



AI Outsider

POWERED BY

Section 2

Text Normalization

Want more? Head over to www.AIOutsider.com





Section Overview



Text Mining & NLP
Sentiment Analysis
Google Colab
Dataset Overview
Descriptive Statistics

Text Normalization
Features Cleaning
Tokenization
Stemming
Lemmatization





AIOutsider.com

WHAT IS NORMALIZATION?



TEXT NORMALIZATION



Reduce randomness in a particular piece of text

Definition

NON-NORMALIZED



- « @AIOutsider soooo fun learning NLP!!! »
- « @AIOutsider I learn NLP. So fun! »
- « Learning NLP is so funny with @AIOutsider »

NORMALIZED



- « so fun learn nlp »
- « learn nlp so fun »
- « learn nlp so fun »





AIOutsider.com

TEXT CLEANING (1/2)



TEXT CLEANING: TWITTER SPECIFIC



Remove or replace all items that do not provide additional information

Definition

NON-NORMALIZED



« RT @AI0utsider: happy to learn!!! 🤖 https://AI0utsider.com #NLP »

1

Retweet

2

User Tag

3

Emojis

4

URLs

5

Hashtags

NORMALIZED



« twitter_user: happy to learn!!! cool_emoji NLP »





AIOutsider.com

TEXT CLEANING (2/2)



TEXT CLEANING: GENERAL



Remove or replace all the items that do not provide additional information

Definition

NON-NORMALIZED



« Soooo HAPPY ! Won't stop buzzing !!!! »

1

Repetition

2

Capitalization

3

Contraction

4

Repetition

NORMALIZED



« so happy! will not stop buzzing! »





AIOutsider.com

TOKENIZATION



HOW IMPORTANT IS TOKENIZATION



Tokenization is a way to separate text into smaller chunks

Definition



Can you understand this?

« vousparlezfrancais »



« you speak french »



you

speak

french

« vou spar lezfrancais »

« vous parl ezfrancais »

« vous parlez francais »



vous

parlez

français



:/> _ TOKENIZATION EXCEPTIONS



Punctuation

« hello! ~~!!!!!!!!!!!!~~ »



Stop words

« ~~I~~ am happy »



Numbers

« I won ~~\$50~~, so nice »

« almost won ~~\$1M~~, so angry »





AIOutsider.com

STEMMING



WHAT IS STEMMING



Stemming is the process of reducing words to their root form

Definition



Manag

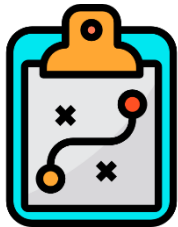
~~Manager~~

Manag

~~Management~~

Manag

~~Managing~~



Rule-based
Approach



Faster but
chops words



Meaning is
less important

► Over-stemming

~~universal~~

~~university~~

universe

► Under-stemming

~~alumnus~~

alumni

~~alumnae~~





AIOutsider.com

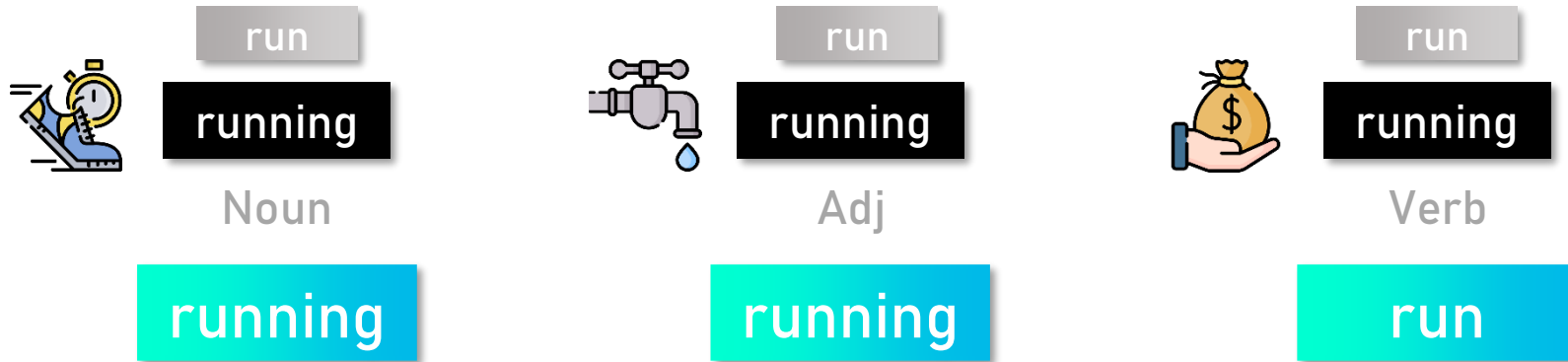
LEMMATIZATION




:> _ WHAT IS LEMMATIZATION




Lemmatization serves the same purpose as Stemming but makes use of word context Definition



Dictionary-based
Approach



Slower but
uses context



Meaning is
important



 FOLLOW US!



AIOutsider.com



twitter.com/AIOutsider_

All Trademarks referred to are the property of their respective owners.

© 2021 – All rights reserved.

Disclaimer

