**Documentation for CLAMP**

**Introduction:**

The CLAMP System is a comprehensive clinical Natural Language Processing software that enables recognition and automatic encoding of clinical information in narrative patient reports. In addition to running a clinical concept extraction pipeline as well as an annotation pipeline, the individual components of the system can also be used as independent modules. The system lends itself for diverse applications in a broad range of clinical domains.

**NLP Pipelines:**

CLAMP components builds on a set of high performance NLP components that were proven in several clinical NLP challenges such as i2b2 , ShARe/CLEF , and SemEVAL. A pipeline can be created and customized by a simple drag and drop on the individual CLAMP components in the order that is desired. Upon creation of the pipeline, CLAMP checks for errors in sequence and directs the user to the appropriate logical order with insertion of the required components for a working pipeline.

CLAMP also has Machine Learning approaches, however they are not very well built and are under development

**Installation process:**

Go to Clamp website and into their downloads folder. Over there, you can request a version(GUI or Command Line) either for Mac or Windows. You will need to fill their form where you provide your details. <https://clamp.uth.edu/get-clamp.php>

After a few days, your details get approved and you get a download link on your email Id to install CLAMP directly to your laptop.

Using the license provided by them in the same mail, CLAMP can be easily installed.

**Components of CLAMP:**

MyCorpus: contains the customized corpus built by the users.

MyPipeline: contains the customized pipeline created by users for clinical notes processing.

PipelineLibrary: contains the built-in pipelines ready to use for a series of common clinical applications.

Log: Includes CLAMP run-time log files

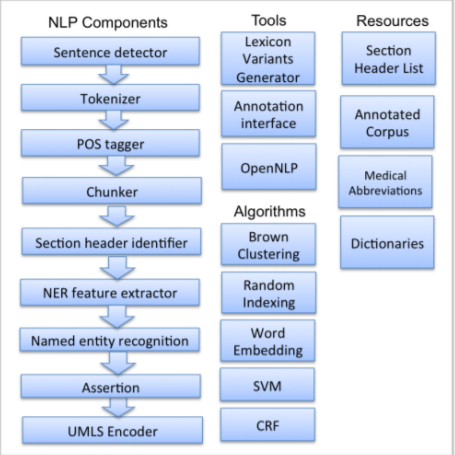
Metadata: The metadata used by CLAMP are included in this folder.

Resources: This folder includes third-party libraries. Currently it has two items:

CRFSuite: the CRF implementation for Name Entity Recognition tasks

Umls\_index: the Lucene index built for CLAMP based on the UMLS thesaurus.

If you want to use UMLS terminologies, then you will need to create an UMLS account. Please follow the following link to create an UMLS account if you do not have any. [https://uts.nlm.nih.gov//license.html](https://uts.nlm.nih.gov/license.html)



**How to use Clamp:**

The main link to learn using clamp is as follows: <https://clamp.uth.edu/manual.php>

CLAMP, mainly, extracts 3 types of clinical information-

The problem/disease- The medical issues faced by the patient eg- synovitis, bruits etc

The treatment -The medication provided by the doctor eg- prednisone

Tests- What kind of tests the doctors have prescribed to a patient eg-(Ultrasound study)

The documentation to using the basic pipeline as well as inbuilt pipelines in clearly provided in this link from section 9 onwards. It is intuitive. <https://clamp.uth.edu/manual.php>

There are many videos available to learn how to use CLAMP:

<https://clamp.uth.edu/tutorial.php>

I have used CLAMP on your i2b2 files in order to extract medications, diseases and tests. I have also used CLAMP to extract smoking information. The F1-score was 85%.

After this, I decided to use my own ruled based system as well as other systems such as word2vec and tfidf on the i2b2 files to get the smoking scores. They are provided in the notebook that I have shared.

There are 3 ipynb files.

1. The first file converts i2b2 files from xml to txt (only the notes part from the entire xml file)

**Notebook name: Converting xml\_to\_txt**

1. The second file extracts the smoking information from the i2b2 files (The actual labels)

**Notebook name: extracting smoking labels**

1. The third file has my rule based system as well as tf-idf and word2vec to extract predicted smoking labels

**Notebook name: Ruled based, tfidf, word2vec**