| **Nama :**  **Nim :** | **Praktikum Penambangan Teks** | **MODUL 2**    **Nama Dosen :**  **Syandra Sari, S.Kom, M.Kom** |
| --- | --- | --- |
| **Hari dan Tanggal :** | **Nama Asisten Laboratorium :**   1. **Ridho Rachmat Giffary (064001800006)** 2. **Tasya Aulia (065001800022)** |

**Praktikum 2**

**Case Folding & Cleansing**

**KODE POKOK BAHASAN:**

**DESKRIPSI POKOK BAHASAN:**

Mengenalkan dan mengajarkan cara melakukan salah satu proses pada tahapan Pre-Processing yaitu Case folding dan Cleansing yang akan digunakan untuk membersihkan data yang telah ditarik dari platform social media twitter sebelumnya.

| No | Elemen Kompetensi | Indikator Kerja | Jml Jam | Hal |
| --- | --- | --- | --- | --- |
| 1 | Melakukan proses CaseFolding | Mampu melakukan proses casefolding terhadap data yang telah dihimpun | 30 menit | 2 |
| 2 | Melakukan proses Cleansing | Mampu melakukan proses Cleansing terhadap baris yang tidak diperlukan | 60 menit |  |

**TUGAS PENDAHULUAN:**

Hal yang harus dilakukan dan acuan yang harus dibaca sebelum praktikum:

1. Mengetahui pemahaman dasar tentang Kegunaan penarikan data.
2. Mempersiapkan Aplikasi R atau Jupyter Notebook yang sudah terinstall di Laptop masing - masing.

**TEORI SINGKAT:**

| Tuliskan Teori singkat terkait Casefolding dan Cleansing pada Text Mining minimal 2 Paragraf |
| --- |

**LAB SETUP:**

Hal yang harus disiapkan dan dilakukan oleh praktikan untuk menjalankan praktikum modul ini.

**Alat dan Bahan:** Laptop

**Software** : R atau Jupyter Notebook

**ELEMEN KOMPETENSI I:**

**Deskripsi:**

Melakukan proses CaseFolding

**Kompetensi Dasar:**

Mampu melakukan proses casefolding terhadap data yang telah dihimpun

**Latihan 1:**

**Penjelasan Singkat:**

Pada latihan ini anda diminta untuk melakukan proses casefolding terhadap data yang telah dihimpun pada tugas sebelumnya.

Python

| import pandas as pd  import numpy as np  import nltk  from nltk.tokenize import word\_tokenize  from nltk.probability import FreqDist  import string  import re  # drop rows  dataclean.drop(["Datetime", "Tweet Id", "Username"], axis = 1, inplace = True)  dataclean.columns = dataclean.columns.str.strip().str.lower() |
| --- |

| # casefolding  dataclean['tweet'] = dataclean['tweet'].str.lower() |
| --- |

R

| library(tm)    #Load the data  teks <- read.csv("namadata\_kelompok.csv", header = T)  View(teks)  Nama\_Kelompok <- iconv(teks$text, to="UTF-8") #Pengenal Character UTF-8  View(Nama\_Kelompok)    #Case Folding  library(tm)  Nama\_Kelompok 2 <- Corpus(VectorSource(Nama\_Kelompok))  Nama\_Kelompok <- tm\_map(Nama\_Kelompok2, tolower)  inspect(Nama\_Kelompok [10:15])  Nama\_Kelompok2 <- unlist(Nama\_Kelompok)  View(Nama\_Kelompok 2)  write.csv(Nama\_Kelompok2, file = 'C:/Users/ibrahim/Downloads/latihan TA/cleanvaksinsinovac.csv', row.names = F) |
| --- |

**ELEMEN KOMPETENSI 2:**

**Deskripsi:**

Melakukan proses Cleansing

**Kompetensi Dasar:**

Mampu melakukan proses Cleansing terhadap baris yang tidak diperlukan

**Latihan 2:**

**Penjelasan Singkat:**

Pada latihan ini anda diminta untuk melakukan proses Cleansing terhadap baris yang tidak diperlukan.

