> Nama: Laksmi Dyah Nurlita

> Kelas: S1SD02A

> NIM: 21110023

```
import numpy as np
import pandas as pd
import re
import re as reg
import matplotlib.pyplot as plt
%matplotlib inline
```

1. PREPROCESSING

```
import csv
data=pd.read_csv('dataset.csv', delimiter=';', encoding='latin1')
data
```

	tweet_akhir
0	Badan Meteorologi Klimatologi dan Geofisika (B...
1	Update Infografis percepatan penanganan COVID-...
2	Peringatan Dini Cuaca DIY Tanggal 07 April 202...
3	Mitigasi berbasis ekosistem
4	Perkembangan penanganan Pandemi COVID-19 Indon...
...	...
1276	Update sebaran kejadian bencana alam di Indone...
1277	Sebanyak 912 jiwa diungsikan setelah Kilang Mi...
1278	Selamat malam sobatkriskes berikut perkembanga...
1279	Sebanyak 932 jiwa diungsikan setelah Kilang Mi...
1280	Salam santun Daerah Sebaran Kasus Positif CoVi...

1281 rows × 1 columns

```
pip install Sastrawi
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting Sastrawi
  Downloading Sastrawi-1.0.1-py2.py3-none-any.whl (209 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 209.7/209.7 KB 5.3 MB/s eta 0:00:00
Installing collected packages: Sastrawi
Successfully installed Sastrawi-1.0.1
```

```
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
```

```
slangs={'yg':'yang', 'tdk':'tidak', 'pd':'pada', 'mlh':'malah', 'jgn':'jangan', 'jg':'juga', 'tp':'tapi', 'blkg': 'belakang', 'dr':'dari'
        'dlm':'dalam','dgn':'dengan', 'poto':'foto', 'g':'tidak', 'n':'dan', 'ad':'ada', 'brp': 'berapa', "abis": "habis", "ad": "ada", '
        "ahaha": "haha", "aj": "saja", "ajep-ajep": "dunia gemerlap", "ak": "saya", "akika": "aku", "akkoh": "aku", "akuwh": "aku", "alay
        "ancur": "hancur", "anjrit": "anjing", "anter": "antar", "ap2": "apa-apa", "apasih": "apa sih", "apes": "sial", "aps": "apa", "ac
        "aseekk": "asyik", "asekk": "asyik", "asem": "asam", "aspal": "asli tetapi palsu", "astul": "asal tulis", "ato": "atau", "au ah":
        "ayank": "sayang", "b4": "sebelum", "bakalan": "akan", "bandes": "bantuan desa", "bangedh": "banget", "banpol": "bantuan polisi",
        "bcanda": "bercanda", "bdg": "bandung", "begajulan": "nakal", "beliin": "belikan", "bencong": "banci", "bentar": "sebentar", "ber
        "bosan", "beud": "banget", "bg": "abang", "bgmn": "bagaimana", "bgt": "banget", "bijimane": "bagaimana", "bintal": "bimbingan mer
        "blegug": "bodoh", "blh": "boleh", "bln": "bulan", "blum": "belum", "bnci": "benci", "bnran": "yang benar", "bodor": "lucu", "bok
        "bohong", "boljug": "boleh juga", "bonek": "bocah nekat", "boyeh": "boleh", "br": "baru", "brg": "bareng", "bro": "saudara laki-l
        "bt": "buat", "btw": "ngomong-ngomong", "buaya": "tidak setia", "bubbu": "tidur", "bubu": "tidur", "bumil": "ibu hamil", "bw": "b
        "cabal": "sabar", "cadas": "keren", "calo": "makelar", "can": "belum", "capcus": "pergi", "caper": "cari perhatian", "ce": "cewek
        "cengengesan": "tertawa", "cepet": "cepat", "cew": "cewek", "chuyunk": "sayang", "cimeng": "ganja", "cipika cipiki": "cium pipi k
        "ckp": "cakep", "cmiiw": "correct me if i'm wrong", "cmpur": "campur", "cong": "banci", "conlok": "cinta lokasi", "cowwyy": "maaf
        "cucok": "cocok", "cuex": "cuek", "cumi": "Cuma miscall", "cups": "culun", "curanmor": "pencurian kendaraan bermotor", "curcol":
        "d": "di", "dah": "deh", "dapet": "dapat", "de": "adik", "dek": "adik", "demen": "suka", "deyh": "deh", "dgn": "dengan", "diancur
