Date Subject

# Regular Expressions

- Formally, a regular expression is an algebraic notation of characterizing a set of strings. particularly used for searching in texts, when we have a pattern to search for and a corpus of texts to search through

- Before we take processing words, we need to decide what counts as a word??

- Corpus ((pl: corpora)): a computer readable collection of text or speech. it is the dataset we use for NLP

- token: Each individual occurance of a word in a text, including ponctuation is counted ((depending on context))

- type: unique words in corpus ((vocabulary size)) capitalization and punctuation may or may not count as seperate tokens depending on NLP tasks.

- Disfluency in speech:

- Disfluency: unplanned pause or corrections in speech.
  - broken-off words: main-
  - Fillers: uh, um

Date Subject

## Lemma vs wordforms:

- lemma: the base form of a word, representing all its inflected wordforms.
- wordform: a fully inflected or derived form
  eg: "cat" is lemma, "cats" is a word form.

## Vocabulary size:

Type (|V|) : size of corpus
Tokens (N) : total running words

## Harlans Law / Heaps Law

- observation: as a corpus gets larger, new word types keep appearing, but at a decreasing rate.
- Formula: $|V| = KN^{\beta}$

  $|V|$ = types, $|N|$ = tokens

  $K, \beta$ = constants

  $0 < \beta < 1$ usually 0.65 - 0.75

## Corpora and variations in language:

(I) context of words

(II) language varieties
- american mainstream english
- afro-american english

(III) code switching
- using more than one language in a stream. like hindi words in english

(IV) genre variation
- different genres, scientific, poetic, etc.

JAHAN

Date Text normalization Subject

- Tokenization: Spliting text into words or meaningful units.
- word normalization: Standardizing word forms.
- Sentence segmentation: Spliting text into sentences.

- word tokenization:
  - punctuation seperation:
    "Hello, world!" → ["Hello", ",", "word"]
  - clitic expression:
    "we're" → ["we", "are"]
  - Special cases:
    "PH.D." → ["PH.D"]
  - Numbers:
    "$45.55" → ["$45.55"]
- Subword Tokenization (BPE)

Date lexical Semantics Subject

## Lemma and Senses:

- Lemma; (citation form): the base dictionary form of a word
  - sing is lemma for sing, sang, sung

- Wordform: any specific inflected form of a lemma
  - sung, carpets, duermes

- Sence: a distinct meaning of a lemma
  - mouse = rodent, mouse = computer device

- lemmas can be polysemous (multiple senses), this will create interpretation problems.

- Word senses disambiguation (WSD):
  Task of deciding which sense applies in context

## Synonymy:

Two sences are synonyms if they can substitute for each other in any sentence without changing its truth conditions.

**No two words are truly identical**

- Differences in linguistic form correspond to some to some difference in meaning.

Date _____ Subject _____

## Word similarity

- not the same as synonymy:
  -- cat and dog are not synonyms but are similar

## Word Relatedness:

- words can be related without being similar. these words often have a semantic Relation with eachother.

## Semantic Frames and Roles:

- Semantic frame : a set of words representing roles in an event.

eg: Commercial transaction →
- verbs: buy, sell, pay
- noun: buyer
- Roles: buyers, sellers, money

## Connotation:

- emotional part of the meaning, positive or negative.
- words with similar meaning, can differ in connotation

eg: copy — copy seems positive/neutral
bake — seems negative.