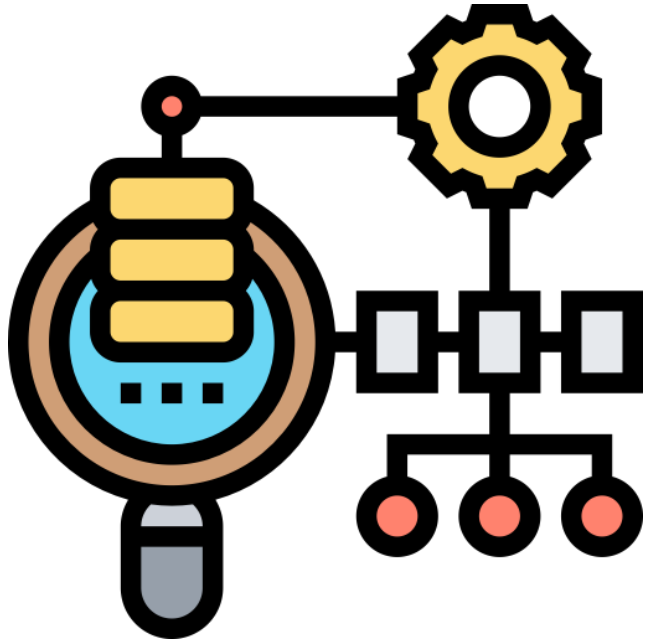
# Text Data Mining
# For Business Decisions

