

PEC 3 - Desarrollo del trabajo - Fase 2

Endo-Mining: herramienta web para la búsqueda automatizada de genes potencialmente relacionados con la endometriosis a través de minería de textos

Jorge Vallejo Ortega

13/05/2021

Índice

1	Descripción del avance del proyecto	2
1.1	Grado de cumplimiento de los objetivos y resultados previstos en el plan de trabajo	2
1.1.1	Objetivos generales	2
1.1.2	Objetivos específicos	2
2	Relación de las actividades realizadas	3
2.1	Actividades previstas en el plan de trabajo	3
2.1.1	Tarea 1. Definir esquema de la aplicación web	3
2.1.2	Tarea 2. Página de visualización de abstracts	8
2.2	Actividades no previstas y realizadas	8
3	Relación de las desviaciones en la temporización y acciones de mitigación si procede y actualización del cronograma si procede	9
3.1	Desviaciones	9
3.2	Acciones de mitigación	9
3.3	Actualización del cronograma	9
4	Listado de los resultados parciales obtenidos hasta el momento (entregables que se adjuntan)	10
5	Apéndices	11
5.1	Apéndice A: Código	11
5.2	Apéndice B: Reproducibilidad	11
6	Referencias	12

1 Descripción del avance del proyecto

1.1 Grado de cumplimiento de los objetivos y resultados previstos en el plan de trabajo

Los objetivos planteados generales y específicos son:

1.1.1 Objetivos generales

1. Encontrar genes relacionados con la endometriosis aplicando técnicas de minería de textos.

1.1.2 Objetivos específicos

1. Desarrollar un script que permita realizar un procedimiento de minería de textos automáticamente, desde la recopilación de datos en bruto hasta la presentación de resultados.
2. Desarrollar una aplicación web implementando el script de minería de textos que resultó del objetivo anterior.

2 Relación de las actividades realizadas

2.1 Actividades previstas en el plan de trabajo

2.1.1 Tarea 1. Definir esquema de la aplicación web

El objetivo de esta tarea era tener una idea clara de la estructura que tendría la aplicación una vez finalizada; los *inputs* necesarios, los *outputs*, los diferentes controles al alcance del usuario y una idea aproximada del aspecto final. A continuación incluyo los diseños, realizados con LibreOffice Draw, de las diferentes secciones que planeé para la aplicación web:

ENDO-MINING

Búsqueda en PubMed

Frecuencia palabras

Frecuencia genes

Gráficas de frecuencia

Caracterización de genes

Publicaciones por término

Publicaciones por gen

Acerca de

Búsqueda en PubMed

Palabras clave
endometriosis

Rango de fechas
21-04-2011 hasta 21-04-2021 ☒ Activa rango de fechas

Consultar
endometriosis AND 2011/04/21:2021/04/21[dp]

Buscar

Número de citas recuperadas: 11729

PMID	Publicaciones
33882252.00	1. Br J Radiol. 2021 Apr 21;20201441. doi: 10.1259/bjr.20201441. [Epub ahead of
33880938.00	2. J Comp Eff Res. 2021 Apr 21. doi: 10.2217/ce-2020-0243. [Epub ahead of print]
33879147.00	3. BMC Womens Health. 2021 Apr 20;21(1):167. doi: 10.1186/s12905-021-01318-0.

Figura 1: Pantalla inicial. Incluye los campos para entrada de palabras clave de búsqueda y rango de fechas. Muestra la cantidad de citas encontradas y una tabla con los datos de publicación de todas ellas.

