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

22/04/2021

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 **2**
2.1 Actividades previstas en el plan de trabajo 2
2.1.1 Tarea 1. Definir esquema de la aplicación web 2
2.2 Actividades no previstas y realizadas 7

3 Relación de las desviaciones en la temporización y acciones de mitigación si procede y actualización del cronograma si procede **7**
3.1 Desviaciones 7
3.2 Acciones de mitigación 7
3.3 Actualización del cronograma 7

4 Listado de los resultados parciales obtenidos hasta el momento (entregables que se adjuntan) **7**

5 Apéndices **7**
5.1 Apéndice A: Código 7
5.2 Apéndice B: Reproducibilidad 7

6 Referencias **8**

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 y una idea aproximada del aspecto final. Lo que empezó como unos garabatos en una libreta terminó convirtiéndose en una serie de diapositivas diseñadas con LibreOffice Draw que incluyo a continuación:

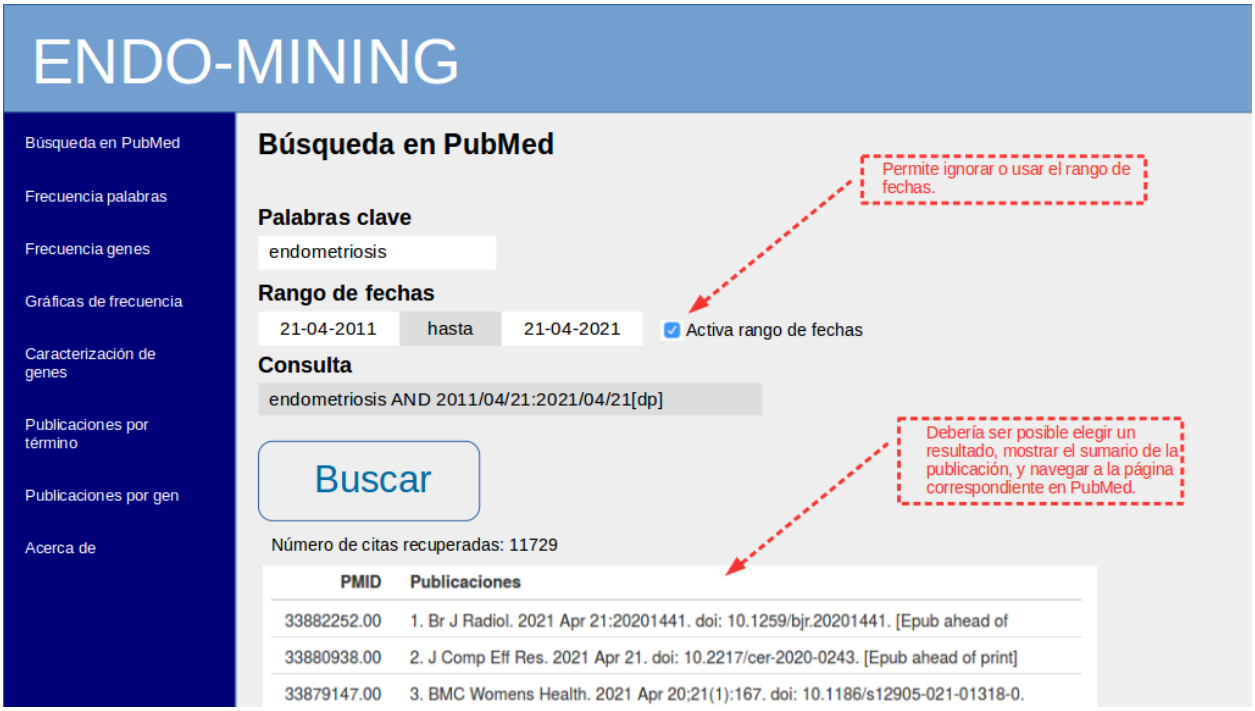

Figura 1: Pantalla inicial. Incluye los campos para entrada de palabras clave de búsqueda y rango de fechas. Muestra la cantidad de citas encontradas y una tabla con los datos de publicación de todas ellas.

