In [17]:
```
import numpy as np
import pandas as pd
import re
```

Dataset

In [18]:
```
data = pd.read_csv("dataset-bnpb-BERSIH-Dupli (1).csv", encoding="ISO-8859-1") # read file csv nya dulu yang java heritage
data
```

Out[18]:

	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

Import untuk Pembersihan menggunakan NLTK dan Sastrawi

In [19]:
```
import Sastrawi
import re
import nltk
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from nltk.corpus import stopwords

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

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\USER\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\USER\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[19]:
```
True
```

Pre processing

In [22]:
```
slangs={"@": "di", "abis": "habis", "ad": "ada", "adlh": "adalah", "afaik": "as far as i know",
        "ahaha": "haha", "aj": "saja", "ajep-ajep": "dunia gemerlap", "ak": "saya", "akika": "aku",
        "akkoh": "aku", "akuwh": "aku", "alay": "norak", "alow": "halo", "ambilin": "ambilkan",
        "ancur": "hancur", "anjrit": "anjing", "anter": "antar", "ap2": "apa-apa", "apasih": "apa sih",
        "apes": "sial", "aps": "apa", "aq": "saya", "aquwh": "aku", "asbun": "asal bunyi", "aseekk": "asyik",
        "asekk": "asyik", "asem": "asam", "aspal": "asli tetapi palsu", "astul": "asal tulis", "ato": "atau",
        "au ah": "tidak mau tahu", "awak": "saya", "ay": "sayang", "ayank": "sayang", "b4": "sebelum",
        "bakalan": "akan", "bandes": "bantuan desa", "bangedh": "banget", "banpol": "bantuan polisi",
        "banpur": "bantuan tempur", "basbang": "basi", "bcanda": "bercanda", "bdg": "bandung",
        "begajulan": "nakal", "beliin": "belikan", "bencong": "banci", "bentar": "sebentar", "ber3": "bertiga",
        "beresin": "membereskan", "bete": "bosan", "beud": "banget", "bg": "abang", "bgmn": "bagaimana",
        "bgt": "banget", "bijimane": "bagaimana", "bintal": "bimbingan mental", "bkl": "akan",
        "bknnya": "bukannya", "blegug": "bodoh", "blh": "boleh", "bln": "bulan", "blum": "belum", "bnci": "benci",
        "bnran": "yang benar", "bodor": "lucu", "bokap": "ayah", "boker": "buang air besar", "bokis": "bohong",
        "boljug": "boleh juga", "bonek": "bocah nekat", "boyeh": "boleh", "br": "baru", "brg": "bareng",
        "bro": "saudara laki-laki", "bru": "baru", "bs": "bisa", "bsen": "bosan", "bt": "buat", "btw": "ngomong-ngomong",
        "buaya": "tidak setia", "bubbu": "tidur", "bubu": "tidur", "bumil": "ibu hamil", "bw": "bawa",
        "bwt": "buat", "byk": "banyak", "byrin": "bayarkan", "cabal": "sabar", "cadas": "keren",
        "calo": "makelar", "can": "belum", "capcus": "pergi", "caper": "cari perhatian", "ce": "cewek",
```

```
"cekal": "cegah tangkal", "cemen": "penakut", "cengengesan": "tertawa", "cepet": "cepat",
"cew": "cewek", "chuyunk": "sayang", "cimeng": "ganja", "cipika cipiki": "cium pipi kanan cium pipi kiri",
"ciyh": "sih", "ckepp": "cakep", "ckp": "cakep", "cmiiw": "correct me if i'm wrong", "cmpur": "campur",
