Introduction to NLP
Chapter 2:
Text Preprocessing

Muhammad Okky Ibrohim, S.Mat., M.Kom.

Bisa.ai
PT. Bisa Artifisial Indonesia

Kampus Merdeka
INDONESIA JAYA

Credit

This slide was built using **Blue Connections Cordelia Presentation Template** theme (https://www.slidescarnival.com/cordelia-free-presentation-template/216).

Several images and other media from Google Images that maybe I forgot to add specific credit (if there is belong yours, please contact me if you want me to add specific credit to you or remove it).

Video course

For those who can not attend the synchronous session or want to relearn this material, you can watch this video for this slide presentation: https://youtu.be/DSjUpWnkdno

Outline

2.1 Why we need to preprocess our dataset?

2.2 Basic text preprocessing in text dataset

2.3 Hands-on: Basic text preprocessing using several NLP libraries (NLTK, Sastrawi, etc.) -> Will be presented using Notebook (Jupyter Notebook or Google Colab)

2.1 Why we need to preprocess our dataset? Definition and motivation

- Definition

  Text preprocessing is a process to make our dataset more “clean” so that more "ready" for feature extraction.

- Motivation
  - Almost all data in a real case is not clean and ready to use
  - To reduce memory needs and fasten the computation
  - To improve the model performance

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing

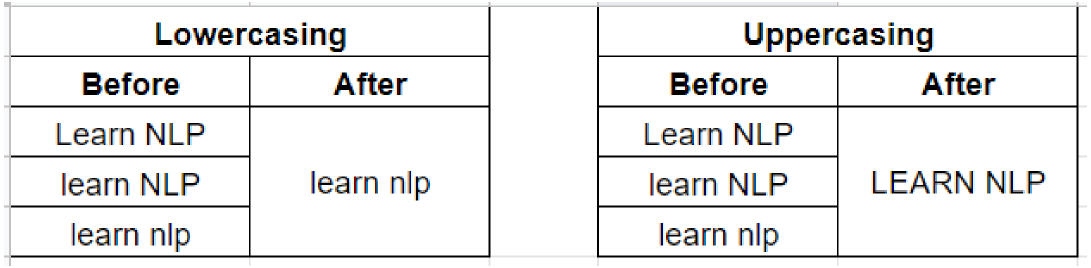
- Text Case Uniforming (Uppercasing OR Lowercasing)
- Emoji and Emoticon Normalization
- Contraction Normalization
- Digit Normalization
- Token Masking
- Text Cleaning
- Basic Text Normalization (Using Dictionary)
- Stop Words Removal
- Stemming and Lemmatization

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing:
Text case uniforming (uppercasing or lowercasing)

Definition

Text case uniforming is a preprocessing step to convert all characters into the same cases, whether to lowercase or uppercase (almost all NLP research is using lowercase, especially when using a pre-trained word embedding).

Example



Lowercasing	
Before	**After**
Learn NLP	learn nlp
learn NLP	
learn nlp	

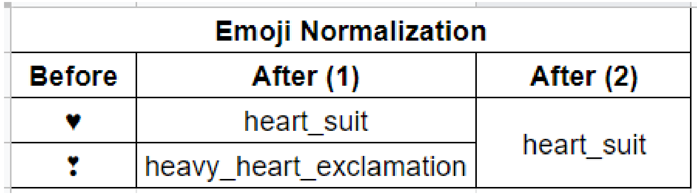
Uppercasing	
Before	**After**
Learn NLP	LEARN NLP
learn NLP	
learn nlp	

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing:
Emoji and emoticon normalization

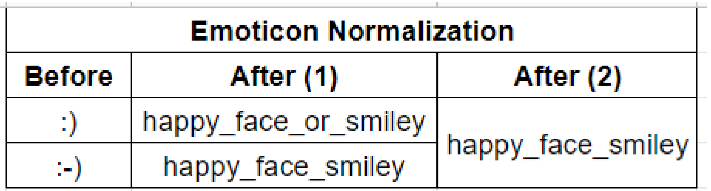
Definition

Emoji and emoticon normalization is a preprocessing step to convert emoji and emoticon into a “verbal” form.

Example



Emoji Normalization		
Before	**After (1)**	**After (2)**
♥	heart_suit	heart_suit
❣	heavy_heart_exclamation	



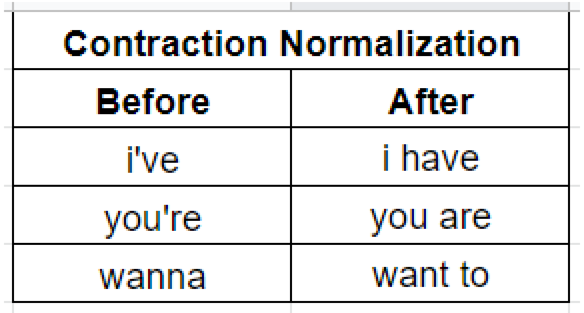
Emoticon Normalization		
Before	**After (1)**	**After (2)**
:)	happy_face_or_smiley	happy_face_smiley
:-)	happy_face_smiley	

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing:
Contraction normalization

◎ Definition

Contraction normalization is a preprocessing step to convert the contraction into the more formal one (just case in several languages).

