-> Word Tokenization: Words ≈ Token

-> I have a can opener, but I can't open the cans.

↳ Word Token: 11 words

↳ Word Type: Different realization of word.

10 word Types.

-> Issues w/ W. Tokenization:

-> Finland's -> Finland, Finlands, Finland's, Finlands'

-> What've, I'm, shouldn't -> What have
I am, Should not

-> San Franciso

-> m.p.h -> ✓

-> Handling Hyphenation:

-> End-of-line Hyphen: show-time, initia-tive

-> Lexical Hyphen:

Prefix: eg: co-, pre-, meta-, multi-

-> Sententially Determined Hyphenation:

-> Case-based, hand-delivered

-> three-to-five-year plan.

-> Normalization: U.S.A & USA

-> Case-Folding:

↳ Reduce all text to lower case.

↳ US (us)

↳ among us there is an impostor

↳ the us is a financial superpower.

↳ Exceptions (Task Dependent)

↳ MG, GM

-> Morphology: (Morphemes)

Uncondiontionelly

-> Lemmatization: Reduces inflections or variant forms to base forms.

-> car, cars, car's, cars' -> car

-> Morphemes are divided in 2 categories:

-> Stem: Core meaning bearing unit.

-> Affixes:

-> Affixes :

-> Prefix : un-, anti-, Sanskrit (a-, ati-, pra-)

-> Suffix : -ity , -ation

-> Infix : Abso-damn-lutely

Un-freakin-believable

-> Stemming : Crude chopping of affixes.

It is always language dependent.

-> Porter's Algorithm :