"cong": "banci", "conlok": "cinta lokasi", "cowwyy": "maaf", "cp": "siapa", "cpe": "capek", "cppe": "capek",
"cucok": "cocok", "cuex": "cuek", "cumi": "Cuma miscall", "cups": "culun",
"curanmor": "pencurian kendaraan bermotor", "curcol": "curahan hati colongan",
"cwek": "cewek", "cyin": "cinta", "d": "di", "dah": "deh", "dapet": "dapat", "de": "adik",
"dek": "adik", "demen": "suka", "deyh": "deh", "dgn": "dengan", "diancurin": "dihancurkan",
"dimaafin": "dimaafkan", "dimintak": "diminta", "disono": "di sana", "dket": "dekat",
"dkk": "dan kawan-kawan", "dll": "dan lain-lain", "dlu": "dulu", "dngn": "dengan", "dodol": "bodoh",
"doku": "uang", "dongs": "dong", "dpt": "dapat", "dri": "dari", "drmn": "darimana", "drtd": "dari tadi",
"dst": "dan seterusnya", "dtg": "datang", "duh": "aduh", "duren": "durian", "ed": "edisi",
"egp": "emang gue pikirin", "eke": "aku", "elu": "kamu", "emangnya": "memangnya", "emng": "memang",
"endak": "tidak", "enggak": "tidak", "envy": "iri", "ex": "mantan", "fax": "facsimile",
"fifo": "first in first out", "folbek": "follow back", "fyi": "sebagai informasi", "gaada": "tidak ada uang",
"gag": "tidak", "gaje": "tidak jelas", "gak papa": "tidak apa-apa", "gan": "juragan", "gaptek": "gagap teknologi",
"gatek": "gagap teknologi", "gawe": "kerja", "gbs": "tidak bisa", "gebetan": "orang yang disuka",
"geje": "tidak jelas", "gepeng": "gelandangan dan pengemis", "ghiy": "lagi", "gile": "gila",
"gimana": "bagaimana", "gino": "gigi nongol", "githu": "gitu", "gj": "tidak jelas", "gmana": "bagaimana",
"gn": "begini", "goblok": "bodoh", "golput": "golongan putih", "gowes": "mengayuh sepeda",
"gpny": "tidak punya", "gr": "gede rasa", "gretongan": "gratisan", "gtau": "tidak tahu",
"gua": "saya", "guoblok": "goblok", "gw": "saya", "ha": "tertawa", "haha": "tertawa",
"hallow": "halo", "hankam": "pertahanan dan keamanan", "hehe": "he", "helo": "halo", "hey": "hai",
"hlm": "halaman", "hny": "hanya", "hoax": "isu bohong", "hr": "hari", "hrus": "harus",
"hubdar": "perhubungan darat", "huff": "mengeluh", "hum": "rumah", "humz": "rumah",
"ilang": "hilang", "ilfil": "tidak suka", "imho": "in my humble opinion", "imoetz": "imut",
"item": "hitam", "itungan": "hitungan", "iye": "iya", "ja": "saja", "jadiin": "jadi",
"jaim": "jaga image", "jayus": "tidak lucu", "jdi": "jadi", "jem": "jam", "jga": "juga",
"jgnkan": "jangankan", "jir": "anjing", "jln": "jalan", "jomblo": "tidak punya pacar",
"jubir": "juru bicara", "jutek": "galak", "k": "ke", "kab": "kabupaten", "kabor": "kabur",
"kacrut": "kacau", "kadiv": "kepala divisi", "kagak": "tidak", "kalo": "kalau", "kampret": "sialan",
"kamtibmas": "keamanan dan ketertiban masyarakat", "kamuwh": "kamu", "kanwil": "kantor wilayah",
"karna": "karena", "kasubbag": "kepala subbagian", "katrok": "kampungan", "kayanya": "kayaknya",
"kbr": "kabar", "kdu": "harus", "kec": "kecamatan", "kejurnas": "kejuaraan nasional",
"kekeuh": "keras kepala", "kel": "kelurahan", "kemaren": "kemarin", "kepengen": "mau", "kepingin": "mau",