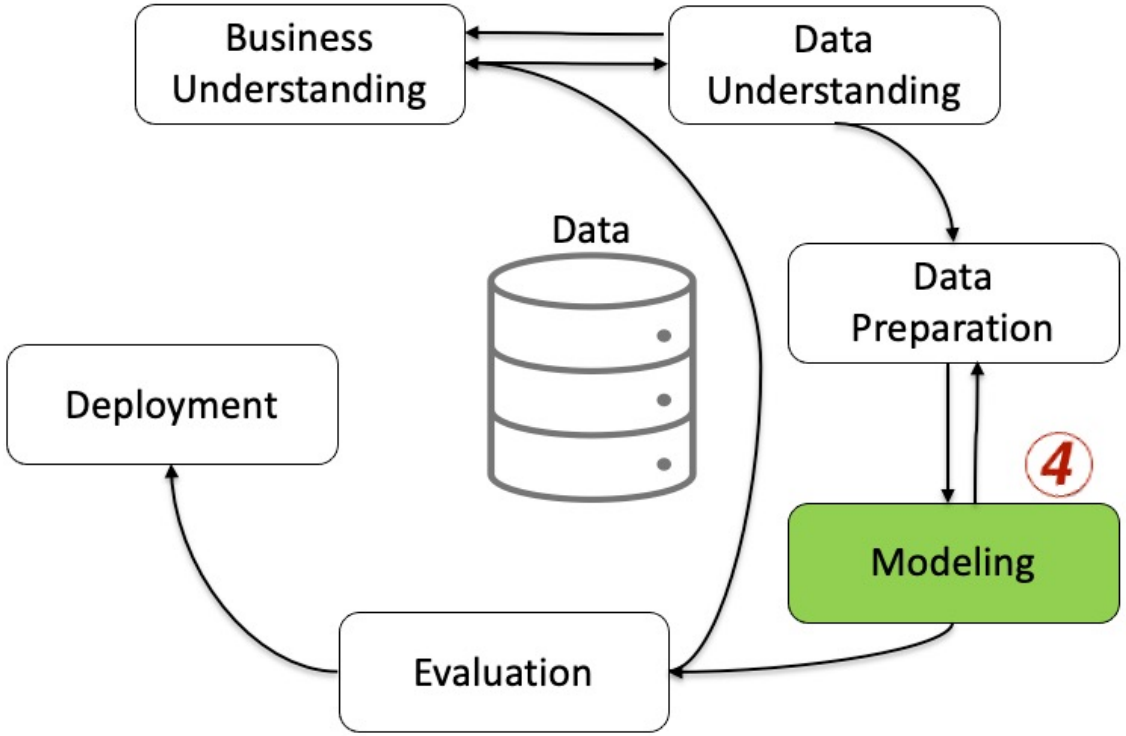


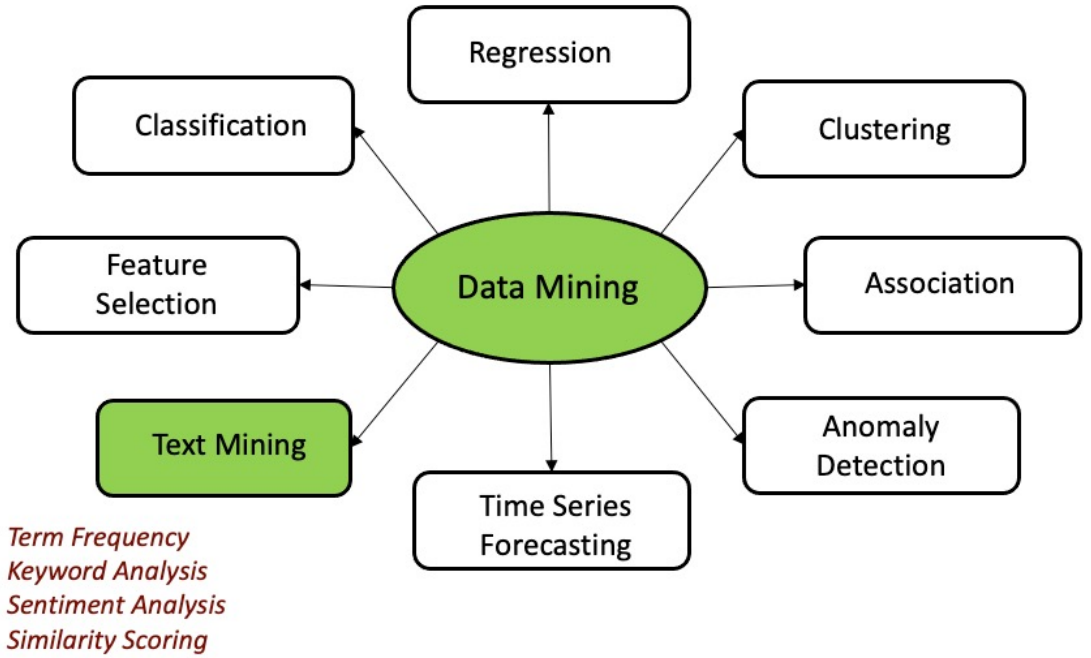
Module 10

Named Entity Recognition

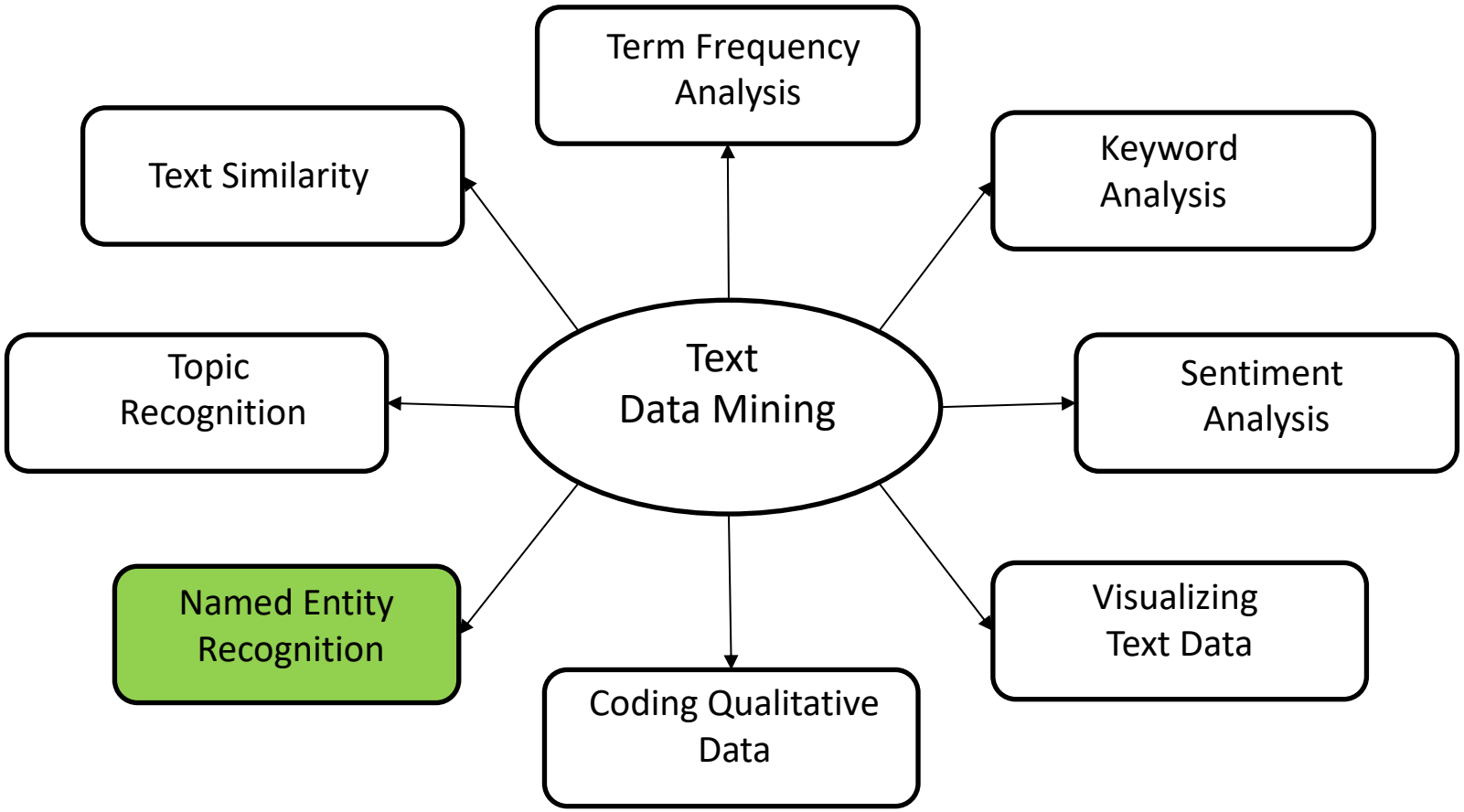
How do we extract the entities from text data?

# Data Mining- Continuing with Model-Making



# Named Entity Recognition



# What is NER?

- *Named Entity Recognition*, NER, is a standard NLP problem dealing with information extraction.
- The primary objective is to locate and classify named entities in a text into predefined categories such as he names of persons, organizations, locations, events, expressions of times, quantities, monetary values, and percentages.
- NER systems extract real-world entities from the text, such as a person's name, an organization, or an event. NER is also known as entity identification, entity chunking, or entity extraction.
- Extracting the leading entities in a text helps sort unstructured data and detect important information, which is crucial to deal with large datasets.

# How NER works

- Most NER systems are programmed to take an unannotated block of text, such as

  *“Mary bought 300 acres of Blueberry Hill Farm in 2020”*

- and produce an annotated block of text that highlights the names of entities:

  *“[Mary]Person bought 300 acres of [Blueberry Hill Farm]Organization in [2020]Time.”*

- In this example, a person’s name consisting of one token, a three token company name, and a temporal expression were detected and classified.

# How is NER used?

- NER is part of the emerging field of information extraction, a critical area in our current information-driven world.
- Rather than indicating which documents need to be read by a user, it extracts pieces of information that are important to the user's needs.
- Links between the extracted information and the original documents are maintained to allow the user to reference context.
- The kinds of information that these systems extract vary in detail and reliability.
  - For example, named entities such as persons and organizations can be extracted with a high reliability but do not provide attributes, facts, or events that those entities have or participated in.
- In our case, we concentrate on named entity extraction.