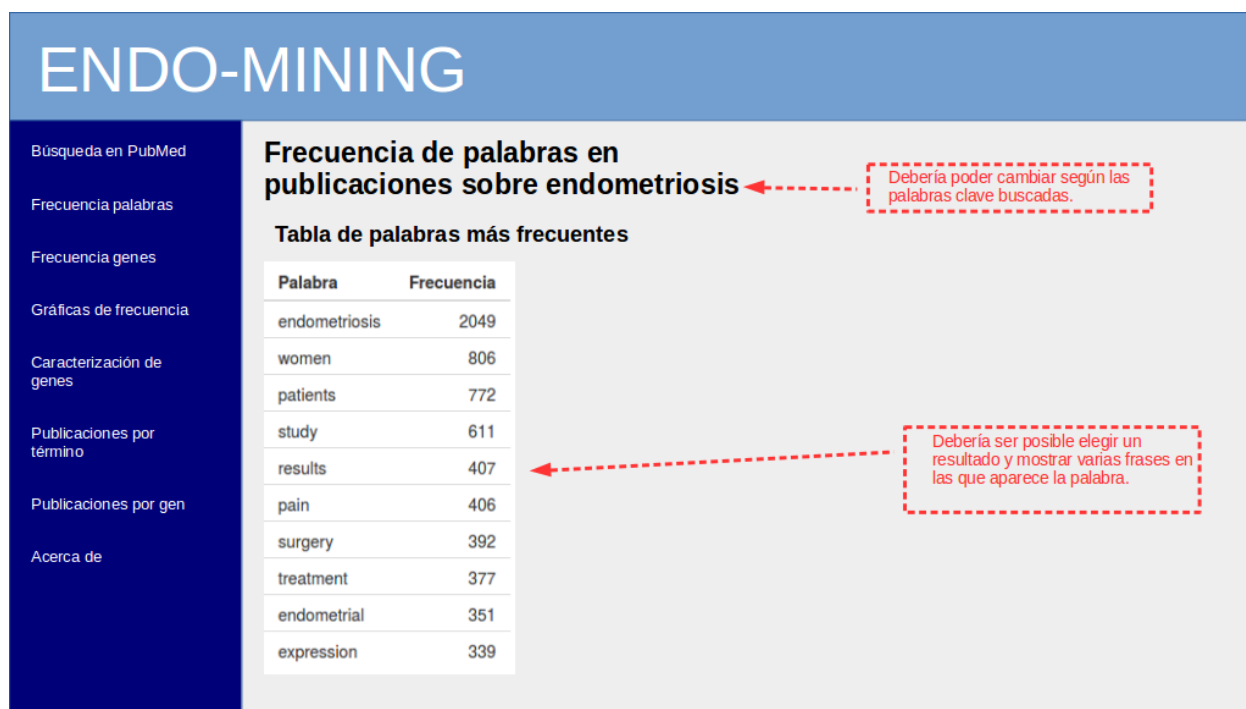


Figura 2: Frecuencia de palabras. Muestra una tabla con la frecuencia de las palabras que componen el corpus primario.