◎ Example



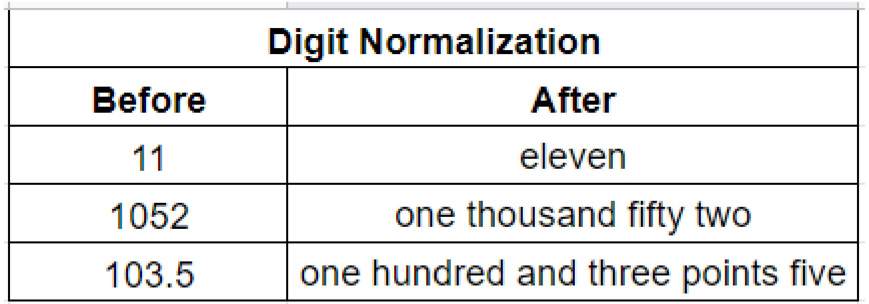
Contraction Normalization	
Before	**After**
i've	i have
you're	you are
wanna	want to

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing:
Digit normalization

- Definition

  Digit normalization is a preprocessing step to convert a digit into a “verbal” form (usually used in text-to-speech task).

- Example



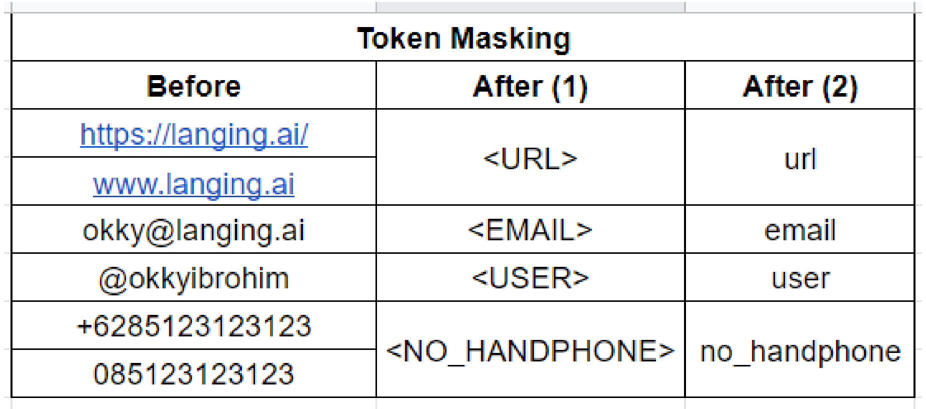
Digit Normalization	
Before	**After**
11	eleven
1052	one thousand fifty two
103.5	one hundred and three points five

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing:
Token masking

Definition

Token masking is a preprocessing step to convert a specific “token” (word, phrases, or a special term) into a verbal form.

Example



Token Masking		
Before	**After (1)**	**After (2)**
https://langing.ai/	<URL>	url
www.langing.ai		
okky@langing.ai	<EMAIL>	email
@okkyibrohim	<USER>	user
+6285123123123	<NO_HANDPHONE>	no_handphone
085123123123		

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing:
Text cleaning (1)