        "dimintak": "diminta", "disono": "di sana", "dket": "dekat", "dkk": "dan kawan-kawan", "dll": "dan lain-lain", "dlu": "dulu", "dr
        "dongs": "dong", "dpt": "dapat", "dri": "dari", "drmn": "darimana", "drtd": "dari tadi", "dst": "dan seterusnya", "dtg": "datang"
        "egp": "emang gue pikirin", "eke": "aku", "elu": "kamu", "emangnya": "memangnya", "emng": "memang", "endak": "tidak", "enggak": '
        "fifo": "first in first out", "folbek": "follow back", "fyi": "sebagai informasi", "gaada": "tidak ada uang", "gag": "tidak", "ga
        "gan": "juragan", "gaptek": "gagap teknologi", "gatek": "gagap teknologi", "gawe": "kerja", "gbs": "tidak bisa", "gebetan": "orar
        "gepeng": "gelandangan dan pengemis", "ghiy": "lagi", "gile": "gila", "gimana": "bagaimana", "gino": "gigi nongol", "githu": "git
        "gn": "begini", "goblok": "bodoh", "golput": "golongan putih", "gowes": "mengayuh sepeda", "gpny": "tidak punya", "gr": "gede ras
        "gua": "saya", "guoblok": "goblok", "gw": "saya", "ha": "tertawa", "haha": "tertawa", "hallow": "halo", "hankam": "pertahanan dar
        "hlm": "halaman", "hny": "hanya", "hoax": "isu bohong", "hr": "hari", "hrus": "harus", "hubdar": "perhubungan darat", "huff": "me
        "ilfil": "tidak suka", "imho": "in my humble opinion", "imoetz": "imut", "item": "hitam", "itungan": "hitungan", "iye": "iya", "
        "jayus": "tidak lucu", "jdi": "jadi", "jem": "jam", "jga": "juga", "jgnkan": "jangankan", "jir": "anjing", "jln": "jalan", "jomb
        "jutek": "galak", "k": "ke", "kab": "kabupaten", "kabor": "kabur", "kacrut": "kacau", "kadiv": "kepala divisi", "kagak": "tidak",
        "kamtibmas": "keamanan dan ketertiban masyarakat", "kamuwh": "kamu", "kanwil": "kantor wilayah", "karna": "karena", "kasubbag": "
        "kayanya": "kayaknya", "kbr": "kabar", "kdu": "harus", "kec": "kecamatan", "kejurnas": "kejuaraan nasional", "kekeuh": "keras ke
        "kepengen": "mau", "kepingin": "mau", "kepsek": "kepala sekolah", "kesbang": "kesatuan bangsa", "kesra": "kesejahteraan rakyat",
        "kibul": "bohong", "kimpoi": "kawin", "kl": "kalau", "klianz": "kalian", "kloter": "kelompok terbang", "klw": "kalau", "km": "ka
        "knp": "kenapa", "kodya": "kota madya", "komdis": "komisi disiplin", "komsov": "komunis sovyet", "kongkow": "kumpul bareng teman-
        "kpn": "kapan", "krenz": "keren", "krm": "kirim", "kt": "kita", "ktmu": "ketemu", "ktr": "kantor", "kuper": "kurang pergaulan", "
        "lam": "salam", "lamp": "lampiran", "lanud": "landasan udara", "latgab": "latihan gabungan", "lebay": "berlebihan", "leh": "bole
        "lgsg": "langsung", "liat": "lihat", "litbang": "penelitian dan pengembangan", "lmyn": "lumayan", "lo": "kamu", "loe": "kamu", "
        "lp": "lupa", "luber": "langsung, umum, bebas, dan rahasia", "luchuw": "lucu", "lum": "belum", "luthu": "lucu", "lwn": "lawan", "
        "kptsan":"keputusan", "krik": "garing", "krn": "karena", "ktauan": "ketahuan", "ktny": "katanya", "kudu": "harus", "kuq": "kok",
        "lambreta": "lambat", "lansia": "lanjut usia", "lapas": "lembaga pemasyarakatan", "lbur": "libur", "lekong": "laki-laki", "lg": "
        "linmas": "perlindungan masyarakat", "lmyan": "lumayan", "lngkp": "lengkap", "loch": "loh", "lol": "tertawa", "lom": "belum", "l
        "luchu": "lucu", "luff": "cinta", "luph": "cinta", "lw": "kamu", "lwt": "lewat", "maaciw": "terima kasih", "mabes": "markas besa
        "maen": "main", "mahatma": "maju sehat bersama", "mak": "ibu", "makasih": "terima kasih", "malah": "bahkan", "malu2in": "memaluk