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

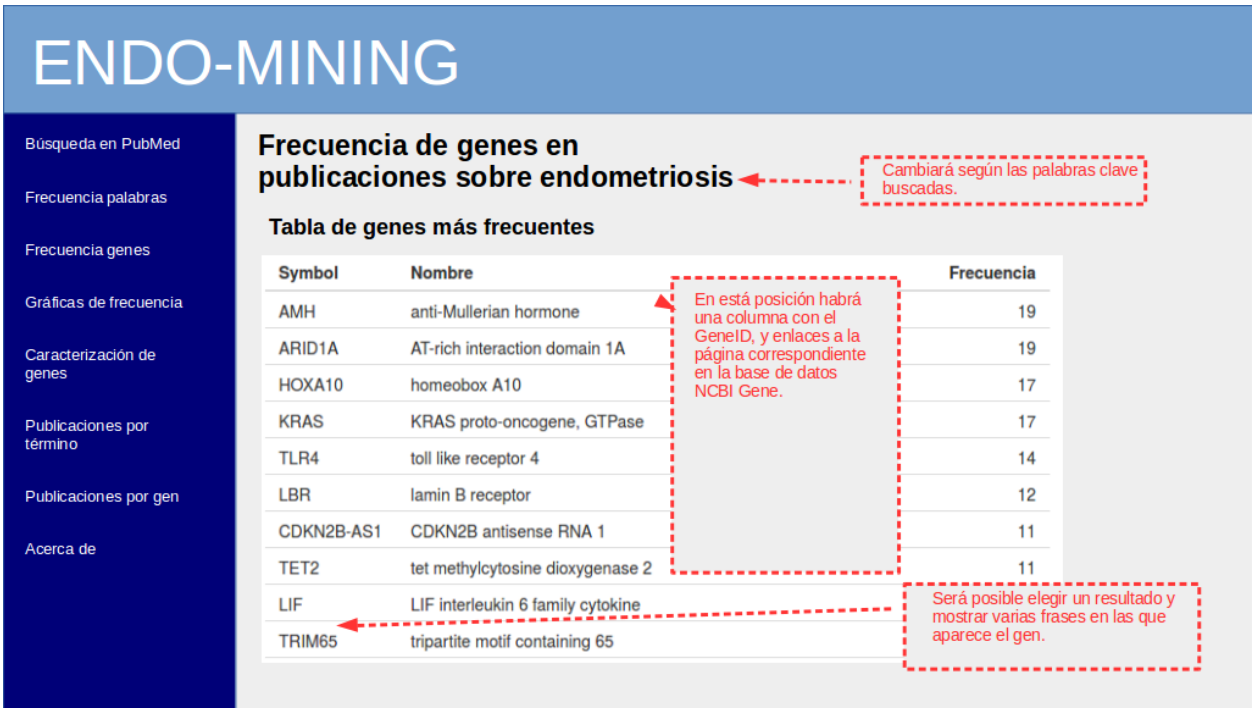

Figura 3: Frecuencia de genes. Muestra una tabla con los genes recuperados del corpus primario y su frecuencia, ordenados de mayor a menor.