# What Does This Mean for Business?

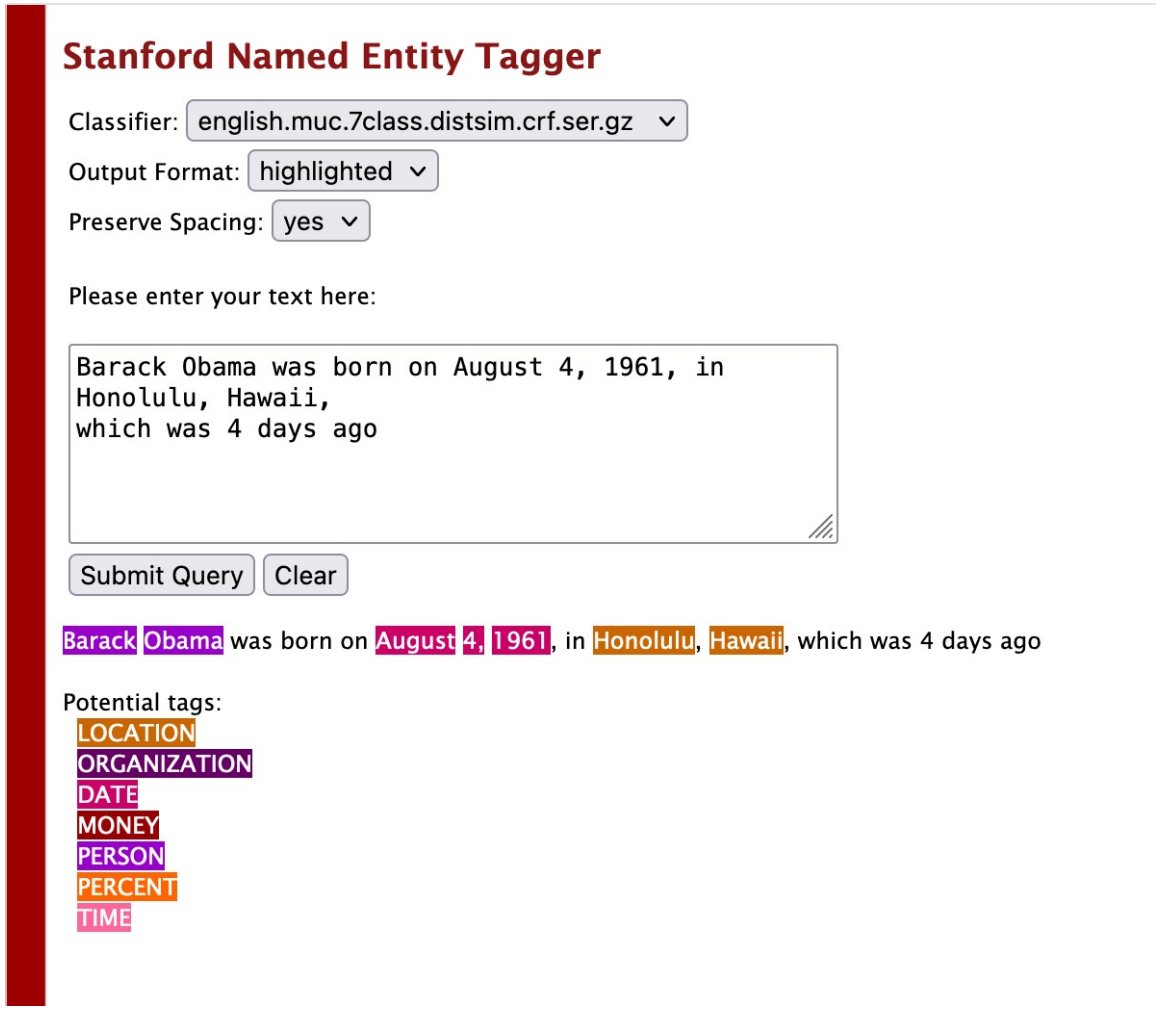
- A NER may add semantic understanding to any large body of the text.
- It has multiple business use-cases, such as classifying and prioritizing news content for newspapers.
- It can also generate candidate shortlists from a large number of CVs for recruiters, for example.
- When integrated into email and chat systems, the use of NER technology could enable a business to extract and collate information from large amounts of documentation across multiple communication channels in a much more streamlined, efficient manner.
- Using a NER allows you to instantly view trending topics, companies, or stock tickers, and provides you with a full overview of all your information channels containing relevant content, such as meeting notes shared via email or daily discussions over chat systems.
- In a world where business managers can send and receive thousands of emails per day, removing the noise and discovering the true value in relevant content may be the difference between success and failure.