        "markus": "makelar kasus", "mba": "mbak", "mending": "lebih baik", "mgkn": "mungkin", "mhn": "mohon", "miker": "minuman keras", "
        "mnt": "minta", "moge": "motor gede", "mokat": "mati", "mosok": "masa", "msh": "masih", "mskpn": "meskipun", "msng2": "masing-ma
        "mumet": "pusing", "muna": "munafik", "munaslub": "musyawarah nasional luar biasa", "musda": "musyawarah daerah", "muup": "maaf",
        "naon": "apa", "napol": "narapidana politik", "naq": "anak", "narsis": "bangga pada diri sendiri", "nax": "anak", "ndak": "tidak"
        "nelfon": "menelepon", "ngabis2in": "menghabiskan", "ngakak": "tertawa", "ngambek": "marah", "ngampus": "pergi ke kampus", "ngan
        "ngaruh": "berpengaruh", "ngawur": "berbicara sembarangan", "ngeceng": "kumpul bareng-bareng", "ngeh": "sadar", "ngekos": "tingg
        "ngemeng": "bicara terus-terusan", "ngerti": "mengerti", "nggak": "tidak", "ngikut": "ikut", "nginep": "menginap", "ngisi": "men
        "ngomongin": "membicarakan", "ngumpul": "berkumpul", "ni": "ini", "nyasar": "tersesat", "nyariin": "mencari", "nyiapin": "memper
        "ok": "ok", "priksa": "periksa", "pro": "profesional", "psn": "pesan", "psti": "pasti", "puanas": "panas", "qmo": "kamu", "qt": "
        "red": "redaksi", "reg": "register", "rejeki": "rezeki", "renstra": "rencana strategis", "reskrim": "reserse kriminal", "sni": "
        "sosbud": "sosial-budaya", "sospol": "sosial-politik", "sowry": "maaf", "spd": "sepeda", "sprti": "seperti", "spy": "supaya", "s
        "sumbangin": "sumbangkan", "sy": "saya", "syp": "siapa", "tabanas": "tabungan pembangunan nasional", "tar": "nanti", "taun": "ta
        "tekor": "rugi", "telkom": "telekomunikasi", "telp": "telepon", "temen2": "teman-teman", "tengok": "menjenguk", "terbitin": "ter
        "thd": "terhadap", "thx": "terima kasih", "tipi": "TV", "tkg": "tukang", "tll": "terlalu", "tlpn": "telepon", "tman": "teman", "
        "tnda": "tanda", "tnh": "tanah", "togel": "toto gelap", "tp": "tapi", "tq": "terima kasih", "trgntg": "tergantung", "trims": "te
        "reklamuk": "reklamasi", "sma": "sama", "tren": "trend", "ngehe": "kesal", "mz": "mas", "analisise": "analisis", "sadaar": "sada
        "zonk": "bodoh", "rights": "benar", "simiskin": "miskin", "ngumpet": "sembunyi", "hardcore": "keras", "akhirx": "akhirnya", "sol
        "masy": "masyarakat", "still": "masih", "tauk": "tahu", "mbual": "bual", "tioghoa": "tionghoa", "ngentotin": "senggama", "kentot
        "rubahnn": "rubah", "trlalu": "terlalu", "nyela": "cela", "heters": "pembenci", "nyembah": "sembah", "most": "paling", "ikon": "
        "setting": "atur", "seting": "akting", "next": "lanjut", "waspadalah": "waspada", "gantengsaya": "ganteng", "parte": "partai", "
        "jentelmen": "berani", "buangbuang": "buang", "tsangka": "tersangka", "kurng": "kurang", "ista": "nista", "less": "kurang", "koa
        "tahi": "kotoran", "tirani": "tiran", "tilep": "tilap", "happy": "bahagia", "tak": "tidak", "penertiban": "tertib", "uasai": "ku
        "taik": "tahi", "wkwkkw": "tertawa", "ahokncc": "ahok", "istaa": "nista", "benarjujur": "jujur", "mgkin": "mungkin", 'ga':'tidak'
        'ny':'nya', 'htm':'harga tiket masuk', 'cm':'cuma', 'slalu':'selalu', 'tingi':'tinggi','neng':'senang'}
processed_comments = []

for sentence in data['tweet_akhir']:
  # Remove all the special characters
  processed_comment = re.sub(r'\W', ' ', str(sentence))

  # Converting to Lowercase
  processed_comment = processed_comment.lower()