"kepsek": "kepala sekolah", "kesbang": "kesatuan bangsa", "kesra": "kesejahteraan rakyat",
"ketrima": "diterima", "kgiatan": "kegiatan", "kibul": "bohong", "kimpoi": "kawin", "kl": "kalau",
"klianz": "kalian", "kloter": "kelompok terbang", "klw": "kalau", "km": "kamu", "kmps": "kampus",
"kmrn": "kemarin", "knal": "kenal", "knp": "kenapa", "kodya": "kota madya",
"komdis": "komisi disiplin", "komsov": "komunis sovyet", "kongkow": "kumpul bareng teman-teman",
"kopdar": "kopi darat", "korup": "korupsi", "kpn": "kapan", "krenz": "keren", "krm": "kirim","kt": "kita", "ktmu": "ketemu
"kyk": "seperti", "la": "lah", "lam": "salam", "lamp": "lampiran", "lanud": "landasan udara",
"latgab": "latihan gabungan", "lebay": "berlebihan", "leh": "boleh", "lelet": "lambat",
"lemot": "lambat", "lgi": "lagi", "lgsg": "langsung", "liat": "lihat", "litbang": "penelitian dan pengembangan",
```

```
"lmyn": "lumayan", "lo": "kamu", "loe": "kamu", "lola": "lambat berfikir", "louph": "cinta", "low": "kalau",
"lp": "lupa", "luber": "langsung, umum, bebas, dan rahasia", "luchuw": "lucu", "lum": "belum", "luthu": "lucu",
"lwn": "lawan", "maacih": "terima kasih", "mabal": "bolos", "macem": "macam", "macih": "masih",
"maem": "makan", "magabut": "makan gaji buta", "maho": "homo", "mak jang": "kaget", "maksain": "memaksa",
"malem": "malam", "mam": "makan", "maneh": "kamu", "maniez": "manis", "mao": "mau", "masukin": "masukkan",
"melu": "ikut", "mepet": "dekat sekali", "mgu": "minggu", "migas": "minyak dan gas bumi",
"mikol": "minuman beralkohol", "miras": "minuman keras", "mlah": "malah", "mngkn": "mungkin",
"mo": "mau", "mokad": "mati", "moso": "masa", "mpe": "sampai", "msk": "masuk", "mslh": "masalah",
"mt": "makan teman", "mubes": "musyawarah besar", "mulu": "melulu", "mumpung": "selagi",
"munas": "musyawarah nasional", "muntaber": "muntah dan berak", "musti": "mesti", "muupz": "maaf",
"mw": "now watching", "n": "dan", "nanam": "menanam", "nanya": "bertanya", "napa": "kenapa",
"napi": "narapidana", "napza": "narkotika, alkohol, psikotropika, dan zat adiktif ",
"narkoba": "narkotika, psikotropika, dan obat terlarang", "nasgor": "nasi goreng", "nda": "tidak",
"ndiri": "sendiri", "ne": "ini", "nekolin": "neokolonialisme", "nembak": "menyatakan cinta",
"ngabuburit": "menunggu berbuka puasa", "ngaku": "mengaku", "ngambil": "mengambil",
"nganggur": "tidak punya pekerjaan", "ngapah": "kenapa", "ngaret": "terlambat", "ngasih": "memberikan",
"ngebandel": "berbuat bandel", "ngegosip": "bergosip", "ngeklaim": "mengklaim", "ngeksis": "menjadi eksis",
"ngeles": "berkilah", "ngelidur": "menggigau", "ngerampok": "merampok", "ngga": "tidak", "ngibul": "berbohong",
"ngiler": "mau", "ngiri": "iri", "ngisiin": "mengisikan", "ngmng": "bicara", "ngomong": "bicara",
"ngubek2": "mencari-cari", "ngurus": "mengurus", "nie": "ini", "nih": "ini", "niyh": "nih", "nmr": "nomor",
"nntn": "nonton", "nobar": "nonton bareng", "np": "now playing", "ntar": "nanti", "ntn": "nonton",
"numpuk": "bertumpuk", "nutupin": "menutupi", "nyari": "mencari", "nyekar": "menyekar", "nyicil": "mencicil",