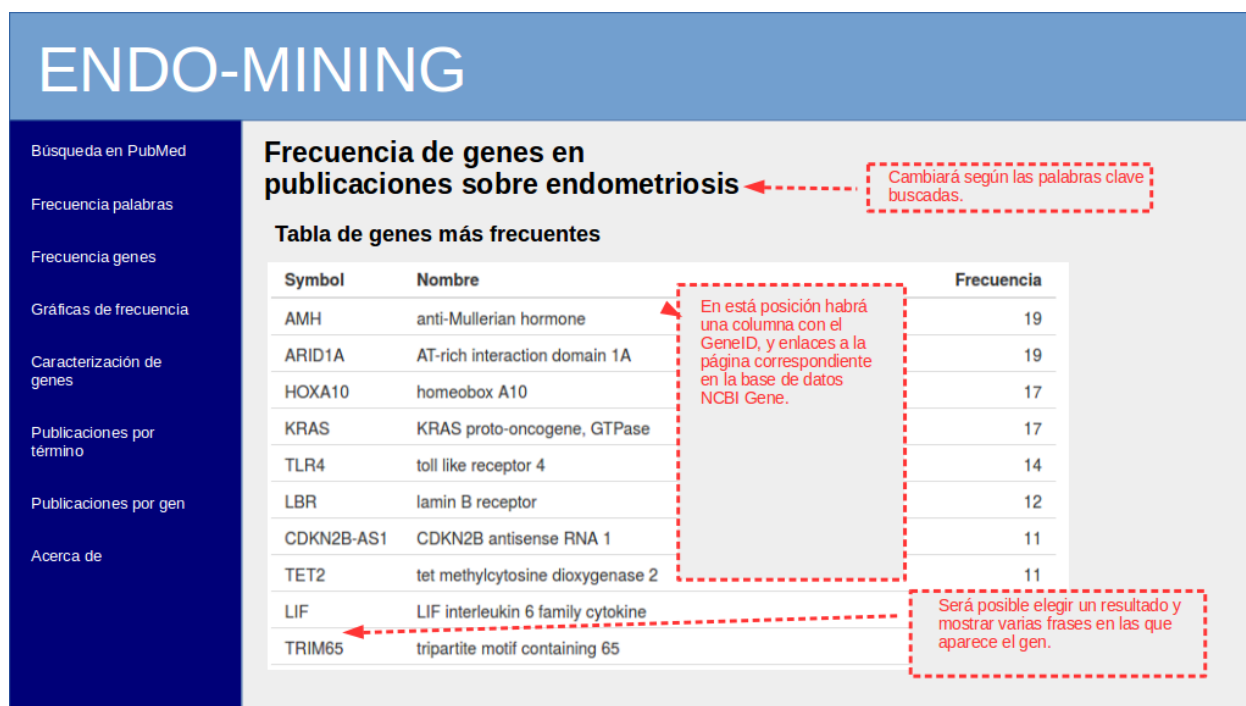


Figura 3: Frecuencia de genes. Muestra una tabla con los genes recuperados del corpus primario y su frecuencia, ordenados de mayor a menor.

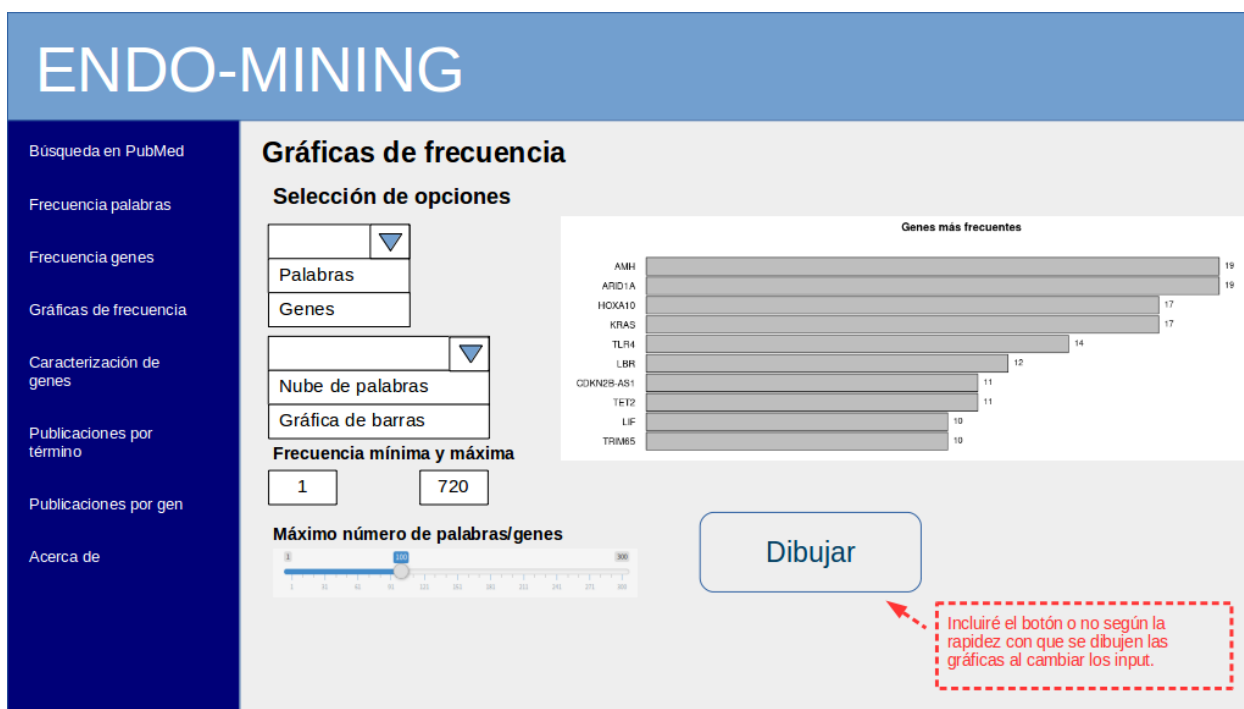


Figura 4: Gráficas de frecuencia de palabras y genes en el corpus primario. Listas desplegables permitirán elegir el tipo de gráfica (nube de palabras o gráfica de barras) y la entidad mostrada (palabras o genes). Se podrá elegir la frecuencia máxima y mínima de las entidades representadas, y cuántas entidades aparecerán en las gráficas.