# What is a named entity?

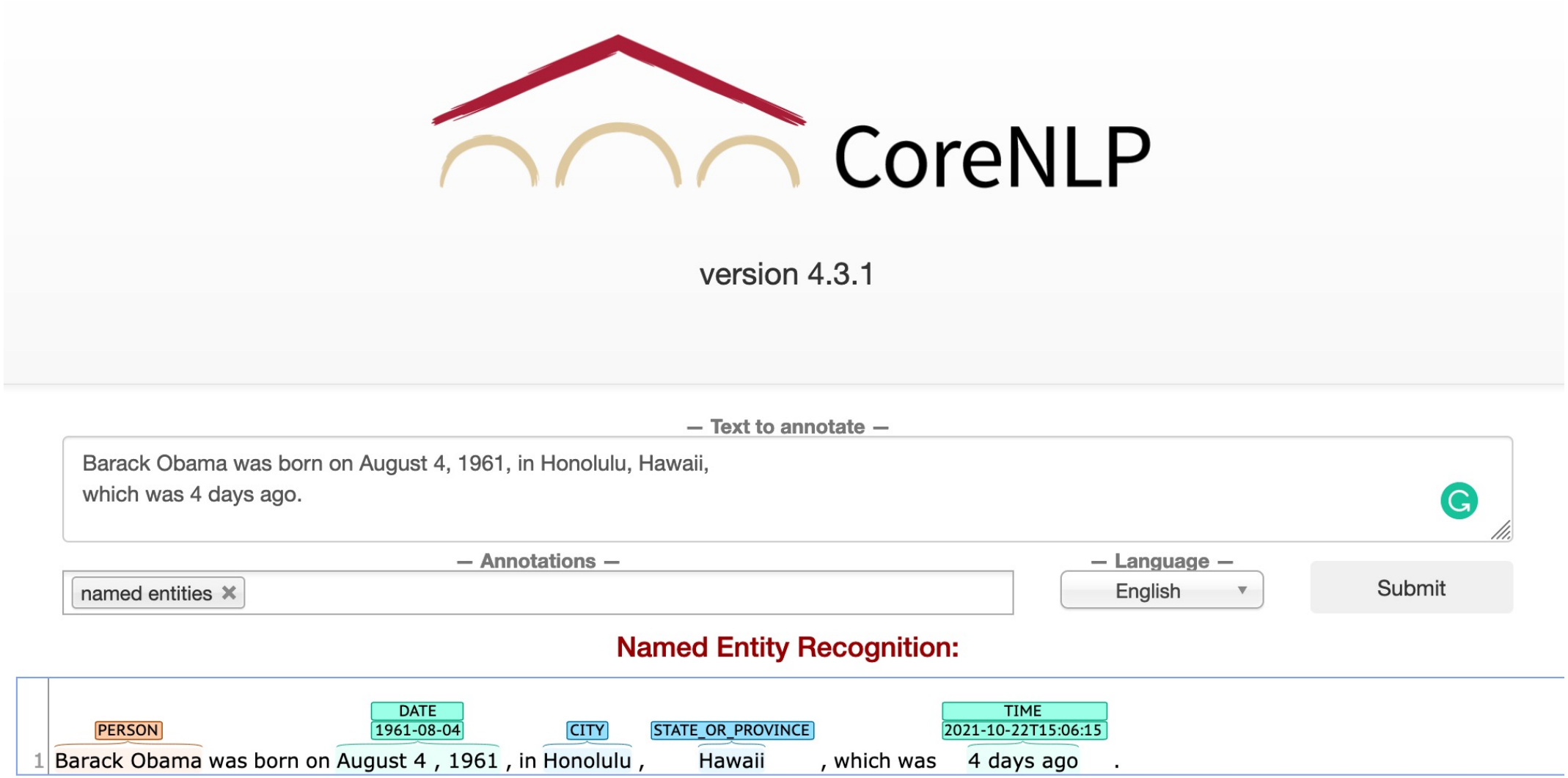
- In information extraction, a named entity is a real-world object, such as people, locations, organizations, or products, that can be denoted with a proper name.
- It can be abstract (company) or have a physical existence (person).
- It may also include time data, such as dates.
  - As an example, consider the sentence, “Washington was a president of the United States.” Both “Washington” and the “United States” are named entities since they refer to specific objects (George Washington and the United States).
- However, “president” is not a named entity since it can refer to many different items in different worlds (in different presidential periods referring to different persons, or even in different countries or organizations referring to different people).
- Rigid designators usually include proper names as well as certain natural terms, like biological species and substances.

