Normalization

Normalization

- Morpheme : base form of a word
- Structure of token : <prefix> <morpheme> <suffix>

  Example : Antinationalist : Anti + national + ist
- Normalization : Process of converting a token into its base form (morpheme)
- Helpful in reducing data dimensionality, text cleaning
- Types : Stemming and Lemmatization

Normalization: Stemming

- Elementary rule based process of removal of inflectional forms from a token

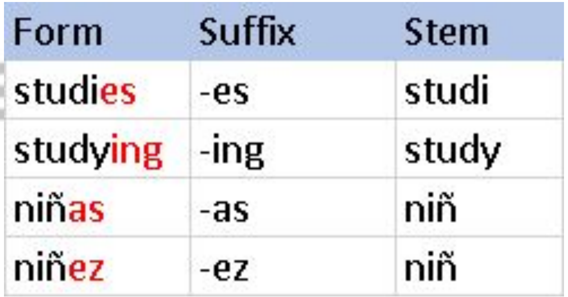
- Outputs the stem of a word

  “laughing”, “laughed”, “laughs”, “laugh” >>> “laugh”

- May generate non-meaningful terms

  his teams are not winning

  >> hi team are not winn



Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

Normalization: Lemmatization

- Systematic process for reducing a token to its lemma

- Makes use of vocabulary, and morphological analysis

- Example :

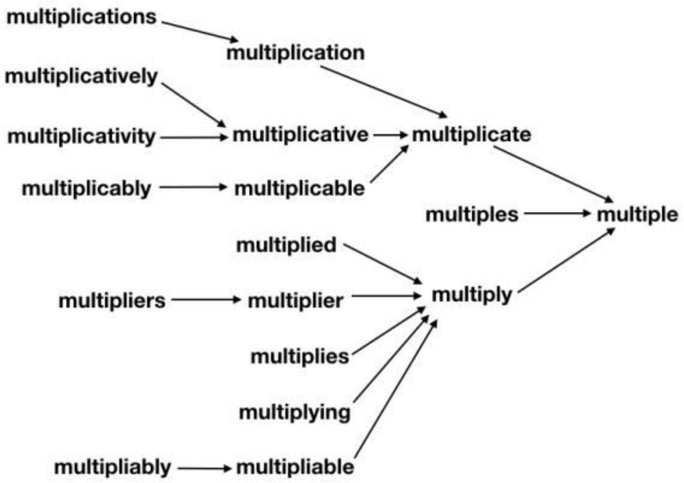
  am, are, is >> be

  running , ran , run , rans >> run

- Running, ‘verb’ >> run

  Running, ‘noun’>> running

- Slower than Stemming



Normalization: Lemmatization

Open the Jupyter Notebook

Thank You