PEC 2 - Desarrollo del trabajo - Fase 1

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

Jorge Vallejo Ortega

15/04/2021

Índice

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.

El **objetivo general** podríamos decir que está completo en un 60-70%. Hemos generado las rutinas necesarias para, a partir de palabras clave (en nuestro caso "endometriosis", aunque se pueden usar otras) y rangos de fechas, recuperar sumarios de publicaciones científicas de la base de datos PubMed. A partir de dichos sumarios, las mencionadas rutinas reconocen los genes que en ellos aparecen y los organizan tanto en forma de tabla de contingencias como en forma de diagrama de barras para mostrarlos al usuario.

Para completar el objetivo general en un 100% planeamos implementar las siguientes funcionalidades:

- Filtrado que mantenga fuera de la lista genes que hayan aparecido con una frecuencia estadísticamente no significativa y puedan ser falsos positivos.

- Ofrecer al usuario la descarga del listado de genes en forma de archivo CSV.

En cuanto a los **objetivos específicos**, consideramos que el desarrollo del script está completo en un 30-40% y la implementación en forma de aplicación web completa en un 20-30%.

En estos momentos el **script** es capaz de, a partir de palabras clave y un rango de fechas, recuperar información de la base de citas bibliográficas PubMed, generar un corpus de sumarios a partir de dicha información, extraer los símbolos HGNC[1] de los genes nombrados en el corpus y su frecuencia, y mostrar esta información en formato tabla y como gráfico de barras. Además calcula también la frecuencia de aparición de palabras en el corpus, mostrando las más frecuentes en forma de tabla y de gráfico de barras.

Para completar el script en un 100% según lo planeado necesitamos implementar lo siguiente:

- **Filtrado estadístico** que permita mostrar sólo aquellos genes que se encuentren sobrerrepresentados en la información recuperada, con respecto al total de citas en la base de datos PubMed.

- **Caracterización funcional** de la lista de genes según términos de ontología génica (GO, *gene ontology*).

- Representación de los resultados de frecuencia de genes y palabras en forma de **nubes de palabras**.

- **Visualización** de los resultado de caracterización funcional.

Por su parte, la **aplicación web** ha conseguido implementar las mismas funcionalidades que actualmente ofrece el script. Hemos podido publicar, en la plataforma de hosting Shinyapps.io, un prototipo esquemático (https://endo-mining.shinyapps.io/shinyapp/) que recoge input del usuario (palabras clave y rango de fechas), recupera información de la base de datos PubMed, la procesa y muestra los resultados en pantalla.

Para completar la aplicación web en un 100% según lo planeado necesitamos implementar lo siguiente:

[1] HUGO Gene Nomenclature Commitee (Comité de Nomenclatura de Genes de la HUGO), https://www.genenames.org/

- Las funciones que se incluirán próximamente en el script (filtrado estadístico, caracterización funcional, visualización de frecuencias en forma de nube de palabras y visualización de los resultados de caracterización funcional).

- **Mensajes de error** que sean de mayor utilidad al usuario (por ejemplo, sugiriendo aumentar o reducir el rango/especificidad de búsqueda cuando no se recuperen genes, o se recuperen demasiados resultados, respectivamente).

- Posibilidad de **descargar** resultados en formato CSV.

- Diseño web que combine **usabilidad** sencilla y **estética** agradable a la vista.

1.2 Justificación de los cambios

De momento, el único cambio con respecto a lo planeado es la inclusión en los resultados de las **frecuencias de las palabras**. Esto se ha hecho por dos razones. Por una parte, barajamos la posibilidad de incluir el **análisis de factores de riesgo no genéticos** como parte de los resultados de la minería de texto. Para realizar eso son necesarios los datos de frecuencia de palabras clave. A estas alturas del proyecto, aunque no hemos descartado completamente realizar dicho análisis de los factores de riesgo, consideramos poco probable llevarlo a cabo debido a restricciones en el tiempo disponible.

En segundo lugar, ofrecer al usuario una visualización de las palabras más frecuentes en el corpus da una **idea general e intuitiva** del material recuperado en la búsqueda. La recuperación y representación de esos datos, sin ser trivial, es bastante sencilla y consideramos que valía la pena implementarlo como una herramienta simple para ayudar a la comprensión y familiarización con los datos.