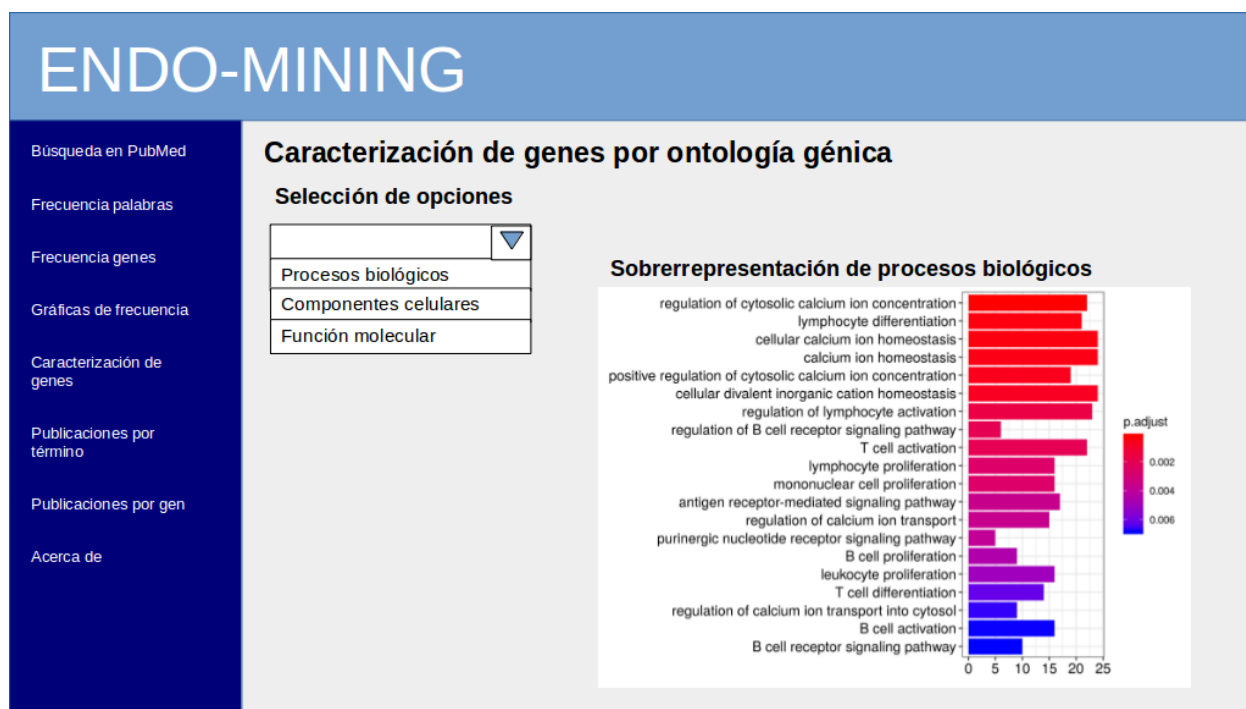


Figura 5: Caracterización por ontología génica. Muestra en una gráfica de barras los datos de ontología génica sobrerrepresentados en la muestra de genes. Con una lista desplegable se podrá elegir representar los datos correspondientes a procesos biológicos, componentes celulares o función molecular.