  #Remove number
  processed_comment = re.sub(r'\d+', ' ', processed_comment)

  # remove all single characters
  processed_comment = re.sub(r'\s+[a-zA-Z]\s+', ' ', processed_comment)

  #remove duplicate character
  pattern=reg.compile(r"(.)\1{1,}",reg.DOTALL)
  processed_comment=pattern.sub(r"\1",processed_comment)

  #Corrected Slang words
  words = processed_comment.split()
  rfrm=[slangs[word] if word in slangs else word for word in words]
  processed_comment= " ".join(rfrm)

  #remove stopword
  factory = StopWordRemoverFactory()
  more_stopword = ['tak', 'jd', 'per', 'nya', 'terjemah', 'diterjemahkan', 'oleh', 'gogle', 'google' ,'nan', 'baik', 'sangat', 'batas', '
                   'ada','bersih', 'salur', 'baru', 'purwokerto', 'batas', 'hotel', 'coba', 'putus', 'ada',
                   'com', 'kamu', 'http', 'https', 'htps', 'htp', 'gak', 'jadi', 'lebih', 'kalau', 'banyak', 'jangan', 'iya'] #menambahk
  stopwords = factory.get_stop_words() + more_stopword
  temp = [t for t in re.findall(r'\b[a-z]+-?[a-z]+\b',processed_comment) if t not in stopwords]
  processed_comment = ' '.join(temp)

  #stemming
  stemmer = StemmerFactory().create_stemmer()
  processed_comment = stemmer.stem(processed_comment)
  # Substituting multiple spaces with single space
  processed_comment = re.sub(r'\s+', ' ', processed_comment, flags=re.I)
```

```
processed_comments
```

```
      'ingat dini gelombang tinggi wilayah air samudera hindia selatan jawa barat jawa tengah yogyakarta',
      'peingatan dini cuaca diy tangal april pukul wib infocuacajogja bmkgdiy',
      'salah satu twet mutual pihak bmkg jawab inti bukan tugas',
      'sembuh kasus barulbh sektr hari suspek bwh ribu postv rate hari sekt',
      'ta sen satu',
      'presiden segera perintah bnpb basarnas menteri sosial menteri sehat tni polri seger',
      'lihat mobil bantuanya pak',
      'strategi cepat tangan covid sbg in putsuport ekspektasi ukur bangsa negara',
      'sat resmi perintah daerah bencana',
      'pimpin alamiah jelas multi',
      'salam santun daerah sebar kasus positif covid indonesia tanggal apr',
      'yth bapak ibu ikut sampai prakiran cuaca esok hari beberapa daerah wisata diy selasa april',
      'salam santun daerah sebar kasus positif covid indonesia tanggal apr',
      'gempa palu tenda ramah perempuan anak mampu bantu prempuan anak selamat ancam leceh',
      'banjir bandang terjang kabupaten flores timur nt',
      'kepala badan nasional penangulangan bencana bnpb letnan jendral tni doni monardo tengah terima lapora',
      'alhamdulilah selalu lihat update moga makin turun terus kasus hari',
      'inabuoybpt rupa salah satu inovasi dukung ekosistem ina tews sama',
      'galau saudara laki tf dulu kirim',
      'maskapai terbang pelita air service pas salah satu anak usaha bumn pertamina jalan misi kemanusian me',
      'moga lemah alah swt kekuatanya mulai hilang penularanya tahap',
      'anjng sinte satu juragan',
      'mingu april jembatan kamba niru waingapu sumba timur nt roboh terjang banjir bapak bapak basuki',
      'kan kemensos mas alam urusin bansos',
      'update tg nt sat bantu tiba prayfornt',
      'update tinggi muka air bendung wilayah kabupaten kendal senin april sumber',
      'update bencana nusa tengara timur sama juang pulih cepat informasi lokasi dampak mungkin',
      'sedih banget deh ber bulan bulan bayar',
      'kemarin beneran cuman rekap data ketingalan',
      'mulai lambat penularanya',