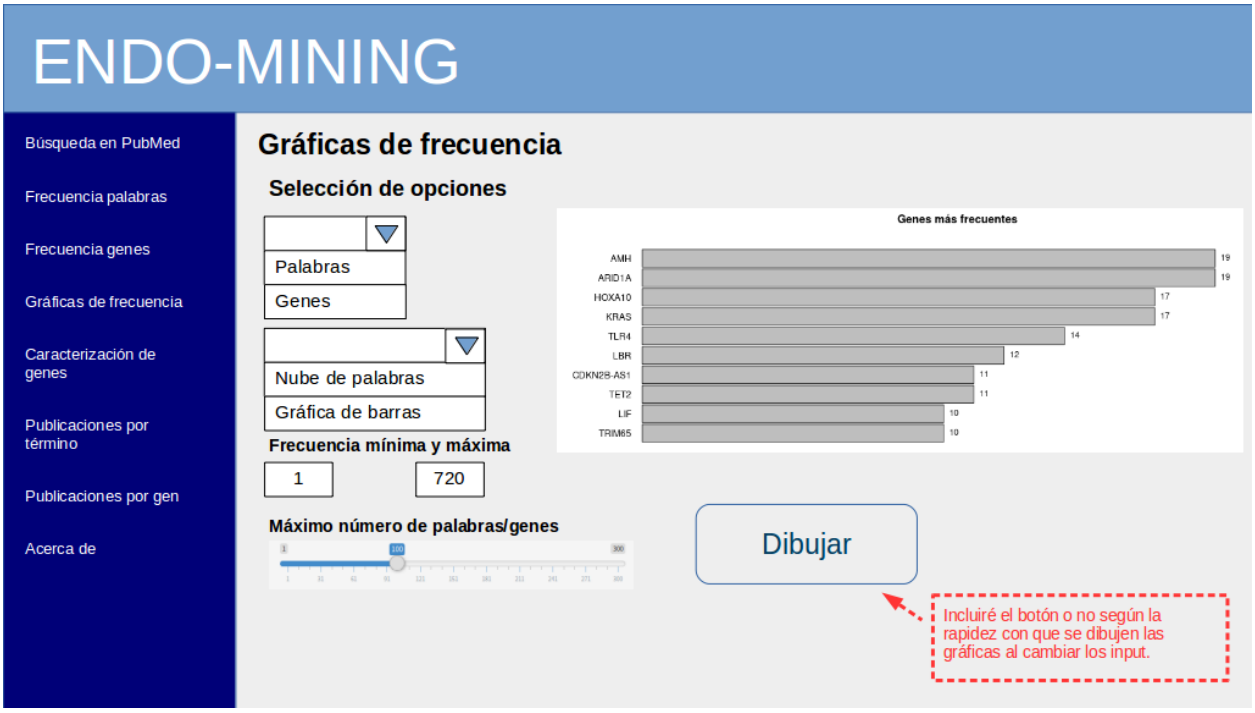

Figura 4: Gráficas de frecuencia de palabras y genes en el corpus primario. Listas desplegables permitirán elegir el tipo de gráfica (nube de palabras o gráfica de barras) y la entidad mostrada (palabras o genes). Se podrá elegir la frecuencia máxima y mínima de las entidades representadas, y cuántas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

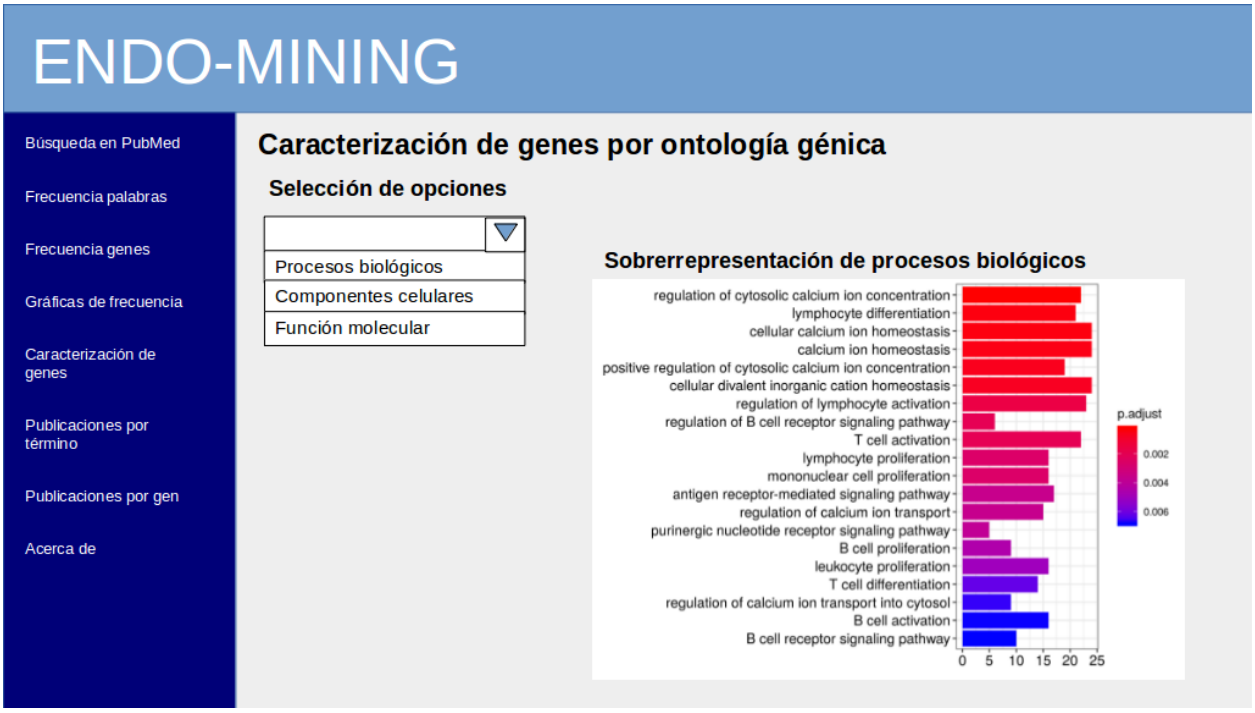

Figura 5: Caracterización por ontología génica. Muestra en una gráfica de barras los datos de ontología génica sobrerrepresentados en la muestra de genes. Con una lista desplegable se podrá elegir representar los datos correspondientes a procesos biológicos, componentes celulares o función molecular.

