

PubExplorer Text Mining Analysis



PubExplorer es una aplicación web desarrollada en R Shiny para el estudio y análisis de la **enfermedad de Crohn** mediante minería de textos de PubMed. La aplicación permite a los usuarios buscar artículos relevantes, analizar la frecuencia de palabras y genes, visualizar datos en tablas, gráficos y nubes de palabras, y explorar la co-ocurrencia de términos en los abstracts de las publicaciones.

📋 Tabla de Contenidos

1. [Descripción](#)
2. [Objetivos](#)
3. [Instalación](#)
4. [Uso](#)
5. [Estructura del Proyecto](#)
6. [Paquetes y librerías](#)
7. [Créditos](#)
8. [Capturas](#)

📝 Descripción

PubExplorer facilita la búsqueda y análisis de artículos relacionados con la enfermedad de Crohn en PubMed. Utilizando técnicas de minería de textos, la aplicación extrae información relevante de los abstracts, analiza la frecuencia de palabras clave y genes asociados, y proporciona visualizaciones para una mejor comprensión de los datos.

🎯 Objetivos

El objetivo principal es el desarrollo de una aplicación web interactiva, que use el sistema de búsqueda de PubMed, y permita a los usuarios extraer la información contenida en los abstracts de las publicaciones, aplicando técnicas de minería de textos.

- Facilitar la búsqueda de artículos en PubMed sobre la enfermedad de Crohn.

- Proporcionar estadísticas sobre la frecuencia de palabras clave y genes asociados a la enfermedad.
- Analizar la proximidad de co-ocurrencia entre diferentes términos en los abstracts.
- Generar visualizaciones: nubes de palabras, gráficos de barras y tablas para facilitar el análisis e interpretación de datos.
- Acceso a enlaces directos: PubMed, DOI, QuickGO, UniProt.

🔧 Instalación

⌚ Requisitos Previos

- **R** (versión 4.0 o superior)
- **RStudio** (opcional pero recomendado)

📦 Paquetes Necesarios

Es necesario tener instalados los siguientes paquetes de R. Se pueden ejecutar con el siguiente código de R:

```
install.packages(c(  
  "shiny",  
  "shinydashboard",  
  "shinyjs",  
  "fs",  
  "DT",  
  "wordcloud",  
  "ggplot2",  
  "pubmed.mineR",  
  "easyPubMed",  
  "lsa",  
  "tokenizers",  
  "viridis"  
)
```

🔧 Configuración del Proyecto

Repository github

Para clonar el repositorio:

```
git clone https://github.com/megamcald/PubExplorer.git
```

📁 Estructura del Proyecto

```
TFM-project/  
├── app.R                      # Archivo principal de la aplicación Shiny  
├── README.md                    # README  
└── www/  
    ├── styles.css                # Estilos CSS  
    └── images/                   # Capruras aplicación de Shiny (desarrollo)
```

```
├── logos/          # Logos
│   ├── lupa.png
│   ├── logo_appv2.png
│   └── logo_uoc.png
└── spinnerv2.gif    # GIF del spinner de carga
01_data/
├── raw/
│   └── crohns_disease/  # Datos crudos descargados de PubMed
└── processed/        # Datos procesados
03_results/
├── word_frequency/
│   ├── word_tokens.txt
│   └── word_tokens_barplot.png
└── gene_analysis/
    ├── genes_df.txt
    └── genes_barplot.png
```

Paquetes y librerías

- *Shiny*
 - shiny: Creación de aplicaciones web interactivas.
 - shinydashboard: Creación de paneles de control atractivos.
 - shinyjs: Facilita el uso de JavaScript en Shiny.
- *Manejo de rutas*
 - fs: Manipulación de rutas de archivos y directorios.
- *Visualización* DT: Creación de tablas interactivas. wordcloud: Generación de nubes de palabras. ggplot2: Creación de gráficos estáticos y dinámicos.
- *Acceso PubMed*
 - pubmed.mineR: Acceso y análisis de datos de PubMed.
 - easyPubMed: Consultas y descargas desde PubMed.
- *Minería de textos*
 - Isa: Análisis semántico latente.
 - tokenizers: Tokenización de textos.
- *Otros*
 - viridis: Paletas de colores para visualizaciones.

Créditos

Este proyecto fue desarrollado como parte del Trabajo Final del Máster en Bioinformática y Bioestadística en la UOC-UB (Universitat Oberta de Catalunya - Universitat de Barcelona).