      'presiden segera perintah bnpb basarnas menteri sosial menteri sehat tni polri un',
      'gainget ser pas galau mesenya',
      'kerja udh kelar kerja sesuai sop kerja tarung nyawa gaji ntarin melulu',
      'badan nasional penangulangan bencana bnpb lalu deputi bidang logistik alat kirim bantu',
      'nyimak om',
      'moga jumlah mati hari tekan digit digit',
      'terima kasih bpb prayfornt',
      'bpbd kabupaten lembata catat wil dampak banjir adl desa waowala desa tanjung batu desa amakaka camat ile apa',
      'badan penangulangan bencana daerah bpbd kabupaten lembata nt lapor warga meningal dunia akibat banjir bandan',
      'ditangkep',
      'samping korban jiwa banjir bandang akibat jembatan puluh rumah warga timbun lumpur se',
      'data mingu pukul wib banjir bandang landa empat desa tiga camat kabupaten fl',
      'alhamdulilah turun tambah kasus konfirmasi positif hari cukup drastis',
      'nt terjang lah nina moga saudara saudara sana prayfornt',
      'bpbd kabupaten flores timur informasi warga kira hilang akibat banjir bandang mingu dini har',
      'yth bapak ibu ikut sampai prakiran cuaca esok hari kabupaten sleman selasa april moga ber',
      'ingat dini gelombang tinggi wilayah air samudera hindia selatan jawa barat jawa tengah yogya',
      'alhamdulilah turun tambah kasus konfirmasi positif hari cukup drastis yait',
      'selamat malam sobatkriskes ikut kembang covid indonesia tangal april pkl wib covi',
      'yth bapak ibu ikut sampai prakiran cuaca esok hari kabupaten sleman selasa april semo',
      'mingu april warga sekitar noelbaki kupang tengah mulai ungsi akibat air laut mulai naik al',
      'yth bapak ibu ikut sampai prakiran cuaca esok hari lereng gunung rapi selasa apri',
      'yth bapak ibu ikut sampai prakiran cuaca esok hari kabupaten bantul selasa april semo',
      'yth bapak ibu ikut sampai prakiran cuaca esok hari kabupaten kulon progo selasa april',
      'yth bapak ibu ikut sampai prakiran cuaca esok hari beberapa daerah wisata diy selasa apr',
      'kejadianx hampir saman dn sebab jatuhx korban jiwa',
      'yth bapak ibu ikut sampai prakiran cuaca esok hari wilayah kota yogyakarta selasa april',
      'saran baik dinformasikan berapa persentase pasien meningal covi',
      'yth bapak ibu ikut sampai prakiran cuaca esok hari kabupaten gunungkidul selasa april',
```

```
import nltk
```

```
nltk.download('punkt')
```

```
     [nltk_data] Downloading package punkt to /root/nltk_data...
     [nltk_data]   Unzipping tokenizers/punkt.zip.
     True
```

```
from nltk.tokenize import word_tokenize

docs = ' '.join(processed_comments)
hasil_tokenizing = nltk.word_tokenize(docs)
hasil_tokenizing
```

```
       'darah',
       'pmi',
       'kota',
       'bekas',
       'rabu',
       'april',
       'stok',
       'darah',
       'waktu',
       'waktu',
       'ubah',
       'mantra',
       'coronareda',
       'cinta',
       'bulan',
       'mulia',
       'ajak',
       'korona',
       'hormat',
       'ramadhan',
       'serang',
       'umat',
       'pak',
       'segera',
       'respon',
       'bantu',
       'ribet',
       'update',
       'citra',
       'radar',
       'cuaca',
       'diy',
       'tangal',
       'april',
       'pukul',
       'wib',
       'infocuacajogja',
       'bmkgdiy',
       'wes',
       'mongo',
       'sak',
       'kerso',
       'panjenengan',
       'prayfornt',
       'lurah',
       'besar',
       'simalungun',
       'turut',
```

