Web Tool for Phonemes Week-2

University at Buffalo The State University of New York

Agenda

- Educator word frequency guide
- WordNet for categories
- Solr Demo
- Klatt and IPA
- Iphod.com
- Word2Vec and Doc2Vec
- Manner of Articulation.
- Paper

Educator Word Frequency Guide

- Study of word frequency, and the number of words analyzed (over 17 million tokens and 164,000 types)

- The guide is organized into four sections:
  1. Technical aspects.
  2. Words with frequencies of 1 or greater, including additional statistics by grade level.
  3. The third section covers words with frequencies less than 1,
  4. It presents all words from the corpus in descending order of frequency.

Word Categorization

- NLTK CMU dictionary provides words and the phoneme information
- NLTK also contains wordnet that contains many words, semantics, hypernym and hyponyms and many more
- Wordnet has Hypernym(broad meaning)
- Took words from CMU to get the phonemes information and used hypernym to get the category of the data
- Other data sources: https://www.dictionaryapi.dev/

Solr

- Used the data provided in NLTK and wordnet and classified the words
- Wrote some Scripts to create schema in solr and create cores
- Indexed the data to the solr core
- Sample Query

  Issues with query and data

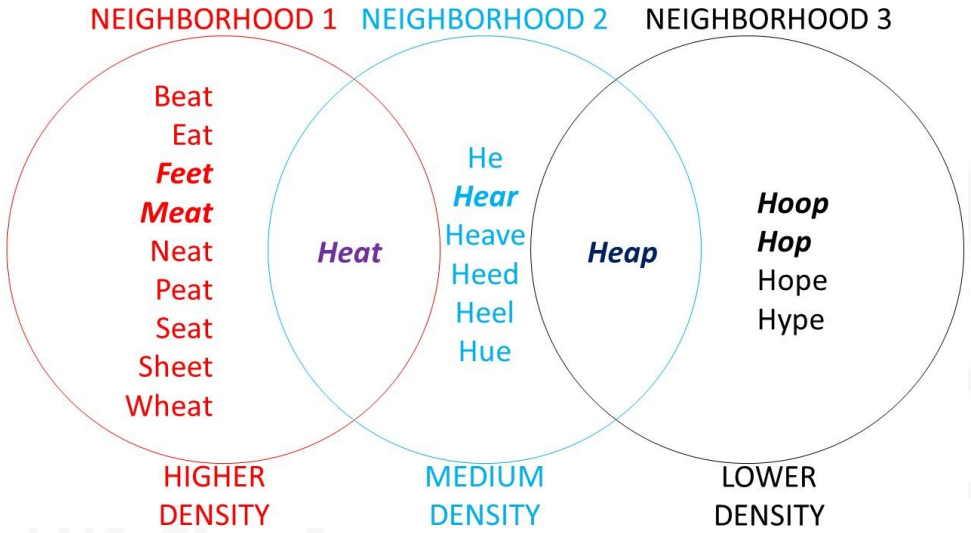
  Orange, oranges, tiger, tigers → we intent to use stemming and lemmatization in order to find the root word and make them unique.

Klatt and IPA

- Klatt is a speech synthesis system.
- It was named after Dr. Dennis Klatt
- System designed to model and produce speech sounds based on acoustic parameters.
- Represents speech sounds
- Allows the generation of artificial speech from text or phonetic input.

- IPA is a standardized system of symbols used to represent the sounds (phonemes) of human language.
- A transcription system designed to accurately depict the speech sounds found in languages.

IPHOD:



- It provides the following:

1. word frequency
2. phonotactic probability:
3. neighborhood density : https://calculator.ku.edu/density/English/words

Word2Vec

- Vector representations of the words
- Semantic similarity of the words is represented between their vectors
- Based on cosine similarity of word
- Classification and information retrieval
- dataset = [

  ['P', 'AH0', 'R', 'K'], # PARK

  ['K', 'AH0', 'R'], # CAR

  #]

https://radimrehurek.com/gensim/models/word2vec.html

can use it to analyze the phonemes in various ways, such as finding **similar phonemes**, phoneme analogy tasks, or visualizing phoneme vectors.

Doc2vec

Doc2Vec extends the Word2Vec algorithm by generating vector representations for entire documents instead of individual words.

Paper on Minimal, Maximal and Multiple Oppositions

- Minimal pairs are pairs that differ in one phoneme
- Maximal Oppositions are the words that have contrasting phonemes which differ in as many distinctive features as possible
- Features are divided in non major class and major class
- Multiple Opposition are sets of maximal opposition pairs
- https://www.speech-language-therapy.com/index.php?option=com_content&view=article&id=133&catid=9&Itemid=101

Articulation Manner

As the data used by the us is following IPA format we will be using the same articulation manners as of now.

If we find any other linguistic traditions that use different nomenclature we will add later to our SOLR