Desarrollado por: [Megam Calderón](#) 

Capturas

Pantalla principal

PubExplorer

Búsqueda en PubMed

Puedes ingresar palabras clave y un rango de fechas para buscar artículos relevantes

Parámetros de búsqueda

Palabras clave: Crohn's disease

Rango de fechas: 24-12-2019 hasta 22-12-2024

Consulta a PubMed: Crohn's disease[MH] AND 2019-12-24:2024-12-22[dp]

Buscador

2024 | TFM Máster Universitario en Bioinformática y Bioestadística (UOC-UB)

Pantallas Palabras

PubExplorer

Tabla Frecuencia Palabras

Aquí se muestra una tabla con las palabras más frecuentes

words	Frecuencia
patients	3579
disease	3140
cd	2164
university	2069
s	2013
ibd	1656
crohn	1597
hospital	1504
clinical	1236
gastroenterology	1105

Mostrando registros del 1 al 10 de un total de 28,674 registros

Anterior 1 2 3 4 5 ... 2,868 Siguiente

2024 | TFM Máster Universitario en Bioinformática y Bioestadística (UOC-UB)

PubExplorer

Tabla Frecuencia Palabras

Aquí se muestra una tabla con las palabras más frecuentes

Palabras	Frecuencia
patients	3579
disease	3140
cd	2164
university	2069
s	2013
ibd	1656
crohn	1597
hospital	1504
clinical	1236
gastroenterology	1105

Palabras

2024 | TFM Máster Universitario en Bioinformática y Bioestadística (UOC-UB)

PubExplorer

Tabla Frecuencia Palabras

Aquí se muestra una tabla con las palabras más frecuentes

Ver como:

Nube de palabras

Frecuencia mínima:

2

Máximo número de palabras:

50

Dibujar Nube

2024 | TFM Máster Universitario en Bioinformática y Bioestadística (UOC-UB)

Pantallas Genes

PubExplorer

Tabla Frecuencia de Genes

Aquí se muestra una tabla con los genes más frecuentes

Ver como:

Tabla

Frecuencia mínima:

2

Máximo número de genes:

50

Dibujar Nube

Mostrar	10	registros	Buscar:	Freq
Gene_symbol	Genes			
1	NHS	NHS actin remodeling regulator	77	
2	ADA	adenosine deaminase	52	
3	UST	uronyl 2-sulfotransferase	52	
4	HR	HR lysine demethylase and nuclear receptor corepressor	45	
5	CRP	C-reactive protein	44	
6	POR	cytochrome p450 oxidoreductase	33	
7	TNF	tumor necrosis factor	25	
8	SCD	stearoyl-CoA desaturase	18	
9	MSC	muzculin	15	
10	MPO	myeloperoxidase	14	

Mostrando registros del 1 al 10 de un total de 314 registros

Anterior 1 2 3 4 5 ... 32 Siguiente

Información Gene_symbol: NHS

[Link QuickGO](#)

[Link UniProt](#)

PubExplorer

Tabla Frecuencia de Genes

Aquí se muestra una tabla con los genes más frecuentes

Ver como:

Grafico

Frecuencia mínima:

2

Máximo número de genes:

50

Dibujar Nube

2024 | TFM Máster Universitario en Bioinformática y Bioestadística (UOC-UB)



Pantallas Navegación

The screenshot shows the PubExplorer interface with a sidebar on the left containing navigation links: Búsqueda en PubMed, Palabras, Genes, Navegación Genes, Navegación palabras, Co-ocurrencia de términos, and Acerca de. The main area is titled "Navegación Genes" with the sub-instruction "Aquí se puede navegar por aquellas publicaciones que contienen determinados términos (Gene_symbol) en el abstracto de los artículos relacionados con la enfermedad de Crohn." It includes filters for "Artículos Gene_symbol: NHS", "Link QuickGO | Link UniProt", "Mostrar: 10 registros", and a "Buscar:" input field. The results table has columns for PMID and Publicación. The first result is highlighted in blue: "1 39567783 participate The study had been performed in accordance with the Declaration of Helsinki and had been approved by the medical ethics committee of The Second Hospital of Jiaxing (No.).". Below the table is a note: "participate: The study had been performed in accordance with the Declaration of Helsinki and had been approved by the medical ethics committee of The Second Hospital of Jiaxing (No.)." and a link to "https://pubexplorer.com/declaration-of-helsinki".

The screenshot shows the PubExplorer interface with a sidebar on the left containing navigation links: Búsqueda en PubMed, Palabras, Genes, Navegación Genes, Navegación palabras, Co-ocurrencia de términos, and Acerca de. The main area is titled "Navegación términos" with the sub-instruction "Aquí se puede navegar por aquellas publicaciones que contienen determinados términos (word) en el abstracto de los artículos relacionados con la enfermedad de Crohn." It includes filters for "Artículos que contienen la palabra: patients", "Mostrar: 10 registros", and a "Buscar:" input field. The results table has columns for PMID and Publicación. The first result is highlighted in blue: "1 39706833 patients competing interests. Ethics approval: All experiments were performed in". Below the table is a note: "patients: competing interests. Ethics approval: All experiments were performed in" and a link to "https://pubexplorer.com/competing-interests".

Pantalla Co-ocurrencia términos

Pantalla Acerca de