"nyoblos": "mencoblos", "nyokap": "ibu", "ogah": "tidak mau", "ol": "online", "ongkir": "ongkos kirim",
"oot": "out of topic", "org2": "orang-orang", "ortu": "orang tua", "otda": "otonomi daerah",
"otw": "on the way, sedang di jalan", "pacal": "pacar", "pake": "pakai", "pala": "kepala",
"pansus": "panitia khusus", "parpol": "partai politik", "pasutri": "pasangan suami istri", "pd": "pada",
"pede": "percaya diri", "pelatnas": "pemusatan latihan nasional", "pemda": "pemerintah daerah",
"pemkot": "pemerintah kota", "pemred": "pemimpin redaksi", "penjas": "pendidikan jasmani",
"perda": "peraturan daerah", "perhatiin": "perhatikan", "pesenan": "pesanan", "pgang": "pegang", "pi": "tapi",
"pilkada": "pemilihan kepala daerah", "pisan": "sangat", "pk": "penjahat kelamin", "plg": "paling",
"pmrnth": "pemerintah", "polantas": "polisi lalu lintas", "ponpes": "pondok pesantren", "pp": "pulang pergi",
"prg": "pergi", "prnh": "pernah", "psen": "pesan", "pst": "pasti", "pswt": "pesawat", "pw": "posisi nyaman",
"qmu": "kamu", "rakor": "rapat koordinasi", "ranmor": "kendaraan bermotor", "re": "reply", "ref": "referensi",
"rehab": "rehabilitasi", "rempong": "sulit", "repp": "balas", "restik": "reserse narkotika", "rhs": "rahasia",
"rmh": "rumah", "ru": "baru", "ruko": "rumah toko", "rusunawa": "rumah susun sewa", "ruz": "terus",
"saia": "saya", "salting": "salah tingkah", "sampe": "sampai", "samsek": "sama sekali", "sapose": "siapa",
"satpam": "satuan pengamanan", "sbb": "sebagai berikut", "sbh": "sebuah", "sbnrny": "sebenarnya",
"scr": "secara", "sdgkn": "sedangkan", "sdkt": "sedikit", "se7": "setuju", "sebelas dua belas": "mirip",
"sembako": "sembilan bahan pokok", "sempet": "sempat", "sendratari": "seni drama tari", "sgt": "sangat",
"shg": "sehingga", "siech": "sih", "sikon": "situasi dan kondisi", "sinetron": "sinema elektronik",
"siramin": "siramkan", "sj": "saja", "skalian": "sekalian", "sklh": "sekolah", "skt": "sakit",
"slesai": "selesai", "sll": "selalu", "slma": "selama", "slsai": "selesai", "smpt": "sempat", "smw": "semua",
"sndiri": "sendiri", "soljum": "sholat jumat", "songong": "sombong", "sory": "maaf", "sosek": "sosial-ekonomi",
"sotoy": "sok tahu", "spa": "siapa", "sppa": "siapa", "spt": "seperti", "srtfkt": "sertifikat",
```

```
"stiap": "setiap", "stlh": "setelah", "suk": "masuk", "sumpek": "sempit", "syg": "sayang", "t4": "tempat",
"tajir": "kaya", "tau": "tahu", "taw": "tahu", "td": "tadi", "tdk": "tidak", "teh": "kakak perempuan",
"telat": "terlambat", "telmi": "telat berpikir", "temen": "teman", "tengil": "menyebalkan", "tepar": "terkapar",
"tggu": "tunggu", "tgu": "tunggu", "thankz": "terima kasih", "thn": "tahun", "tilang": "bukti pelanggaran",
"tipiwan": "TvOne", "tks": "terima kasih", "tlp": "telepon", "tls": "tulis", "tmbah": "tambah",
"tmen2": "teman-teman", "tmpah": "tumpah", "tmpt": "tempat", "tngu": "tunggu", "tnyta": "ternyata",
"tokai": "tai", "toserba": "toko serba ada", "tpi": "tapi", "trdhulu": "terdahulu", "trima": "terima kasih",
"trm": "terima", "trs": "terus", "trutama": "terutama", "ts": "penulis", "tst": "tahu sama tahu",
