1. Unit 1
   1. Natural Language
      1. Natural languages are those that evolved or emerged gradually over time, largely unconsciously.
      2. English, Spanish, sign language
   2. Artificial Language
      1. Artificial languages are those that were designed, crafted, or invented with conscious purpose, largely all at once and not gradually.
      2. Klingon, Elvish
   3. NLP: Natural Language Processing
      1. NLP refers to the development and use of machine-based methods to process content in natural language.
   4. NLU: Natural Language Understanding
      1. NLU refers to the branch of NLP that uses computer software to understand input made in the form of natural language.
   5. NLG: Natural Language Generation
      1. NLG refers to the branch of NLP that uses computer software to generate usable, natural language output that is not just identical to its input.
2. Unit 2
   1. Lexical Analysis
      1. Fundamental foundation of all analyses
      2. Vocabulary or words like analysis. Inflection of plural v singular.
      3. Analysis goes up and down through these levels to help correct errors and fill-in gaps for all other levels of analysis.
   2. Syntactic analysis
      1. Grammar like analysis. Putting full sentences through a program that can draw structure of grammar.
      2. Analysis goes up and down through these levels to help correct errors and fill-in gaps for all other levels of analysis.
      3. Second level of NLP that focuses on grammatical analysis of documents
   3. Semantic analysis
      1. Entire meanings based on usage in a full like sentence way. 90% of data science projects stop here.
      2. Analysis goes up and down through these levels to help correct errors and fill-in gaps for all other levels of analysis.
   4. Discourse / entailment analysis
      1. Entire meanings that can be inferred based on another semantic meaning
      2. Drawing inferences is advanced NLP and what everyone is working towards
      3. Nuances of human conversation. So much is metaphorical, ironic, comical, etc.
   5. Lexicon
      1. Machine readable dictionary. Key vocabulary words for the program to complete lexical analysis
   6. Morphology
      1. Study of Morphemes
   7. Morphemes
      1. The units that our words are made of
   8. Token
      1. Word – sequence of characters
   9. Tokenization
      1. Covert a string of characters into a sequence of tokens
   10. Separator
       1. Tabs, spaces
       2. Punctuations, hyphens, apostrophes
   11. Stemming
       1. Type of discourse analysis
   12. Stemmer
       1. Making an algorithm that when inputted “running” automatically gives you the root morpheme “run” is called a stemmer
   13. Corpus derived meta-data
       1. Looking up words in our established vocabulary in order to gather together word clouds in order to help with things like:
          1. Word frequency score
          2. Common collocations
          3. Common co-occurring words, context words, etc.
   14. Collocations
       1. Words commonly occurring together that take on a different meaning as a unit when used together
   15. Head words
       1. Words that contain an entry in the lexicon
   16. Polysemous word
       1. Many words have many meanings
   17. WordNet
       1. Lexicon for NLP practices
   18. Apocryphal
       1. Doubtful authenticity
   19. Domain relations
       1. When a word can have more than one meaning, a certain meaning can often be tied to a specific domain, domain relations is a word’s relationship with a specific domain for its meaning.
   20. Spell correction
       1. Using lexicons to comb through documents to fix errors
   21. Terminology extraction
       1. Extracting key terms across a collection of documents
   22. Lexical diversity measurement
       1. The measurement of the word diversity contained within a document
   23. Sentence boundaries
       1. Things like . and ? etc. we cannot search for just those things when dealing with grammatical analysis. Because . can also be used for other things.
       2. Getting the document broken into sentences and defining sentence boundaries can become difficult.
   24. POS Tagging
       1. Part of speech tagging:
   25. Grammar Parsing
       1. Being able to pull a full sentence and determine the noun, verb phrase (VP), noun, preposition….etc breaking it into parts to break it into individual words
   26. Parse trees
       1. Trees showing the full sentence breaking it into its different pieces
       2. This can become complicated because not all people are capable of perfect grammar for the computer program.
   27. Lemmatization
       1. A lemma is the canonical (conventional) form that represents a set of related words forms eg. For run, runs, ran, running, the lemma is “run”.