Python

| # Cleansing process  def remove\_tweet\_special(text):  # remove tab, new line, ans back slice  text = text.replace('\\t'," ").replace('\\n'," ").replace('\\u'," ").replace('\\',"")  # remove non ASCII (emoticon, chinese word, .etc)  text = text.encode('ascii', 'replace').decode('ascii')  # remove mention, link, hashtag  text = ' '.join(re.sub("([@#][A-Za-z0-9]+)|(\w+:\/\/\S+)"," ", text).split())  #remove url (menghapus link)  text = re.sub(r'\w+:\/{2}[\d\w-]+(\.[\d\w-]+)(?:(?:\/[^\s/]))\*', '', text)  # remove incomplete URL  return text.replace("http://", " ").replace("https://", " ")    dataclean['tweet'] = dataclean['tweet'].apply(remove\_tweet\_special)  #remove number (menghapus angka)  def remove\_number(text):  return re.sub(r"\d+", "", text)  dataclean['tweet'] = dataclean['tweet'].apply(remove\_number)  #remove punctuation (menghapus tanda baca)  def remove\_punctuation(text):  return text.translate(str.maketrans("","",string.punctuation))  dataclean['tweet'] = dataclean['tweet'].apply(remove\_punctuation)  # remove single char  def remove\_singl\_char(text):  return re.sub(r"\b[a-zA-Z]\b", "", text)  dataclean['tweet'] = dataclean['tweet'].apply(remove\_singl\_char)  dataclean.head()  #remove whitespace leading & trailing (menghapus spasi awal dan akhir)  def remove\_whitespace\_LT(text):  return text.strip()  dataclean['tweet'] = dataclean['tweet'].apply(remove\_whitespace\_LT)  #remove multiple whitespace into single whitespace  def remove\_whitespace\_multiple(text):  return re.sub('\s+',' ',text)  dataclean['tweet'] = dataclean['tweet'].apply(remove\_whitespace\_multiple)  # remove single char  def remove\_singl\_char(text):  return re.sub(r"\b[a-zA-Z]\b", "", text)  dataclean['tweet'] = dataclean['tweet'].apply(remove\_singl\_char)  dataclean |
| --- |

R

| ##cleaning data  Nama\_Kelompok3 <- read.csv('cleanslangvaksinsinovac.csv')  Nama\_Kelompok 3 <- iconv(Nama\_Kelompok 3$x, to="UTF-8")  Nama\_Kelompok <- Corpus(VectorSource(Nama\_Kelompok3))  Nama\_Kelompok <- tm\_map(Nama\_Kelompok, tolower)  inspect(Nama\_Kelompok [20:25])  removeURL <- function(x) gsub("http[^[:space:]]\*", "", x)  removeAT <- function(x) gsub("@([A-Za-z0-9\_]+)", "", x)  removehash <- function(x) gsub("#([A-Za-z0-9\_]+)", "", x)  removeNL<- function(y) gsub("\n", " ", y)  removeRT <- function(y) gsub("rt", "", y)    Nama\_Kelompok <- tm\_map(Nama\_Kelompok, removeURL)  Nama\_Kelompok <- tm\_map(Nama\_Kelompok,removeAT)  inspect(Nama\_Kelompok [139:144])  Nama\_Kelompok <- tm\_map(Nama\_Kelompok,removehash)  Nama\_Kelompok <- tm\_map(Nama\_Kelompok, removeNL)  Nama\_Kelompok <- tm\_map(Nama\_Kelompok, removeRT)  Nama\_Kelompok <- tm\_map(Nama\_Kelompok, removePunctuation)  Nama\_Kelompok <- tm\_map(Nama\_Kelompok, removeNumbers)  Nama\_Kelompok <- tm\_map(Nama\_Kelompok, stripWhitespace)  Nama\_Kelompok 2 <- unlist(Nama\_Kelompok)  View(Nama\_Kelompok 2)    write.csv(Nama\_Kelompok 2, file = 'C:/Users/ibrahim/Downloads/latihan TA/cleanvaksinsinovac.csv', row.names = F)  View(Nama\_Kelompok 2) |
| --- |



| **No** | **Elemen Kompetensi** | **Penyelesaian**  **Cek List** | |
| --- | --- | --- | --- |
| **Selesai** | **Tidak** |
| **1** | **Melakukan proses CaseFolding** |  |  |
| **2.** | **Melakukan proses Cleansing** |  |  |

1. **Form Umpan Balik**

| **Elemen Kompetensi** | **Waktu Pengerjaan** | **Kriteria** |
| --- | --- | --- |
| **Melakukan proses CaseFolding** |  |  |
| **Melakukan proses Cleansing** |  |  |

Kriteria

1.Sangat Menarik

2.Cukup Menarik

3.Kurang Menarik

4.Sangat Kurang Menarik