2. Bag of Word dan tampilkan dalam bentuk grafik histogram untuk setiap katanya

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(hasil_tokenizing)
print(vectorizer.get_feature_names())
Doc_Term_Matrix = pd.DataFrame(X.toarray(), columns = vectorizer.get_feature_names())
```

```
['abang', 'abg', 'abk', 'abrasi', 'absen', 'acara', 'ace', 'ada', 'adam', 'adan', 'adil', 'adisasmito', 'adl', 'admi', 'admin', 'ad
```

```
Doc_Term_Matrix
```

```
abang  abg  abk  abrasi  absen  acara  ace  ada  adam  adan  ...  yogya  yogyaka  yogyakarta  yoh
```

```
token_freq = {}
for token in hasil_tokenizing:
    if token in token_freq:
        token_freq[token] += 1
    else:
        token_freq[token] = 1
for token, frequency in token_freq.items():
    print(f"{token} = {frequency}")
```

```
puncakmusimkemau = 2
rsdc = 1
wisma = 1
atlet = 1
kawulamoda = 2
kalteng = 2
kh = 2
ma = 2
ruf = 2
eksperimental = 1
peduliklim = 2
astrazeneca = 1
novavax = 1
diplomasi = 1
kemlu = 1
sip = 1
bengkel = 1
sasar = 1
pns = 1
engan = 1
jel = 1
elwasi = 1
bermanfat = 1
sumedang = 1
berintera = 1
hubunganya = 1
nu = 1
pela = 1
frasa = 1
tunjuk = 1
berangkat = 1
redaksi = 1
topik = 1
dala = 1
sdm = 1
yamg = 1
trace = 2
bismilah = 1
last = 1
day = 1
absen = 1
lpj = 1
spj = 1
cair = 1
minimal = 1
febuari = 1
merchandise = 1
hehehe = 1
vakninasi = 1
he = 1
apal = 1
unfaedah = 1
sengsara = 1
walikotanya = 1
kilang = 2
minyak = 2
pt = 2
balong = 4
```

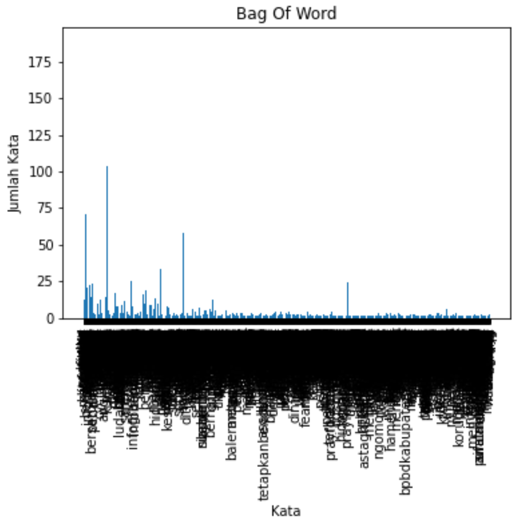
```
keys = list(token_freq.keys())
values = list(token_freq.values())

# Plot the histogram as a bar graph
plt.bar(keys, values)

# Add labels and title
plt.xlabel('Kata')
plt.ylabel('Jumlah Kata')
plt.title('Bag Of Word')

plt.xticks(rotation=90)

# Show the plot
plt.show()
```



3. Vektorisasi menggunakan TF-IDF dan tampilkan hasilnya dalam bentuk daaframe berupa nama fitur dan nilai vektornya

```
pip install sklearn
```

```
ın indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
ıg sklearn
ıding sklearn-0.0.post1.tar.gz (3.6 kB)
ıng metadata (setup.py) ... done
 wheels for collected packages: sklearn
ıg wheel for sklearn (setup.py) ... done
ı wheel for sklearn: filename=sklearn-0.0.post1-py3-none-any.whl size=2344 sha256=ba7665877401fd2b1269257719b04fb31abe91351b0b2ba300
 in directory: /root/.cache/pip/wheels/14/25/f7/1cc0956978ae479e75140219088deb7a36f60459df242b1a72
ılly built sklearn
ıg collected packages: sklearn
ılly installed sklearn-0.0.post1
```

```
pip install scikit-learn
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.8/dist-packages (1.0.2)
Requirement already satisfied: scipy>=1.1.0 in /usr/local/lib/python3.8/dist-packages (from scikit-learn) (1.7.3)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.8/dist-packages (from scikit-learn) (3.1.0)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.8/dist-packages (from scikit-learn) (1.2.0)
Requirement already satisfied: numpy>=1.14.6 in /usr/local/lib/python3.8/dist-packages (from scikit-learn) (1.21.6)
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
vektor = TfidfVectorizer(max_features=400)
vektor
```