ENDO-MINING

- Búsqueda en PubMed
- Frecuencia palabras
- Frecuencia genes
- Gráficas de frecuencia
- Caracterización de genes
- Publicaciones por término
- Publicaciones por gen
- Acerca de

Filtrar publicaciones por gen

Seleccione un gen del desplegable

▼

Filtrar

Número de citas con el gen: 11

PMID	Publicaciones
33882252.00	1. Br J Radiol. 2021 Apr 21;20201441. doi: 10.1259/bjr.20201441. [Epub ahead of
33880938.00	2. J Comp Eff Res. 2021 Apr 21. doi: 10.2217/ce-2020-0243. [Epub ahead of print]
33879147.00	3. BMC Womens Health. 2021 Apr 20;21(1):167. doi: 10.1186/s12905-021-01318-0.
33877643.00	4. Reprod Sci. 2021 Apr 20. doi: 10.1007/s43032-021-00587-2. [Epub ahead of print]
33876904.00	5. Minerva Obstet Gynecol. 2021 Apr 20. doi: 10.23736/S2724-606X.21.04764-X. [Epub
33876611.00	6. Tidsskr Nor Laegeforen. 2021 Apr 19;141(6). doi: 10.4045/tidsskr.20.0551. Print

Tabla con las citas que incluyen el gen seleccionado.

Al seleccionar un resultado se mostrará el resumen y un enlace a la página del artículo

Figura 7: Filtrar publicaciones por gen. El usuario podrá elegir un gen de una lista desplegable, de entre la lista de genes extraídos del corpus primario. El resultado será una selección de las citas del corpus que contienen el gen seleccionado.

ENDO-MINING


- Búsqueda en PubMed
- Frecuencia palabras
- Frecuencia genes
- Gráficas de frecuencia
- Caracterización de genes
- Publicaciones por término
- Publicaciones por gen
- Acerca de

Acerca de

Versión de la aplicación

Enlace a la memoria completa del trabajo de fin de máster

Datos de contacto



Universitat
Oberta
de Catalunya

Figura 8: Esta sección de la aplicación mostrará al usuario información acerca de la propia aplicación, el trabajo de fin de máster en el que tiene su origen y los datos de contacto del creador.

A medida que fui implementando funcionalidades, algunas opciones acabaron siendo descartadas, y se incluyeron otras al juzgar que el usuario las encontraría útiles. XXXX Completar y modificar este párrafo con ejemplos concretos sacados de la aplicación real (ej. los controles en la sección de ontología génica son más completos de previsto) XXX

2.1.2 Tarea 2. Página de visualización de abstracts

El objetivo de esta tarea es permitir que el usuario, desde la aplicación web, pueda leer los sumarios de las citas recuperadas. Además se ofrece la posibilidad de navegar hasta la página origen de la cita en PubMed, a través de un hipervínculo.

En la pantalla inicial de la aplicación los resultados de la búsqueda se muestran como una tabla interactiva. Al seleccionar el usuario cualquiera de los resultados de la tabla, se le muestra la información contenida en el respectivo resumen maqueta de forma que la información de la publicación (título, año, etc.) aparece en letra negrita, al igual que el título del artículo, y el resto del texto del resumen en letra cursiva. También se crea de forma reactiva un hipervínculo a PubMed, donde puede consultarse la cita y, en su caso, el texto completo del artículo.

Endo-Mining

Buscar en PubMed

Frecuencia de palabras

Frecuencia de genes

Palabras clave

endometriosis

Rango de fechas

06-04-2021

hasta

06-05-2021

☐ Seleccionar máximo rango de fechas

Texto consulta a PubMed

endometriosis AND 2021/04/06:2021/05/06[dp]

Buscar en PubMed

Nº de citas recuperadas: 149

Mostrar 10 registros

Buscar:

	PMID	Publicaciones
1	33950556	1. Br J Gen Pract. 2021 May 4; pii: BJGP.2021.0030. doi: 10.3399/BJGP.2021.0030.
2	33949053	2. J Obstet Gynaecol Res. 2021 May 4. doi: 10.1111/jog.14801. [Epub ahead of print]
3	33949042	3. J Obstet Gynaecol Res. 2021 May 4. doi: 10.1111/jog.14819. [Epub ahead of print]
4	33948974	4. Med Res Rev. 2021 May 5. doi: 10.1002/med.21802. [Epub ahead of print]
5	33948927	5. Reprod Sci. 2021 May 4. doi: 10.1007/s43032-021-00595-z. [Epub ahead of print]
6	33948387	6. Am J Cancer Res. 2021 Apr 15;11(4):1754-1769. eCollection 2021.
7	33947672	
8	33946650	
9	33945889	9. J Gynecol Obstet Hum Reprod. 2021 May 1:102158. doi: 10.1016/j.jogoh.2021.102158. [Epub ahead of print] Surgical management of deep pelvic endometriosis in France: do we need to be a pelvic surgeon to deal with DPE?
10	33945001	10. Steroids. 2021 May 1:108856. doi: 10.1016/j.steroids.2021.108856. [Epub ahead of

Mostrando de 1 a 10 de 149 entradas

Anterior

1

2

3

4

5

...