       2. Lemmatizer: must be built to do things better and more advanced than a stemmer tool
   28. Inflections
       1. Syntactic differences between words
   29. Spell correction
       1. Offline:
          1. Indexing tokens: building the index for candidate generation
          2. Build a language model: model that estimates the a priori probability for an intended query
          3. Bundling an error model: edits – edit distance
       2. Query Time:
          1. Candidate generation: identifying the spelling correction for the query
          2. Scoring: computing the score or probability for each candidate
          3. Presenting suggestions: determining whether and how to present a spelling correction.
   30. Discrete Text Field Analysis
       1. Special form of syntax analysis
       2. Used when things like the web are your based documentation
   31. Unitizing
       1. Breaking discrete data points into their units that have previously been put together as one
   32. Normalizing
       1. Getting consistent nomenclature and units of measure and how information is conveyed.
   33. “Smart ETL”
       1. Extract, transform, load
       2. The actions with the data for grabbing, normalizing, and loading into the desired database
   34. Discrete text field analysis
       1. Are fields that are crammed into a single field and need to be parsed out by unitizing it
   35. Word Sense Disambiguation
       1. Is the problem of determining which "sense" (meaning) of a word is activated by the use of the word in a particular context, a process which appears to be largely unconscious in people.
   36. Classification
       1. The ability o use a tree structured graph to place documents into categories
       2. Often it is employed in machine learning models, such as SVM: Support Vector Machines
   37. Taxonomy
       1. A categorization of everything that exists in a particular field or domain
       2. It is difficult to have a good taxonomy because many documents live in many places simultaneous
       3. Most taxonomies must be cleaned up before they can be used.
   38. Semantic Task: Tags
       1. Names that identify the grouping of words in order to determine what a document is about.
   39. Semantic Task: Topic segmentation
       1. To find the junctures in the document as we analyze top to bottom when there is a topic shift.
   40. Semantic Task: Sentiment analysis
       1. The ability to determine if things are good / bad as far as what is being communicated in the document.
   41. Discourse anlaysis
       1. Small minority of NLP handle this level of analysis
       2. It is increasing more and more
       3. NLP can only do this if the previous analysis are near perfectly executed in the project and as an NLP researcher in general
   42. Anaphora resolution
       1. Type of discourse analysis
       2. When the NLP needs to refer to a previous text
       3. Pronouns
   43. Discourse Modeling
       1. Type of discourse analysis
   44. Question answering
       1. Type of discourse analysis
   45. Textual entailment
       1. Type of discourse analysis
   46. Pragmatic analysis
       1. Type of discourse analysis
       2. Practical reason that allows human being to make inferences around what is strictly explicated in what they read.
   47. Anaphora Resolution
       1. the problem of resolving what a pronoun, or a noun phrase refers to. In the following example, 1) and 2) are utterances; and together, they form a discourse.
   48. Discourse Model
       1. A discourse model is a mental object that constitutes an. individual's knowledge of a discourse. It is constructed on the basis of what has occurred in the discourse supplemented by general and specific knowledge.
   49. Textual entailment
       1. in natural language processing is a directional relation between text fragments. The relation holds whenever the truth of one text fragment follows from another text. In the TE framework, the entailing and entailed texts are termed text (t) and hypothesis (h), respectively.
   50. Pragmatic analysis
       1. refers to a set of linguistic and logical tools with which analysts develop systematic accounts of discursive political interactions.
   51. Table

       Description automatically generated
3. Unit 3
   1. Shallow v Deep
      1. Building a robust, exhaustive representation of text meaning, or pick out a few elements needed for the application
   2. Statistical v Symbolic
      1. Leverage powerful, complex statistical methods or Use rules that can be operated on by strict logic
   3. Top Down v Bottom Up
      1. Top down starting with high level categorization v The reverse
   4. Transparent v Opaque
      1. Engineer a system that is explainable to an SME v Build that only data scientists/engineers can understand
   5. Deep Learning
      1. Systems that have 3 or more hidden layers is considered Deep Learning
   6. NP Noun phrase
   7. Verb Phrase
   8. Preposition phrase
   9. Deep parsing
      1. Breaking sentence into individual words
   10. Shallow parsing
       1. Only breaking sentences into chunks
   11. Deep semantics
       1. Elements from each
   12. Shallow semantics
       1. Meaning without deep dive exact meanings
   13. Symbolic Approach
       1. Rule based approach to NLP. Resembles math formulas when written out. Powerful and easy for people to understand
       2. Manual maintanence
   14. Statistical Approach
       1. Uses an algorithm to find words, sentences, documents that are the most similar to each other
   15. LSA/LSI: Latent Semantic Analysis/Latent Semantic Indexing
       1. Building a similar words index using statistical method
       2. Nurse was found to be closer to the word Doctor than Physician because the times when those words were used together were more statistically common
   16. Feature Learning
       1. Involving a human expert to engineer features
   17. Feature Engineering
       1. Not using a human for feature but solely a training set
4. Unit 4
5. Unit 5
6. Unit 6
7. Unit 7
8. Unit 8
9. Unit 9: Midterm
10. Unit 10
11. Unit 11
12. Unit 12
13. Unit 13
14. Unit 14
15. Unit 15: Final