1.探寻数据

在DOSE中引用的DO.db的数据：

数据来源：HumanDO.obo

[HumanDiseaseOntology/src/ontology at main · DiseaseOntology/HumanDiseaseOntology (github.com)](https://github.com/DiseaseOntology/HumanDiseaseOntology/tree/main/src/ontology)

DOANCESTOR: between DO terms and their ancestor (parents and all their parents)

DOOFFSPRING: between DO terms and their offsprings terms

DOPARENTS: between DO terms and their direct parents terms

DOTERM

dotb1.rda

Gene-disease关联信息

数据来源：do\_rif.human.txt <http://projects.bioinformatics.northwestern.edu/do_rif/>

IDMapping.txt http://doa.nubic.northwestern.edu/pages/download.php

DO2EG.rda

EG2DO.rda

EG2ALLDO.rda

DO2ALLEG.rda

DOIC.rda

2.数据的处理

DOTERM

1. 以table形式读取HumanDo.obo
2. 提取每个term的相关信息，“term”、”id”、”name”、”synonym”
3. 建立一个n\*4的数据框，用以存储数据
4. 在for循环中，以“term”为分界线，对每一个term进行处理，将doid,name,syynonym和alt\_id放在相应的位置。有些term的”alt\_id”有很多个，而这里仅是保存了最后一个。在数据框df中，每一行代表着一个term的信息，存在一些空白行。
5. 将含有信息的行提取出来，并去掉一些字符，然后保存。

DOPARENTS

1. 以table形式读取HumanDo.obo
2. 提取每个term的相关信息，“term”、”id”、”is\_a”
3. 建立一个n\*2的数据框，用以存储数据
4. 在for循环中，term有两种情况，一种是有“is\_a”信息的，一种是没有“is\_a”信息的。在这里分情况进行处理，确保每循环一次都匹配到字符串”Term”。然后将数据提取到数据框df中。
5. 最后提取有用的信息行，并手动添加最后一行的信息。因为循环条件的原因，最后一行数据并没有进行处理。