```
TfidfVectorizer(max_features=400)
```

```
df = pd.DataFrame(processed_comments, columns = ['CleanText'])
df
```

	CleanText
0	badan meteorologi klimatologi geofisika bmkg r...
1	update infografis cepat tangan covid indonesia...
2	ingat dini cuaca diy tangal april pukul wib in...
3	mitigasi bas ekosistem
4	kembang tangan pandemi covid indonesia sebar k...
...	...
1276	update sebar jadi bencana alam indonesia perio...
1277	banyak jiwa ungsi kilang minyak milik pt perta...
1278	selamat malam sobatkriskes ikut kembang covid ...
1279	banyak jiwa ungsi kilang minyak milik pt perta...
1280	salam santun daerah sebar kasus positif covid ...

1281 rows × 1 columns

```
#menghitung tf-idf dengan TfidfTransformer
vektor_dt = vektor.fit_transform(df['CleanText'].values.astype('U'))
print (vektor_dt)
print (vektor_dt.shape)
```

```
  (0, 331)	0.4028017069510187
  (0, 378)	0.385932005388073
  (0, 335)	0.36457003261869936
  (0, 151)	0.39111598577938983
  (0, 134)	0.345372702217839
  (0, 58)	0.37659125430727325
  (0, 24)	0.37659125430727325
  (1, 52)	0.42375877016705776
  (1, 395)	0.21485253943992824
  (1, 280)	0.22123794224328966
  (1, 19)	0.2032366218150581
  (1, 351)	0.21398625981543024
  (1, 129)	0.24098164212776252
  (1, 73)	0.4473359311333447
  (1, 352)	0.2725006498106083
  (1, 70)	0.32083700019419187
  (1, 131)	0.40155152247130843
  (1, 388)	0.20397848639675384
  (2, 59)	0.2944180479562987
  (2, 130)	0.2936753161839819
  (2, 93)	0.28934802850588187
  (2, 74)	0.26713464543045307
  (2, 91)	0.44652148598115265
  (2, 135)	0.4382745134117255
  (2, 395)	0.2665921501099091
  :	:
  (1278, 128)	0.2683760472068566
  (1278, 338)	0.3542421812547135
  (1278, 321)	0.3086328268542802
  (1278, 213)	0.3303628584503251
  (1278, 165)	0.30273631191647987
  (1278, 395)	0.20273118548223917
  (1278, 351)	0.20191377882886008
  (1278, 129)	0.22738616036547935
  (1278, 73)	0.42209854192959384
  (1279, 230)	0.45066527037484855
  (1279, 89)	0.3690438729602981
  (1279, 67)	0.3728723804650749
  (1279, 38)	0.40158342151446363
  (1279, 150)	0.2808699705260557
  (1279, 148)	0.3690438729602981
  (1279, 386)	0.3812126137034672
  (1280, 354)	0.39088700850945207
  (1280, 273)	0.37536971580212614
  (1280, 306)	0.40007825102381817
  (1280, 301)	0.37195356041180067
  (1280, 78)	0.3046665769472186
  (1280, 159)	0.3116655940379871
  (1280, 313)	0.3131491616284104
  (1280, 129)	0.2537659389134113
  (1280, 73)	0.23553375595634282
(1281, 400)
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
vektorizer = TfidfVectorizer()
vektor = vektorizer.fit_transform(hasil_tokenizing)
vektor
```

```
<11335x2381 sparse matrix of type '<class 'numpy.float64'>'
	with 11335 stored elements in Compressed Sparse Row format>
```

```
matrix = pd.DataFrame(vektor.toarray(),columns = vektorizer.get_feature_names())
pd.set_option('display.precision',2)
matrix
```

```
/usr/local/lib/python3.8/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function get_fea
  warnings.warn(msg, category=FutureWarning)
```

	abang	abg	abk	abrasi	absen	acara	ace	ada	adam	adan	...	yogya	yogyaka	yogyakarta	yoh
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
11330	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0

4. Pemodelan dengan TOPIC MODELLING

Topik Modelling menggunakan LSA

