**Glossary:**

**Unit 1:**

Natural Language:

Artificial Language:

NLU: Natural Language Understanding is the computers attempt to produce a useful representation of some inputted language

NLG: Natural language generation is the creation of natural seeming language from structured data.

Word2Vec (Mikolov 2013): Each word is embedded in to word vector that represents multiple dimensions.

**Unit 2:**

Lexical Analysis: Morphological and stemming of the data into a lexicon to use for analysis.

Syntactical Analysis: A word level analysis such a POS tagging

Semantic Analysis: A word meaning analysis indicating the intent of a word

Discourse Analysis: The understanding of the intended meaning.

Stemming: A lexical analysis tool to reduce terms down to there root or base word, but that word does not always represent a correct word.

Collocations: Counts of where words or letters occur next to each either.

WordNet: The NLP bible of real-world word senses

POS Tagging: After breaking a text into a sentence we can tag the various parts of speech from the text.

Lemmatization: Convert terms down to there lemma or base form which goes beyond just stemming.

Named Entity Extraction: The identification and extraction of people, places, or things from a given text.

Relationship Extraction: Identification of the relationship between 2 or more different named entities.

Word Sense Disambiguation: Use of context words to clarify the intended meaning of the word with respect to the other terms.

**Unit 3:**

Shallow NLP: Chunking of the of sentences or structures that useful for understanding text at a high level.

Deep NLP: The understanding of every word and what it represents with respect to the context of everything else in the sentence or document.

Statistical NLP: Use of a combination of signals and nodes of varying strengths to develop a final statistical score.

Symbolic NLP: Rule based systems where tests are performed for applying each rule.

Feature Learning: The use of machine techniques to learn the features to feed into a model.

Feature Engineering: The hand crafting of the features or variables that are going to be used for the model.

Top-Down:

Bottom-Up:

Transparent NLP: These are interpretable NLP tasks meaning a user is able to see how a process is working or what features may be driving a prediction.

Opaque NLP: These are “black-box” NLP tools where the actual features or processes that drive the NLP result are unknown.

RNN

**Unit 4:**

NLP Engineer:

Knowledge Engineer:

Data Scientist:

Database Administrator: Utilize NLP commonly for ETL tasks using NLU or generating reports using NLG.

Associative Processing Unit (APU):

**Unit 5:**

LSTM

LSTM Peephole

LSTM Coupled

LSTM GRU

Text Pre-Processing

Sentence Segmentation

Word Tokenization

Stop Words

Stemmers

TF-IDF

Lexical Diversity

Reading Level

n-grams

**Unit 6:**

Lexical Knowledge Base: This is the detailed understanding of the structure in relationships of a lexicon for using NLP applications.

Lexicon: In NLP this goes beyond just dictionary that makes it machine readable and useful to determining major NLP functions

Hyponym: A specific subcatergory of an entity or holonym. (i.e Cockerspanial is the hyponym of of dog)

Hypernym: The more generic group made up multiple hyponym. (i.e dog is the hypernym of cocker spanial)

Holonom: The whole of an object made up of various (i.e Car is holonom of car wheels)

Meronym: The part of a larger object that exist both separately and part of that object. (i.e car wheels are meronyms of car wheels)

WordNet: This is a semantic network that lets you measure some of the relationships between various words.

Ontological Distance: The distance in lexical space between words that provide a measurement of the similarity between two words.

Polysemy: Words with multiple senses or mean more than 1 thing.

Monosemy: Words with only one word sense or only mean 1 thing.

**Unit 7:**

Parts-of-Speech (POS) Tags: Tags the words in sentence or document with the intended usage of the word.

Tag Set: Set used for tagging all of the words to be used in the POS tagger.

Penn Treebank Tagset: Most commonly used tagset with 45 different tagsets.

Brown Corpus: A freely available reference corpus used in NLP

Rule-Based POS Tagger: Simple rules based approach. Begins with assigning the most common POS tag. Then assign simple rules that correct the tag based on the words immediately preceding or following that word.

Hidden Markov Model: A probabilistic path given previous POS it can determine the most probable POS given the weights of the prior POS.

Troponyms: The more specific subset of verbs of a large similar verb class. Like some troponyms of speaking are things like blubbering, whispering, etc.

Named Entity Extraction: Identification of proper names of a specific entity based. This can be rule based by recognizing patterns like consecutive NNP as first name last name or separated by commas.

Conjunctive Elimination: Strategy for NE identification that would eliminate named entities if and only if the NNP is not in some canonical list like Standard and Poors and has a low frequency.

**Unit 8:**

Shallow Parsing: A middle level of parsing or chunking that combines several POS tags into phrases, but does not parse the phrases themselves to show how the chunks relate.

Full Parse Tree: Takes all the chunks from shallow parse and show how all the chunks and POS tags relate to one another to visualize the entire grammatical structure.

Shallow Part Tree: The visualization of the shallow parsing for the phrases of the sentence, but does not break it down further like the full parse tree. This makes it easier to read.

Constituency Parser: Breaks sentence in NP, VP, or PP and then continues to parse those phrases down further.

Dependency Parser: Identifies a key verb and then chunks based on the subject and objects around that verb and completes parsing from those chunks.

Directed Acyclic graph: For parse tree this is a graph of nodes connected by edges with a directionality that explains the relationships & structure.

CYK algorithm: Parse tree algorithms that uses increasingly large windows to match grammar rules to achieve full parse of a sentence.