15

Siguiente

Abrir publicación en PubMed

9. J Gynecol Obstet Hum Reprod. 2021 May 1:102158. doi: 10.1016/j.jogoh.2021.102158. [Epub ahead of print] Surgical management of deep pelvic endometriosis in France: do we need to be a pelvic surgeon to deal with DPE?

Pellerin M(1), Faller E(2), Minella C(1), Garbin O(3), Host A(3), Lecointre L(1), Akladios C(1).

Author information: (1)Gynaecology Unit, Hautepierre Hospital, University Hospitals of Strasbourg, 1 avenue Molière, 67000 Strasbourg France.

(2)Gynaecology Unit, Hautepierre Hospital, University Hospitals of Strasbourg, 1 avenue Molière, 67000 Strasbourg France.

Electronic address: emilie.faller@chru-strasbourg.fr.

(3)Gynaecology Unit, Centre medico chirurgical et obstetrical (CMCO), University Hospitals of Strasbourg, 19 rue Louis Pasteur, 67000 Schilligheim France.

INTRODUCTION: Endometriosis is a common disease in women, which requires a medical and surgical approach.

Figura 9: Pantalla de inicio de la aplicación. Después de introducir las palabras clave, el rango de fechas y pulsar en el botón "Buscar en PubMed" los resultados de la búsqueda se muestran en la zona derecha de la imagen. Seleccionando cualquiera de los resultados de la tabla, el texto del resumen correspondiente se muestra bajo la misma, junto a un hipervínculo para visitar la página en PubMed dedicada al artículo.

2.2 Actividades no previstas y realizadas

8

3 Relación de las desviaciones en la temporización y acciones de mitigación si procede y actualización del cronograma si procede

3.1 Desviaciones

3.2 Acciones de mitigación

3.3 Actualización del cronograma

4 Listado de los resultados parciales obtenidos hasta el momento (entregables que se adjuntan)

5 Apéndices

5.1 Apéndice A: Código

El documento original en formato .Rmd que incluye el código completo en lenguaje R usado para generar este informe (archivo `PEC3_fase2_informe.Rmd`), se puede consultar y descargar desde el siguiente repositorio en Github: [jorgevallejo/endometriosis-text-mining](https://github.com/jorgevallejo/endometriosis-text-mining)

5.2 Apéndice B: Reproducibilidad

```
sessionInfo() # For better reproducibility
```

```
## R version 3.6.3 (2020-02-29)
## Platform: i686-pc-linux-gnu (32-bit)
## Running under: Ubuntu 16.04.7 LTS
##
## Matrix products: default
## BLAS: /usr/lib/libblas/libblas.so.3.6.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8      LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8  LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8     LC_NAME=C
##  [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods    base
##
## other attached packages:
## [1] knitr_1.25
##
## loaded via a namespace (and not attached):
##  [1] rstudioapi_0.13    magrittr_1.5      tidyselect_1.1.0
##  [4] R6_2.4.0           rlang_0.4.10     stringr_1.4.0
##  [7] dplyr_1.0.5        tools_3.6.3      xfun_0.20
## [10] DBI_1.1.0          htmltools_0.5.1.1 ellipsis_0.3.1
## [13] yaml_2.2.0         digest_0.6.27    assertthat_0.2.1
## [16] tibble_3.0.4       lifecycle_1.0.0  crayon_1.3.4
## [19] purrr_0.3.4        vctrs_0.3.7      glue_1.4.2
## [22] evaluate_0.14      markdrive_0.0.2.900 rmarkdown_2.6
## [25] stringi_1.4.3      compiler_3.6.3   pillar_1.4.6
## [28] generics_0.1.0     googledrive_1.0.1 pkgconfig_2.0.3
```

6 Referencias