Natural Language Processing Toolkit Notes

17/02/2017

How NLTK get nouns/ verbs

- Steps:
  - Tokenize text
  - Tag words
  - Chunk sentences

- Tags for nouns and verbs:
  - NN noun, singular 'desk'
  - NNS noun plural 'desks'
  - NNP proper noun, singular 'Harrison'
  - NNPS proper noun, plural 'Americans‘
  - VB verb, base form take
  - VBD verb, past tense took
  - VBG verb, gerund/present participle taking
  - VBN verb, past participle taken
  - VBP verb, sing. present, non-3d take
  - VBZ verb, 3rd person sing. present takes

Chunk structure for a given sentence

- by regular expression
  - <RB.?>*<VB.?>*<NNP>+<NN>?
    - <RB.?>* = "0 or more of any tense of adverb," followed by:
    - <VB.?>* = "0 or more of any tense of verb," followed by:
    - <NNP>+ = "One or more proper nouns," followed by:
    - <NN>? = "zero or one singular noun. "
  - <DT>?<JJ>*<NN>
    - <DT>? = "optional determiner,“ followed by:
    - <JJ>* = "0 or more of any tense of adjective," followed by:
    - <NN>= " one singular noun.“

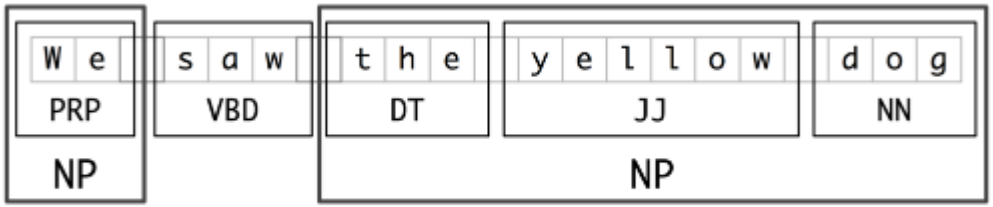
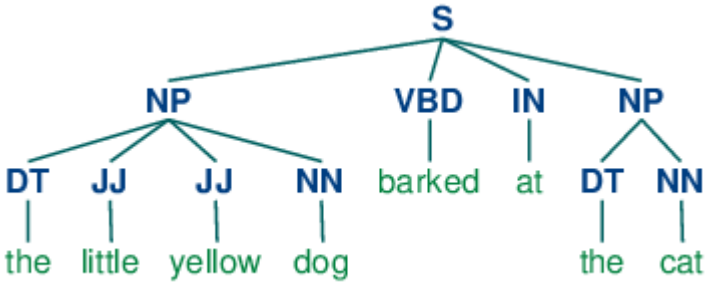


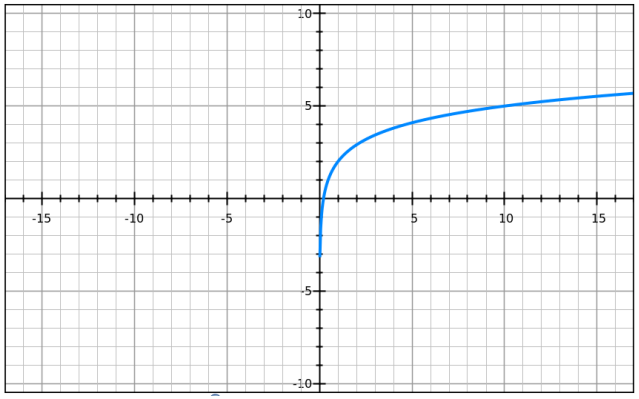
Figure 2.1: Segmentation and Labeling at both the Token and Chunk Levels



Filter by frequency

- Filter out infrequently nouns
  - Counts how many times a word shows up in text
  - Use threshold to cut infrequently words

- A guess about show up frequency of a character in text
  - Frequency might be a function of story length, frequency = *f*(story length)
  - *f*(story length) might be in logarithmic form
    - Story is longer, a character might show up more times
    - But, also when story is long enough, new characters will probably show up



Current Progress

- Current:
  - Get all nouns from text
  - Filter infrequent nouns

- Next:
  - Deal with pronoun
  - Combine nouns and verbs, <Noun, Verb>
  - Find more chunking ways