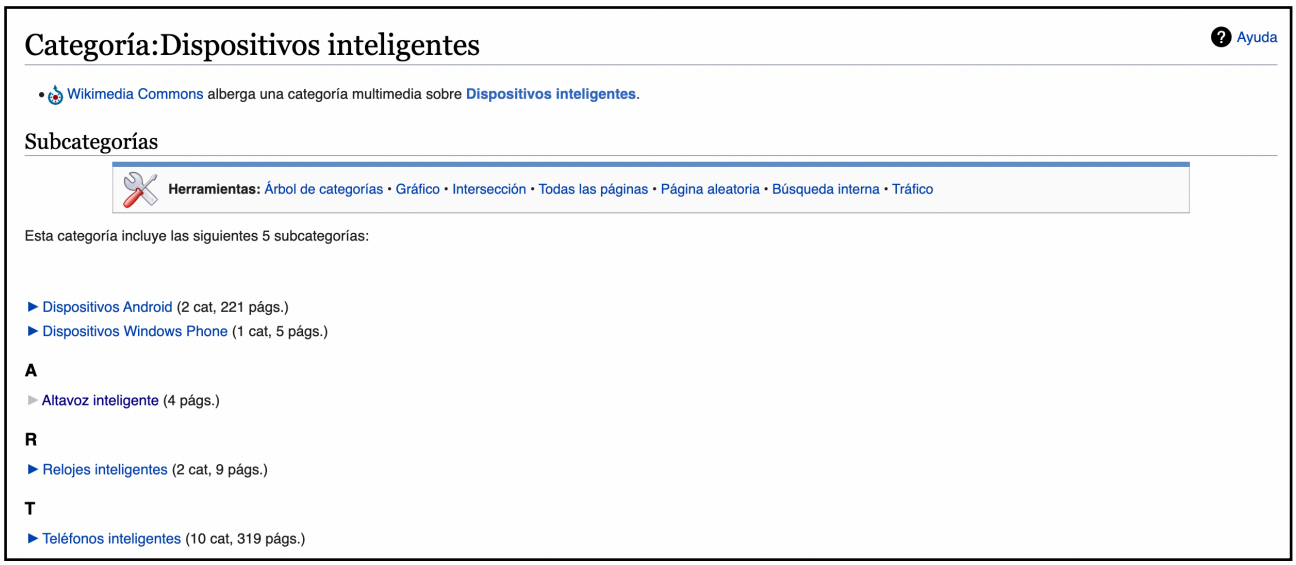
**Estudiante:** Jan Carlos Pérez

**Problema:** La información consultada por Wikipedia, puede ser consumida de varias maneras; una de estas a través de ficheros. Los artículos dentro de la enciclopedia, están marcados por categorías; por ejemplo, el artículo referente a *Altavoz*, pertenece a las categorías Altavoces, Inventos de Alemania del siglo XIX, Periféricos de computadora, Ciencia de 1861, Alemania en 1861 e Introducciones audiovisuales de 1924.

La asignación de categorías permite agrupar artículos, y con ellos, su información. Las categorías se definen al final de cada artículo. Además, Wikipedia brinda páginas donde se muestran las subcategorías que puede contener una categoría dada. Por ejemplo, la imagen siguiente muestra una categoría y sus subcategorías definidas.

**Requerimiento:** Se necesita 2 funciones, implementadas en Python. La primera función, a partir de una categoría, tiene que devolver todas las subcategorías contenidas en esta; y la segunda función debe de retornar todos los textos que pertenecen a una categoría definida.

Se recomienda realizar un análisis previo de los ficheros en los que se definen las categorías, para establecer cierta estructura arbórea que permita la búsqueda de las categorías de manera rápida y además, la implementación de alguna estructura de datos que establezca la relación entre una categoría y los ficheros que pertenezcan a ella, así como la relación entre las categorías.




**Nota:** Lo anterior expuesto es la tarea definida para aprobar la asignatura. En cuanto se termine y el tiempo lo permita, se añadirán otros requerimientos que enriquecerán los resultados y el conocimiento adquirido por parte de los estudiantes.

**Tema #3:** Visualización de distintos vectores de *word embedding*.

**Estudiante:** David Sánchez Iglesias

**Problema:** Con los avances tecnológicos, se han creado algoritmos que tomando un texto, le asignan a los token del mismo vectores numéricos, cuyos vectores capturan información semántica de los token en el texto. Han surgido varios algoritmos, pudiendo sitar Word2Vec, GloVe y Bert, cada cual con sus particularidades en la forma de construir los vectores en el espacio $\mathbb{R}^n$.

**Requerimiento:** Se necesita la implementación de 2 funciones, en Python. La primera, utilizando un corpus definido por parámetro, debe de construir los vectores resultantes de aplicar varios métodos de word embedding; como métodos imprescindible se encuentran: Word2Vec, GloVe, Bert, FastText y Poincaré, y de manera opcional: WordRank y Varembed. Luego del procesamiento de los textos, cada método tiene que trabajar con los mismos datos.

La segunda función consiste en a partir de un fichero con vectores, visualizar las palabras usando TensorFlow y el algoritmo TSNE como reductor de dimensiones.

Se recomienda el uso de las bibliotecas spaCy y gensim.

**Nota:** Lo anterior expuesto es la tarea definida para aprobar la asignatura. En cuanto se termine y el tiempo lo permita, se añadirán otros requerimientos que enriquecerán los resultados y el conocimiento adquirido por parte de los estudiantes.