**Diseases lexicon**

**Resources:**

* Unified Medical Language System (UMLS) Metathesaurus
* Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED - CT)

**UMLS Metathesaurus:**

Compiling UMLS Metathesaurus is a separate project. Please see our project at https://github.com/CutaneousBioinf/LiteratureMiningTool/tree/master/ConceptMap/UMLSMetathesaurusCompiler

Diseases and synonyms can be retrived based on the UMLS semantic types related to DISORDER. We developed a Java program to achieve the same.

$ javac DiseaseDictionaryGenerator.java

$ java DiseaseDictionaryGenerator ALL\_CONCEPTS\_FILE UMLS\_DISEASE\_FILE

**SNOMED - CT:**

We need UMLS Metathesaurus license for downloading and using SNOMED CT. The resource can be downloaded from UMLS Terminology Services (https://uts.nlm.nih.gov//home.html). Download UMLS Metathesaurus and install on server by selecting the option SNOMED CT.

Processing MRCONSO.RRF a file:

$ javac MRCONSOPreferredVocabularyGenerator.java

$ java MRCONSOPreferredVocabularyGenerator ~/MRCONSO.RRF OUTPUT\_FILE1

Processing MRSTY.RRF a file:

A. Extraction of unique concept identifier (CUI) and unique semantic type identifier (TUI)

$ javac MRSTYCuiTuiGenerator.java

$ java MRSTYCuiTuiGenerator ~/MRSTY.RRF OUTPUT\_FILE2

B. Grouping of concepts with same CUI, but different TUI

$ javac MRSTYCuiTuiGrouper.java

$ java MRSTYCuiTuiGrouper OUTPUT\_FILE2 OUTPUT\_FILE3

The processed files from MRCONSO.RRF and MRSTY.RRF are used for generating diseases lexicon. We need semantic group file (i.e. SemGroups\_2018.txt) which can be downloaded from NCBI MetaMap (https://metamap.nlm.nih.gov/Docs/SemGroups\_2018.txt). The execution may take several hours depending on the system or server used to run the program.

$ javac MRCONSODictionaryPreparer.java

$ java OUTPUT\_FILE3 SemGroups\_2018.txt OUTPUT\_FILE1 OUTPUT\_FILE4

We developed a Java program to extract disease concepts from SNOMED CT.

$ javac DiseaseDictionaryGenerator.java

$ java DiseaseDictionaryGenerator OUTPUT\_FILE4 SNOEMED\_DISEASES\_FILE

**UMLS Metathesaurus + SNOMED CT:**

SNOMED CT is from UMLS Metathesaurus only. We assume that the disease concepts from both resources may overlap. We also noticed that the concepts have same CUI\_TUI. For example, ‘acute abdominal pain syndrome’ has CUI\_TUI: C0000727\_T184 in both resources.

We combined the resources using Linux command.

$ cat UMLS\_DISEASE\_FILE SNOEMED\_DISEASES\_FILE | sort | uniq -i > DISEASES\_LEXICON\_v1

**Post processing of diseases lexicon:**

We noticed stop words as disease synonyms (e.g. ‘In’ for ‘Present’). Such synonyms will give false frequency count in the literature. Therefore, it is good to remove the disease synonyms that are actually stop words.

$ javac StopwordsAsDiseaseSynonyms.java

$ java StopwordsAsDiseaseSynonyms DISEASES\_LEXICON\_v1 STOPWORDS\_AS\_DISEASES\_AND\_SYNONYMS

$ javac StopwordsAsDiseaseSynonymsRemover.java

$ java StopwordsAsDiseaseSynonymsRemover STOP\_WORDS DISEASES\_LEXICON\_v1 DISEASES\_LEXICON\_v1\_NO\_STOPWORDS

We collected stopwords from our previous work and various resources. File: stopwords.txt

We noticed that certain disease concepts map to multiple CUIs. This will lead to independent counts for each CUI. Therefore, we combined the multiple CUIs together and assigned a customized CUI.

$ javac DiseasesWithMultipleIDRetriever.java

$ javac DiseasesWithMultipleIDRetriever DISEASES\_LEXICON\_v1\_NO\_STOPWORDS DISEASES\_WITH\_MULTIPLE\_IDs

$ javac DiseasesWithMultipleCUICompiler.java

$ java DiseasesWithMultipleCUICompiler DISEASES\_LEXICON\_v1\_NO\_STOPWORDS DISEASES\_WITH\_MULTIPLE\_IDs OUTPUT\_FILE5

$ javac DiseasesWithModifiedCUIUpdater.java

$ java DiseasesWithModifiedCUIUpdater OUTPUT\_FILE5 DISEASES\_LEXICON\_v1\_NO\_STOPWORDS DISEASES\_LEXICON\_v2