========== Software for Clinical Language Annotation Modeling Processing========

CLAMP is a comprehensive Clinical Natural Language Processing software that enables a pipeline of text analysis for narrative clinical notes. CLAMP provides a wide range of clinical text processing functions, from sentence segmentation and tokenization, to clinical concepts recognition, encoding and assertion identification. This is a brief document for the command line version of CLAMP.

Please take a special note that in order to use this command line version of CLAMP, the user name and password of UMLS (Unified Medical Language System) must be provided.

========== Function modules in CLAMP ==========
During the text analysis process of CLAMP, the following seven functional modules will be executed sequentially as a pipeline:

Sentence detector: identifies the boundaries of each single sentence in text.

Tokenizer: segments the sentence into a sequence of tokens.

POS tagger: assigns part of speech to each token.

Section Identifier: identifies section headers in a clinical note and maps them into general categories. For example, the section header “icd10 code” will be generalized to “icd_code”.

Named entity recognizer: identifies named entities, i.e. clinical concepts, and their semantic types. CLAMP by default allows identification of the following three types of clinical concepts: “problem”, “treatment” and “test”.

Negation assertion recognizer: checks whether there is a negation assertion related to a specific clinical concept in the text. A negation assertion would indicate the absence of the corresponding clinical concept in the specific context, whereas if there is no negation, it indicates the presence of the concept.

UMLS encoder: maps the terms of clinical concepts to their corresponding CUIs in UMLS. For example, the term “breast cancer” will be encoded with the CUI of “C6006142” in UMLS.

========== Quick Start ==========
CLAMP provides a set of command options to start the command line version easily, which are listed below:

-h Print this message
-i <arg> Directory path for the input file folder, only accepts files with a .txt extension.
-I <arg> Directory path for the UMLS index folder.

-o <arg> Directory path for the output file folder.
-p <arg> Path for the pipeline .jar file, if available. The jar file can be generated using CLAMP's GUI version.
-P <arg> UMLS password.
-U <arg> UMLS username. (Register first if you don't have an account with UMLS here at https:/uts.nlm.nih.gov/license.html)

As an example of using CLAMP in conjunction with UMLS, CLAMP provides the run.sh file. Please open the run.sh file and enter in your own UMLS user name and password. If you don't have an UMLS account, please register as described above.

========== Output Format==========
After CLAMP runs through the analysis process, the output of each functional module is saved in two files, namely a txt file and a xmi file.

txt file: provides a view of the tab delimited output. The output information is provided at the section, token and the named entity levels:

-The specific sentence
-Start index: Starting position of the sentence.
-End index: Ending position of the sentence.
-Section: Name of the section the sentence is in.

-The specific Token
-Start index: Starting position of the token.
-End index: Ending position of the token.
-POS: POS tag of the token.

-The specific Named Entity
The detailed information of a recognized clinical concept is listed in one line in the file, the types of information include:
-Start index: Starting position of the named entity.
-End index: Ending position of the recognized concept.
-Semantic type: Semantic type of the recognized concept.
-CUI: The Concept Unique Identifier of the concept in UMLS.
-Assertion: if the named entity is "present" or "absent" according to its context in the clinical note
-Concept mention: mention of the concept, i.e., named entity in the text.

Following is an example of the input text and CLAMP output:
input:

Description: Discharge summary of a patient with mood swings and oppositional and defiant

behavior.

output:

```
NamedEntity 50  61  semantic=problem    assertion=present   cui=C0085633    ne=mood swings
NamedEntity 66  99  semantic=problem    assertion=present   cui=C0029121    ne=oppositional and defiant behavior
Sentence    0   100 section=description
Token   0   11  pos=NN
Token   11  12  pos=:
Token   14  23  pos=NN
Token   24  31  pos=NN
Token   32  34  pos=IN
Token   35  36  pos=DT
Token   37  44  pos=NN
Token   45  49  pos=IN
Token   50  54  pos=JJ
Token   55  61  pos=NNS
Token   62  65  pos=CC
Token   66  78  pos=JJ
Token   79  82  pos=CC
Token   83  90  pos=JJ
Token   91  99  pos=NN
Token   99  100 pos=.
```

xmi file: provides a view of the text annotation. This file is for use in the GUI version of CLAMP. Opening this file in the interface of CLAMP provides a view of the clinical text with highlighted clinical concepts in a new window.