Figura 6: Filtrar publicaciones por término. El usuario podrá introducir una palabra y obtener como resultado un subgrupo del corpus conteniendo dicha palabra.

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

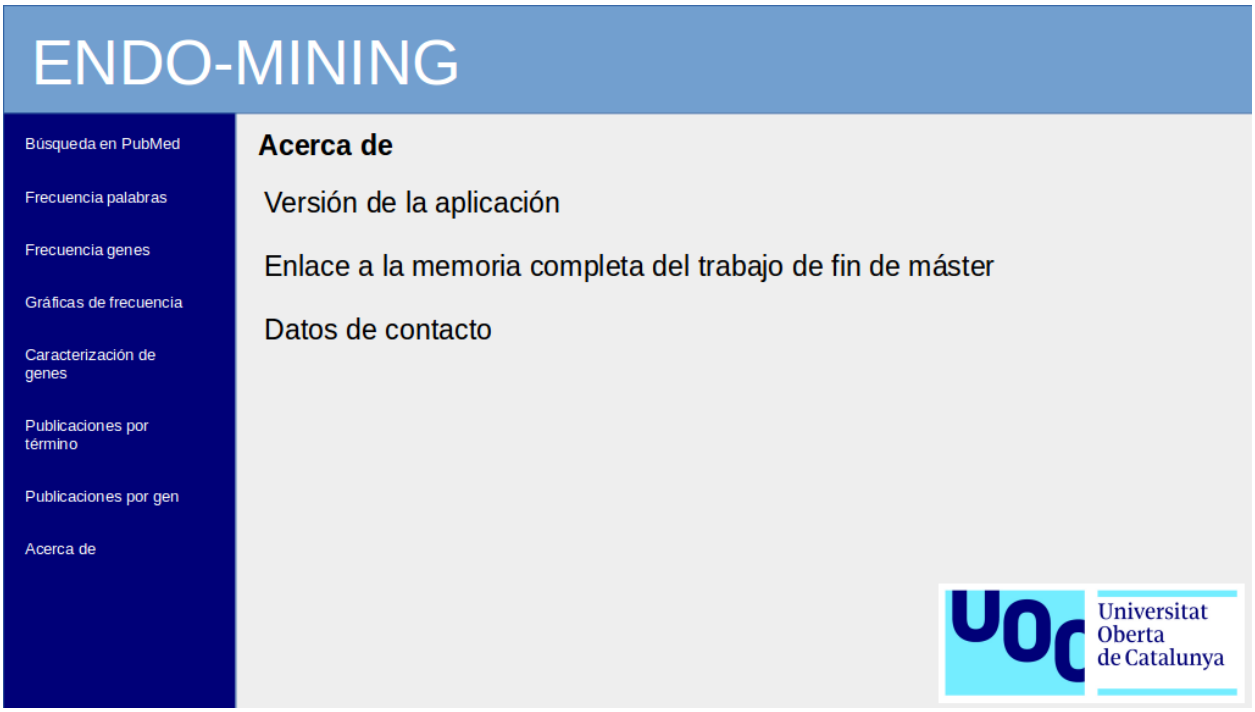

Figura 8: Esta sección de la aplicación mostrará al usuario información acerca de la propia aplicación, el trabajo de fin de máster en el que tiene su origen y los datos de contacto del creador.

2.2 Actividades no previstas y realizadas

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 (fichero `PEC3_fase2_informe.Rmd`), se puede consultar y descargar en el siguiente repositorio de Github: 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] compiler_3.6.3    magrittr_1.5      tools_3.6.3
##  [4] htmltools_0.5.1.1 yaml_2.2.0        stringi_1.4.3
##  [7] rmarkdown_2.6     stringr_1.4.0     xfun_0.20
## [10] digest_0.6.27     rlang_0.4.10      evaluate_0.14
```

6 Referencias