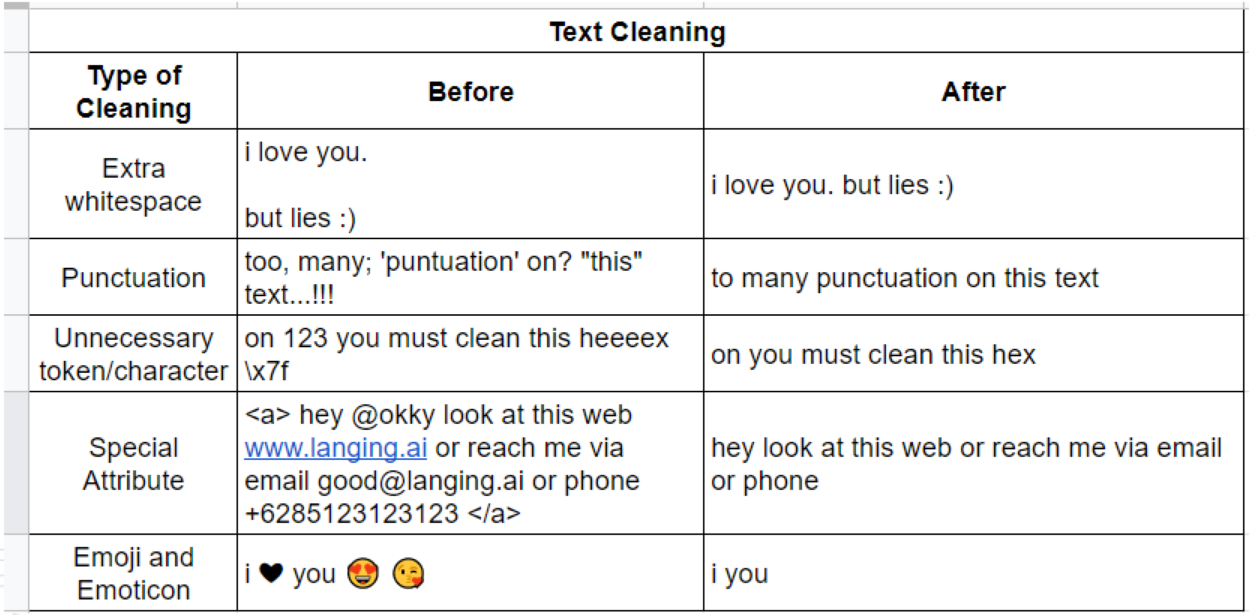
- Definition

  Text cleaning is a preprocessing step remove a specific “token” or “character” from our text.

- Kind of Text Cleaning
  - Extra whitespace (enter, tab, etc.), including triming
  - Punctuation (dot, coma, semicolon, question mark, exclamation mark, etc.)
  - Unnecessary token/character (hex, digit, character repetition, etc.)
  - Special attribute (HTML tag, RT, username, URL, email, No. Handphone, etc.)
  - Emoji and emoticon
  - etc.

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing: Text cleaning (2)

- Example



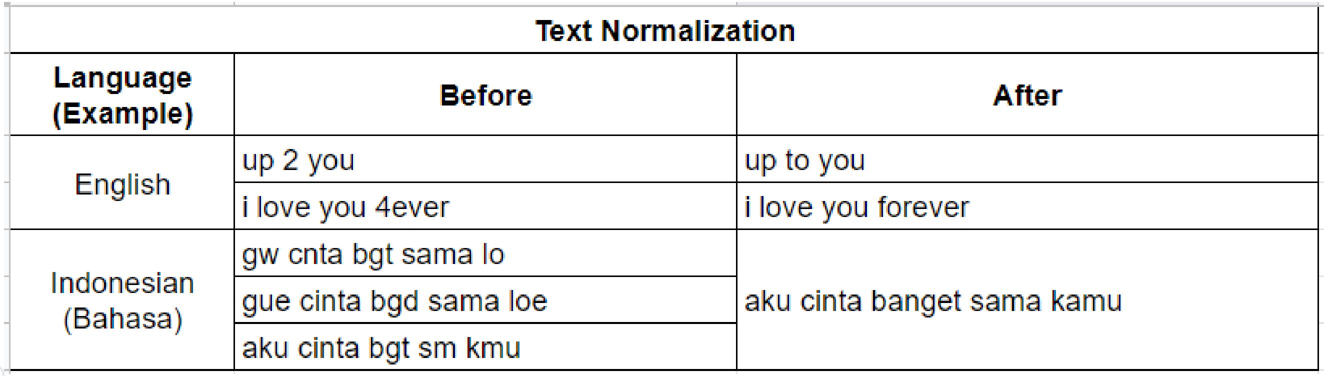
Text Cleaning		
Type of Cleaning	**Before**	**After**
Extra whitespace	i love you. but lies :)	i love you. but lies :)
Punctuation	too, many; 'puntuation' on? "this" text...!!!	to many punctuation on this text
Unnecessary token/character	on 123 you must clean this heeeex \x7f	on you must clean this hex
Special Attribute	<a> hey @okky look at this web www.langing.ai or reach me via email good@langing.ai or phone +6285123123123 </a>	hey look at this web or reach me via email or phone
Emoji and Emoticon	i ❤ you 😍 😘	i you

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing:
Basic text normalization (using dictionary)

Definition

Text normalization is a preprocessing step to normalize a text into the “normal” one (usually to get the formal one).

