Natural Language Processing

1. Introduction to NLP
    - History
    - Motivation for NLP
        - How simple programs can help you manipulate and analyze language data, and how to write these programs
        - How key concepts from NLP and linguistics are used to describe and analyze language
        - How data structures and algorithms are used in NLP
        - How language data is stored in standard formats
        - How data can be used to evaluate the performance of NLP techniques
    - Look at how ML helps NLP
        - What are Automatic Learning Procedures
        - Look at Computational Linguistics
    - Look at how Statistics helps NLP
        - Look at Stochastic, Probabilistic, and Statistical methods
        - Look at Markov models
    - Brief on important tasks of NLP
        - Automatic summarization
        - Coreference resolution
        - Language identification
        - Machine translation (language translation)
        - Named entity recognition (NER)
        - Part-of-speech tagging
        - Parsing
        - Relationship extraction
        - Natural language generation and understanding
        - Optical character recognition (OCR)
        - Question answering

        - Sentiment analysis
        - Speech recognition
        - Topic segmentation and recognition
        - Word sense disambiguation
        - Stemming
        - Text simplification
        - Text-to-speech
        - Text-proofing
        - Natural language search
    - Limitations of NLP
2. Introduction to NLTK
    - Installation
    - Language Processing
        - Accessing pre-built books
        - Searching text using concordance
        - Word Sense Disambiguation
    - Simple Statistics
        - Dispersion plots
        - Frequency distribution
        - Look at Collocations and Bigrams
    - Corpus
        - Accessing Text Corpora
        - Look at famous Corpora
            - Gutenberg Corpus
            - Brown Corpus
            - Web Chat Corpus
        - Text Corpus Structure
            - Isolated
            - Categorized
            - Overlapping
            - Temporal

        - Loading your own Corpus
    - LAB: Conditional Frequency Distributions
        - Conditions and Events
        - Counting Words by Genre
        - Plotting and Tabulating Distributions
    - Lexical Resources
        - Look at Wordlist Corpora
        - What are stopwords?
        - Look at Comparative Wordlists
    - WordNet
        - Look at WordNet Hierarchy
        - Lexical Relations
        - Semantic Similarity
    - The NLP Pipeline
        - Capture User Input
        - Tokenize text and select words of interest
        - Normalize words and build the vocabulary
3. Text Manipulation
    - Capturing User Input
        - Electronic Books
        - Dealing with HTML
        - Processing RSS Feeds
        - Reading Local Files
        - Extracting Text from PDF, MSWord and other Binary Formats
    - String Manipulation
        - Dealing with Unicode text
        - Regular Expression
            - Regular Expressions for Detecting Word Patterns
        - Accessing Individual Characters
        - Accessing Substrings

        - Tokenization
            - Using Regex
            - Using NLTK
        - Normalizing Text
            - Stemming
            - Lemmatization
        - Segmentation
            - Sentence Segmentation
            - Word Segmentation
        - N-Grams
    - LAB: Convert English Text to Pig Latin
        - Definition of Pig Latin Each word of the text is converted as follows: move any consonant (or consonant cluster) that appears at the start of the word to the end, then append ay, e.g. string → ingstray, idle → idleay. http://en.wikipedia.org/wiki/Pig_Latin
        - Write a function to convert a word to Pig Latin.
        - Write code that converts text, instead of individual words.
        - Extend it further to preserve capitalization, to keep qu together (i.e. so that quiet becomes ietquay), and to detect when y is used as a consonant (e.g. yellow) vs a vowel (e.g. style).
4. Categorizing and Tagging Words
    - POS Tagger
    - Standard Part-of-Speech Tagset
    - Look at the English Language Structure
        - ADJ - adjective (new, good, high, special, big, local)
        - ADV - adverb (really, already, still, early, now)
        - CNJ - conjunction (and, or, but, if, while, although)
        - DET - determiner (the, a, some, most, every, no)
        - EX - existential (there, there's)

        - FW - foreign (word dolce, ersatz, esprit, quo, maitre)
        - MOD - modal (verb will, can, would, may, must, should)
        - N - noun (year, home, costs, time, education)
        - NP - proper (noun Alison, Africa, April, Washington)
        - NUM - number (twenty-four, fourth, 1991, 14:24)
        - PRO - pronoun (he, their, her, its, my, I, us)
        - P - preposition (on, of, at, with, by, into, under)
        - TO - the (word to to)
        - UH - interjection (ah, bang, ha, whee, hmpf, oops)
        - V - verb (is, has, get, do, make, see, run)
        - VD - past (tense said, took, told, made, asked)
        - VG - present (participle making, going, playing, working)
        - VN - past (participle given, taken, begun, sung)
        - WH - wh (determiner who, which, when, what, where, how)
    - Defining Dictionaries
        - Default Dictionaries
        - Updating Dictionaries
        - Inverting Dictionaries
    - Tagging
        - The Default Tagger
        - The Regular Expression Tagger
        - The Lookup Tagger
        - N-Gram Tagging
            - Unigram Tagging
            - General N-Gram Tagging
        - Combining Taggers
        - Storing Taggers
    - Determine the Category of a Word
        - Morphological Clues
        - Syntactic Clues
        - Semantic Clues

        - New Words
    - LAB: Process the Brown Corpus to find answers to the following questions,
        - Which nouns are more common in their plural form, rather than their singular form? (Only consider regular plurals, formed with the -s suffix.)
        - Which word has the greatest number of distinct tags. What are they, and what do they represent?
        - List tags in order of decreasing frequency. What do the 20 most frequent tags represent?
        - Which tags are nouns most commonly found after? What do these tags represent?
5. Classifying Text
    - Supervised Classification
        - Gender Identification
        - Document Classification
        - Part-of-Speech Tagging
        - Sequence Classification
        - Sentence Segmentation
        - Identifying Dialogue Act Types
        - Recognizing Textual Entailment
    - Evaluating Models
        - Accuracy
        - Precision and Recall
            - True positives (relevant items that we correctly identified as relevant)
            - True negatives (irrelevant items that we correctly identified as irrelevant)
            - False positives (irrelevant items that we incorrectly identified as relevant)
            - False negatives (relevant items that we incorrectly

identified as irrelevant)
        - Confusion Matrices
        - Cross-Validation
    - Decision Trees
        - Definition
        - Naive Bayes Classifiers
        - Zero Counts and Smoothing
        - Non-Binary Features
        - The Naivete of Independence
        - The Cause of Double-Counting
6. Information Extraction
    - Architecture
        - Sentence Segmentation
        - Tokenization
        - POS Tagging (Parts of Speech)
        - Entity Detection
        - Relation Detection
    - Chunking
        - Noun Phrase Chunking
        - Tagging Patterns
        - Chunking with Regular Expressions
        - Chinking (Exclusion)
    - Representing Chunks: Tags vs Trees
    - A look at NLTK Trees
    - Named Entity Recognition
    - Relation Extraction
    - Analyzing Sentence Structure
        - A look grammar representation in NLTK
            - S - sentence (the man walked)
            - NP - noun phrase (a dog)
            - VP - verb phrase (saw a park)

        - PP - prepositional phrase (with a telescope)
        - Det - determiner(the)
        - N - noun (dog)
        - V - verb (walked)
        - P - preposition (in)
    - Parsing with grammar
    - Writing your own grammars
- Parsing With Context Free Grammar
    - Recursive Descent Parsing
    - Shift-Reduce Parsing
    - The Left-Corner Parser
    - Substring Tables
- LAB: Write a tag pattern to identify places of work from a set of resumes by building your own grammar.