"ttg": "tentang", "tuch": "tuh", "tuir": "tua", "tw": "tahu", "u": "kamu", "ud": "sudah", "udah": "sudah",
"ujg": "ujung", "ul": "ulangan", "unyu": "lucu", "uplot": "unggah", "urang": "saya", "usah": "perlu",
"utk": "untuk", "valas": "valuta asing", "w/": "dengan", "wadir": "wakil direktur", "wamil": "wajib militer",
"warkop": "warung kopi", "warteg": "warung tegal", "wat": "buat", "wkt": "waktu", "wtf": "what the fuck",
"xixixi": "tertawa", "ya": "iya", "yap": "iya", "yaudah": "ya sudah", "yawdah": "ya sudah", "yg": "yang",
"yl": "yang lain", "yo": "iya", "yowes": "ya sudah", "yup": "iya", "7an": "tujuan", "ababil": "abg labil",
"acc": "accord", "adlah": "adalah", "adoh": "aduh", "aha": "tertawa", "aing": "saya", "aja": "saja",
"ajj": "saja", "aka": "dikenal juga sebagai", "akko": "aku", "akku": "aku", "akyu": "aku",
"aljasa": "asal jadi saja", "ama": "sama", "ambl": "ambil", "anjir": "anjing", "ank": "anak", "ap": "apa",
"apaan": "apa", "ape": "apa", "aplot": "unggah", "apva": "apa", "aqu": "aku", "asap": "sesegera mungkin",
"aseek": "asyik", "asek": "asyik", "aseknya": "asyiknya", "asoy": "asyik", "astrojim": "astagfirullahaladzim",
"ath": "kalau begitu", "atuh": "kalau begitu", "ava": "avatar", "aws": "awas", "ayang": "sayang",
"ayok": "ayo", "bacot": "banyak bicara", "bales": "balas", "bangdes": "pembangunan desa", "bangkotan": "tua",
"banpres": "bantuan presiden", "bansarkas": "bantuan sarana kesehatan",
"bazis": "badan amal, zakat, infak, dan sedekah", "bcoz": "karena", "beb": "sayang", "bejibun": "banyak",
"belom": "belum", "bener": "benar", "ber2": "berdua", "berdikari": "berdiri di atas kaki sendiri",
"bet": "banget", "beti": "beda tipis", "beut": "banget", "bgd": "banget", "bgs": "bagus",
"bhubu": "tidur", "bimbuluh": "bimbingan dan penyuluhan", "bisi": "kalau-kalau", "bkn": "bukan", "bl": "beli",
"blg": "bilang", "blm": "belum", "bls": "balas", "bnchi": "benci", "bngung": "bingung", "bnyk": "banyak",
"bohay": "badan aduhai", "bokep": "porno", "bokin": "pacar", "bole": "boleh", "bolot": "bodoh",
"bonyok": "ayah ibu", "bpk": "bapak", "brb": "segera kembali", "brngkt": "berangkat", "brp": "berapa",
"brur": "saudara laki-laki", "bsa": "bisa", "bsk": "besok", "bu_bu": "tidur", "bubarin": "bubarkan",
"buber": "buka bersama", "bujubune": "luar biasa", "buser": "buru sergap", "bwhn": "bawahan", "byar": "bayar",
"byr": "bayar", "c8": "chat", "cabut": "pergi", "caem": "cakep", "cama-cama": "sama-sama",
"cangcut": "celana dalam", "cape": "capek", "caur": "jelek", "cekak": "tidak ada uang", "cekidot": "coba lihat",
"cemplungin": "cemplungkan", "ceper": "pendek", "ceu": "kakak perempuan", "cewe": "cewek", "cibuk": "sibuk",
"cin": "cinta", "ciye": "cie", "ckck": "ck", "clbk": "cinta lama bersemi kembali", "cmpr": "campur",
"cnenk": "senang", "congor": "mulut", "cow": "cowok", "coz": "karena", "cpa": "siapa", "gokil": "gila",
"gombal": "suka merayu", "gpl": "tidak pakai lama", "gpp": "tidak apa-apa", "gretong": "gratis", "gt": "begitu",
