# NLP Basics

1. Feature Engineering
2. Language Processing

* Tokenization
* Lemmatization
* Stemming

---

## Tokenization

→ String breaking

A big dog. } Tokenization based on space.

≠ word based

tokenization } Sub word based tokenization

A b i g d o g : } letter based tokenization

Stemming (Idea is to have least # independent features)

organize, organizes, organizing | democracy democratic

how to deal with similar features!

## Lemmatization [same idea → to have min # features]

lexeme: am, are, is → be (lemma)

car, cars, car's, cars' ⇒ car (lemma)

## Stemming

reducing words to their root.

1. Porter Stemmer
2. Snow ball Stemmer

} part of NLTK

generate → generat

generation → generat

Stemming ⇓

but, generat doesn't have any meaning

but in lemmatization → they both becomes "generate" probably.