GeneDetector

A Program to Dig Out Gene Related to Input Neurodegenerative Diseases

Jiayi Li, Andrew ID: lijiayi

Problem:

The goal of GeneDetector is to analyze abstracts from Pubmed, and identify the related gens according to input diseases.

Progress:

Once we input the gene name and the number of pages you want to download from the PubMed search result webpage(1 page including 10 abstracts), we can get all the gene names and other information about the paper. Though some names are not a name of the gene, and I don't specify which genes are actually related to the input disease.

Current feature:

- Input
  - Keyword about disease name . e.g. Alzheimer’s
  - The number of pages of abstract want to download.
- Output
  - A cvs file including paper title, url, abstract, gene name, mid, doi, and keyword.

Example command:

(There are hint sentences guiding you want you should input)

```
Alzheimer’s
30
```

Changes:

- At first, I limit the type of diseases to 4 kinds of neurodegenerative diseases. Now I think it can accept any disease keyword input.
- Also, I want to change the purpose of my project a little bit. At first, I plan to get the "relationships" between genes and diseases. Now I decide only to get the gene name that actually related to the input keyword. There are already many factors I need to think about in this program in order to make it more accurate.

Challenges:

I think I chose a challenging project for myself. My programming background is weak since I didn’t do any programming project before this class, though I learned a lot from this course. There are a lot of challenges. So far, doing web scraping, analyzing the HTML, downloading the abstract from PubMed, using the library, and finding the right gene name are all difficult. Current results only get all the gene names from the abstract using the "simple pattern matching," which means some are "trash" names, not even a gene name, and I don't specify which genes are related to the input keyword yet. Next step, I want to use natural language processing(NLP) to analyze the sentences in the abstract and finally return the gene name that is more accurately related to the input disease name.