# Common Approaches to Extracting Named Entities

- The Stanford Named Entity Recognizer is an excellent example of a NER.
  - The Stanford NER is implemented in Java.
  - It provides a default trained model for recognizing entities like organizations, people, and locations.
  - It also makes available models trained for different languages and circumstances.
- The Stanford NER is referred to as a CRF (Conditional Random Field) classifier.
  - Conditional Random Field sequence models have been implemented in the software.
- Custom models can be trained with a Stanford NER labeled dataset for various applications with proper instruction.

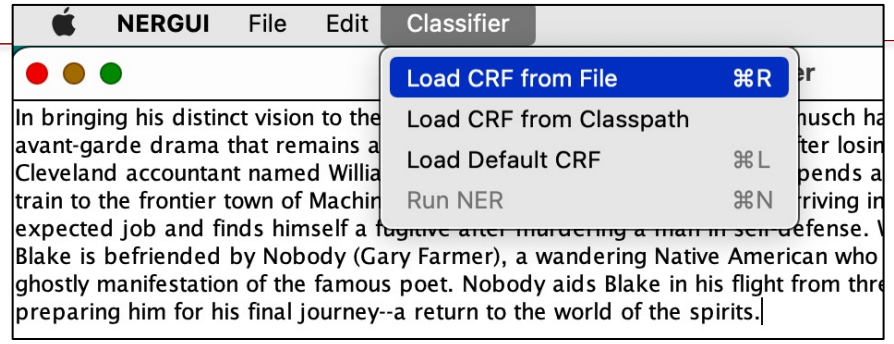



- http://nlp.stanford.edu:8080/ner/process

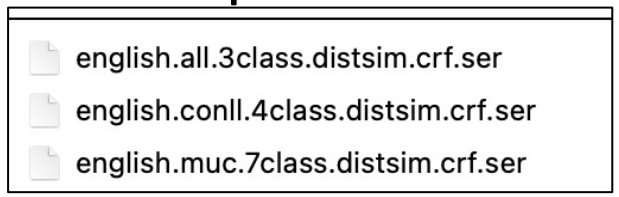



- http://corenlp.run/

# Download and Install Stanford NER




- Download and install software from:
  - https://nlp.stanford.edu/software/CRF-NER.html#Download
- Run *stanford-ner.jar*
  - Go to *Classifier* tab and note the directory that is the Java path to fin the classifiers
  - Copy the 3 *.ser* files to that directory




- Load the .7 classifier and analyze the following sentence to make sure it is working
  - Barack Obama was born on August 4, 1961 in Honolulu, Hawaii, which was 4 days ago.

# Using the Stanford NER