Example



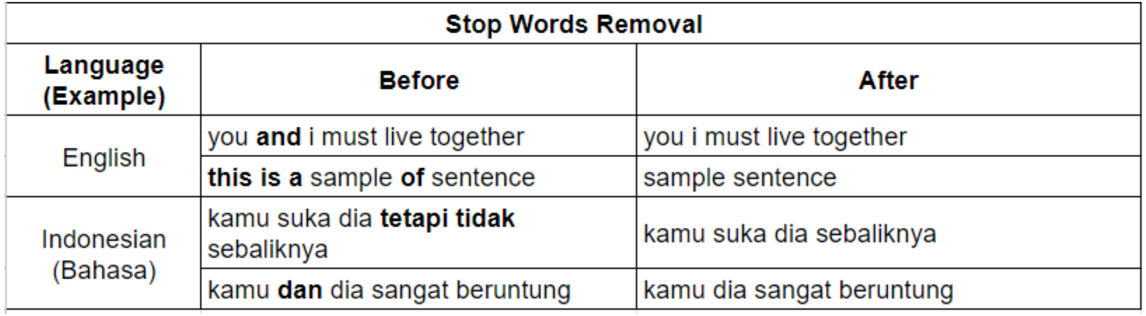
Text Normalization		
Language (Example)	**Before**	**After**
English	up 2 you	up to you
	i love you 4ever	i love you forever
Indonesian (Bahasa)	gw cnta bgt sama lo	aku cinta banget sama kamu
	gue cinta bgd sama loe	
	aku cinta bgt sm kmu	

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing:
Stop words removal

Definition

Stop words removal is a preprocessing step to stop words i.e. words/phrases that do not have meaning in our research (e.g. conjunction, article, to be, or specific words/phrases that we define based on EDA process) from our text.

Example



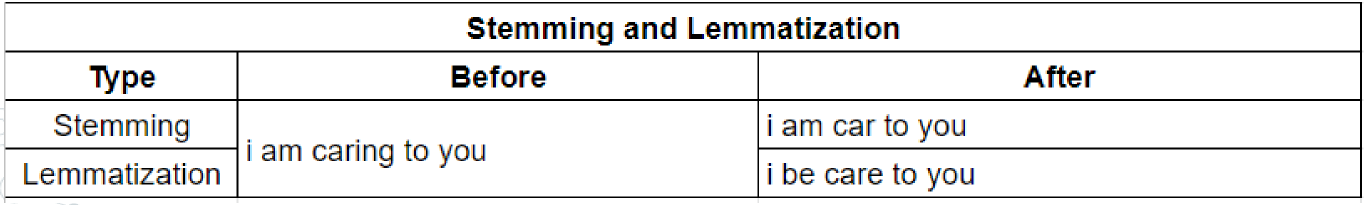
Stop Words Removal

Language (Example)	**Before**	**After**
English	you **and** i must live together	you i must live together
	this is a sample **of** sentence	sample sentence
Indonesian (Bahasa)	kamu suka dia **tetapi tidak** sebaliknya	kamu suka dia sebaliknya
	kamu **dan** dia sangat beruntung	kamu dia sangat beruntung

2.2 Basic text preprocessing in text dataset
Kind of basic text preprocessing: Stemming and lemmatization

- Definition
  - Stemming is a preprocessing step to cut affixed words into the base form.
  - Lemmatization is a preprocessing step to cut affixed words into the base Verb 1 form.
  - In some languages that do not have tenses, stemming is equal to lemmatization.

- Example



Stemming and Lemmatization		
Type	**Before**	**After**
Stemming	i am caring to you	i am car to you
Lemmatization		i be care to you

2.2 Basic text preprocessing in text dataset
Attention in doing text preprocessing (1)

- Order is matter
  Sequence errors will result in non-optimal preprocessing results. For example, stemming/lemmatization must be in the final step of preprocessing.

- Follow your feature extraction process
  Not all preprocessing steps must be done before feature extraction. For example, when you want to use an orthography feature (will be explained in the next chapter), the preprocessing step must be done after you extract the orthography feature.

2.2 Basic text preprocessing in text dataset
Attention in doing text preprocessing (2)

- Source, domain, and language
  Depending on your source (social media, web news, etc.), domain (health, economic, etc.), and languages (English, Indonesian, monolingual, multilingual, etc.) of your dataset, you may need some specific preprocessing steps. For example, preprocessing a news dataset may usually less effort than preprocessing a social media dataset (especially for Twitter). For this reason, a popular text preprocessing library may not be appropriate to your dataset, better you try and look at the output before deciding to use it for your entire dataset.

- Follow the pre-trained word embedding you used
  If you use a particular pre-trained word embedding (will be explained in the next chapter), you must follow the same pre-processing step used by the authors that build the pre-trained word embedding you used.

2.2 Basic text preprocessing in text dataset
Advanced text preprocessing examples

- Text translation
  Some dataset may contain multilinguality whether it is switch-coded or mixed-coded. If we do not have multilingual pre-trained word embedding that cover a particular language of our dataset, text translation often used to normalize it.

- Context aware text normalization
  Actually, a text normalization often can simply done just using dictionary matching. For example, an abbreviation has several interpretations depends on the context. For this case, context aware text normalization e.g. using a seq2seq deep learning model may will be good.

Thank You!

ßisa.ai

PT. Bisa Artifisial Indonesia

Kampus Merdeka

INDONESIA JAYA