Text Preprocessing

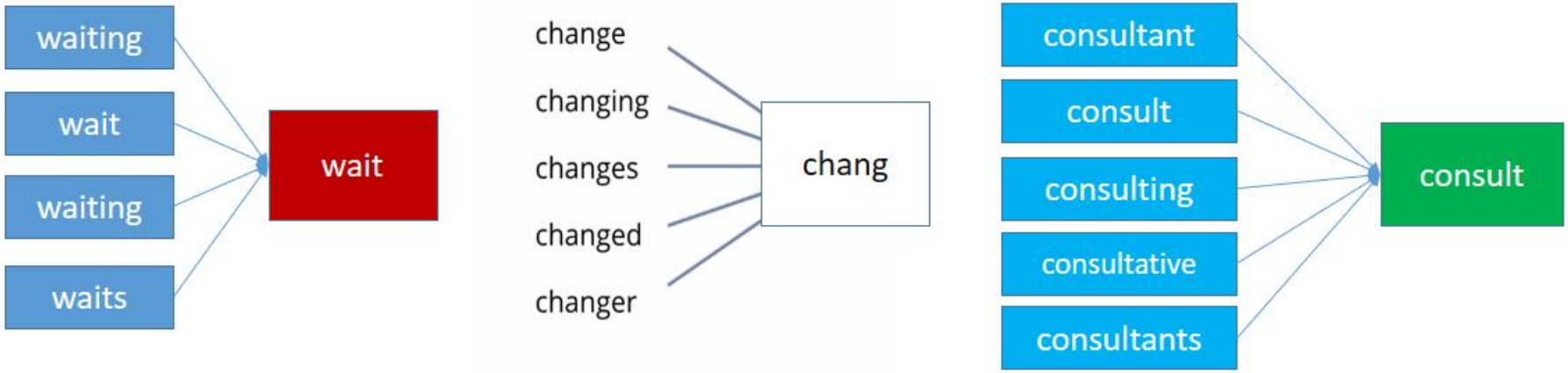
Stemming and Lemmatization in NLP

Stemming and Lemmatization are Text Normalization (or sometimes called Word Normalization) techniques in the field of Natural Language Processing that are used to prepare text, words, and documents for further processing. Stemming and Lemmatization have been studied, and algorithms have been developed in Computer Science since the 1960's. We will learn about Stemming and Lemmatization in a practical approach covering the background, some famous algorithms, applications of Stemming and Lemmatization, and how to stem and lemmatize words, sentences and documents using the Python **nltk package** which is the Natural Language Tool Kit package provided by Python for Natural Language Processing tasks.

Text Preprocessing

Stemming

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding (NLU) and natural language processing (NLP). When a new word is found, it can present new research opportunities. For example-



Text Preprocessing

Stemming

- **Porter Stemmer():** The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and in flexional endings from words in English.
- **Lovins Stemmer**
- **Dawson Stemmer**
- **Krovetz Stemmer**
- **Xerox Stemmer**
- **N-Gram Stemmer**
- **Snowball Stemmer**
- **Lancaster Stemmer**

Text Preprocessing

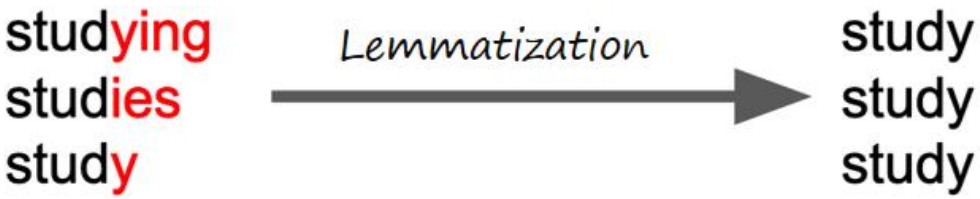
Lemmatization

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma .

- **Word Net Lemmatizer**
- **Spacy Lemmatizer**
- **TextBlob**
- **Gensim Lemmatizer**
- **TreeTagger**

Text Preprocessing

Lemmatization



	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

Text Preprocessing

Stemming vs Lemmatization