```
from sklearn.decomposition import TruncatedSVD
lsa_model = TruncatedSVD(n_components=10, algorithm='randomized', n_iter=10, random_state=42)
lsa_top=lsa_model.fit_transform(vektor_dt)
```

```
print(lsa_top.shape)
```

```
(1281, 10)
```

```
print(lsa_top)
```

```
[[ 0.00282605  0.05738323 -0.03199464 ...  0.00872899  0.03307177
  -0.00898541]
 [ 0.35163366  0.3567197   0.37804094 ...  0.10141704  0.00793036
  -0.0512734 ]
 [ 0.68252738 -0.02734006 -0.02220413 ... -0.0944403  -0.09610605
   0.04900215]
 ...
 [ 0.19288761  0.37550682  0.50284458 ...  0.12517366 -0.11066733
   0.10505252]
 [ 0.0049283   0.09526338 -0.0809934  ... -0.03316647  0.04416228
   0.04736188]
 [ 0.02711611  0.35417459  0.46796484 ... -0.29735536  0.02954897
  -0.17796046]]
```

```
# Memunculkan nilai lsa setiap topik
r = lsa_top[0]
print("Topik-topik:")
for i,topic in enumerate(r):
  print("Topic ",i," : ",topic*100)
```

```
Topik-topik:
Topic  0  :  0.28260502238791735
Topic  1  :  5.738322692780238
Topic  2  :  -3.199463669219448
Topic  3  :  0.40355963070420886
Topic  4  :  -2.6025314303174056
Topic  5  :  3.9387648222701244
Topic  6  :  1.4155685284339983
Topic  7  :  0.8728985724124841
Topic  8  :  3.307176855287998
Topic  9  :  -0.8985410794715755
```

```
# Memunculkan jumlah kata-kata dalam setiap topik
print(lsa_model.components_.shape)
print(lsa_model.components_)
```

```
(10, 400)
[[ 0.00066303  0.00072516  0.000199   ...  0.00138338  0.00189063
   0.00636787]
 [ 0.01564368  0.03794099  0.00927259 ...  0.00561774  0.01007982
   0.03291496]
 [-0.00215973 -0.03984106  0.00472703 ... -0.00072972 -0.00227022
  -0.01583052]
 ...
 [ 0.00140888 -0.04024834 -0.00324476 ... -0.00918064 -0.01555816
  -0.02711291]
 [ 0.00173638  0.02769134  0.00934694 ... -0.00335185 -0.00455883
  -0.02723401]
 [-0.00905372  0.01272199 -0.00149411 ... -0.0039335  -0.00593918
  -0.0104941 ]]
```

```
# Word/ kata paling penting dalam setiap topik
vocab = vektor.get_feature_names()
for i, comp in enumerate(lsa_model.components_):
  vocab_comp = zip(vocab, comp)
  vocab_comp = zip(vocab, comp)
  sorted_words = sorted(vocab_comp, key= lambda x:x[1], reverse=True)[:10]
  print("Topic "+str(i)+" : ")
  for a in sorted_words:
    print(a[0],end=", ")
  print("\n")
```

```
Topic 0 :
infocuacajogja, bmkgdiy, pukul, tangal, diy, wib, citra, radar, cuaca, update, 

Topic 1 :
covid, nt, bencana, indonesia, banjir, tangan, bandang, timur, sebar, kasus, 

Topic 2 :
covid, indonesia, kasus, kembang, sebar, april, sobatkriskes, salam, pkl, positif, 

Topic 3 :
hari, ikut, prakiran, bapak, ibu, esok, sampai, yth, kabupaten, selasa, 

Topic 4 :
banjir, bandang, timur, flores, kabupaten, terjang, meningal, data, covid, warga, 

Topic 5 :
bnpb, kepala, doni, monardo, tangan, bencana, tni, satgas, ketua, nasional, 

Topic 6 :
daerah, sebar, salam, bencana, kasus, positif, santun, tanggal, apr, jadi, 

Topic 7 :
min, hujan, bencana, update, jadi, moga, maret, alam, selamat, periode, 

Topic 8 :
min, pak, update, hujan, apa, mohon, bagaimana, vaksin, kepala, bnpb, 

Topic 9 :
pak, bencana, apa, jadi, alam, bagaimana, periode, nasional, april, januari, 
```

✓ 0s completed at 2:41 PM ● ✕