Overview: https://en.wikipedia.org/wiki/Natural-language_processing

Natural Language Processing (NLP) and the NLTK: See nlp.stanford.edu

1. Part-of-Speech Identification and Tagging
   - Penn Treebank P.O.S. Tags.htm
   - stanford-postagger-3.7.0.jar

2. Lemmatization, Parsing and Parse Trees
   - https://www.twinword.com/api/lemmatizer.php (JSON output)
   - lexparser-gui.sh / ~.bat

3. Named Entity Recognition (NER)
   - stanford-ner-3.7.0.jar
   - Word Find&Replace: /LOCATION>*<LOCATION> => ^p ; remove <, >
     nlp.stanford.edu/software/CRF-NER.html

4. Topic Recognition / Topic Modeling
   - TopicModelingTool.jar
     Based on MALLET (Java): mallet.cs.umass.edu

5. Stylometry
   - LEXOS – Comparative Stylometry: Dendrogram + PCA
     - lexos.wheatoncollege.edu

Other tools and resources:

Voyant Tools: voyant-tools.org

word2vec (Python): radimrehurek.com/gensim/models/word2vec.html

Python NLTK script: in “Python (basic example)” nltk_parse.py
Syntax

Part-of-speech tagging

Given a sentence, determine the part of speech for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb or adjective; and "out" can be any of at least five different parts of speech. Some languages have more such ambiguity than others. Languages with little inflectional morphology, such as English, are particularly prone to such ambiguity. Chinese is prone to such ambiguity because it is a tonal language during verbalization. Such inflection is not readily conveyed via the entities employed within the orthography to convey intended meaning.

Parsing (see also Stochastic grammar)

Determine the parse tree (grammatical analysis) of a given sentence. The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses. In fact, perhaps surprisingly, for a typical sentence there may be thousands of potential parses (most of which will seem completely nonsensical to a human). There are two primary types of parsing, Dependency Parsing and Constituency Parsing. Dependency Parsing focuses on the relationships between words in a sentence (marking things like Primary Objects and predicates), whereas Constituency Parsing focuses on building out the Parse Tree using a Probabilistic Context-Free Grammar (PCFG).

Semantics

Named entity recognition (NER)

Given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, organization). Note that, although capitalization can aid in recognizing named entities in languages such as English, this information cannot aid in determining the type of named entity, and in any case is often inaccurate or insufficient. For example, the first word of a sentence is also capitalized, and named entities often span several words, only some of which are capitalized. Furthermore, many other languages in non-Western scripts (e.g. Chinese or Arabic) do not have any capitalization at all, and even languages with capitalization may not consistently use it to distinguish names. For example, German capitalizes all nouns, regardless of whether they refer to names, and French and Spanish do not capitalize names that serve as adjectives.

OCR

Sentiment Analysis

Topic recognition / Topic Modeling

Stylometry

From: https://en.wikipedia.org/wiki/Natural-language_processing