"gtw": "tidak tahu", "gue": "saya", "guys": "teman-teman", "gws": "cepat sembuh", "haghaghag": "tertawa",
"hakhak": "tertawa", "handak": "bahan peledak", "hansip": "pertahanan sipil", "hellow": "halo", "helow": "halo",
"hi": "hai", "hlng": "hilang", "hnya": "hanya", "houm": "rumah", "hrs": "harus",
"hubad": "hubungan angkatan darat", "hubla": "perhubungan laut", "huft": "mengeluh",
"humas": "hubungan masyarakat", "idk": "saya tidak tahu", "ilfeel": "tidak suka", "imba": "jago sekali",
"imoet": "imut", "info": "informasi", "itung": "hitung", "isengin": "bercanda", "iyala": "iya lah",
"iyo": "iya", "jablay": "jarang dibelai", "jadul": "jaman dulu", "jancuk": "anjing", "jd": "jadi",
```

```
"jdikan": "jadikan", "jg": "juga", "jgn": "jangan", "jijay": "jijik", "jkt": "jakarta", "jnj": "janji",
"jth": "jatuh", "jurdil": "jujur adil", "jwb": "jawab", "ka": "kakak", "kabag": "kepala bagian",
"kacian": "kasihan", "kadit": "kepala direktorat", "kaga": "tidak", "kaka": "kakak",
"kamtib": "keamanan dan ketertiban", "kamuh": "kamu", "kamyu": "kamu", "kapt": "kapten",
"kasat": "kepala satuan", "kasubbid": "kepala subbidang", "kau": "kamu", "kbar": "kabar",
"kcian": "kasihan", "keburu": "terlanjur", "kedubes": "kedutaan besar", "kek": "seperti", "keknya": "kayaknya",
"keliatan": "kelihatan", "keneh": "masih", "kepikiran": "terpikirkan", "kepo": "mau tahu urusan orang",
"kere": "tidak punya uang", "kesian": "kasihan", "ketauan": "ketahuan", "keukeuh": "keras kepala",
"khan": "kan", "kibus": "kaki busuk", "kk": "kakak", "klian": "kalian", "klo": "kalau", "kluarga": "keluarga",
"klwrga": "keluarga", "kmari": "kemari", "kmpus": "kampus", "kn": "kan", "knl": "kenal", "knpa": "kenapa",
"kog": "kok", "kompi": "komputer", "komtiong": "komunis Tiongkok", "konjen": "konsulat jenderal", "koq": "kok",
"kpd": "kepada", "kptsan": "keputusan", "krik": "garing", "krn": "karena", "ktauan": "ketahuan",
"ktny": "katanya", "kudu": "harus", "kuq": "kok", "ky": "seperti", "kykny": "kayanya", "laka": "kecelakaan",
"lambreta": "lambat", "lansia": "lanjut usia", "lapas": "lembaga pemasyarakatan", "lbur": "libur",
"lekong": "laki-laki", "lg": "lagi", "lgkp": "lengkap", "lht": "lihat", "linmas": "perlindungan masyarakat",
"lmyan": "lumayan", "lngkp": "lengkap", "loch": "loh", "lol": "tertawa", "lom": "belum", "loupz": "cinta",
"lowh": "kamu", "lu": "kamu", "luchu": "lucu", "luff": "cinta", "luph": "cinta", "lw": "kamu", "lwt": "lewat",
"maaciw": "terima kasih", "mabes": "markas besar", "macem-macem": "macam-macam", "madesu": "masa depan suram",
"maen": "main", "mahatma": "maju sehat bersama", "mak": "ibu", "makasih": "terima kasih", "malah": "bahkan",
"malu2in": "memalukan", "mamz": "makan", "manies": "manis", "mantep": "mantap", "markus": "makelar kasus",
"mba": "mbak", "mending": "lebih baik", "mgkn": "mungkin", "mhn": "mohon", "miker": "minuman keras",
"milis": "mailing list", "mksd": "maksud", "mls": "malas", "mnt": "minta", "moge": "motor gede",
"mokat": "mati", "mosok": "masa", "msh": "masih", "mskpn": "meskipun", "msng2": "masing-masing",
"muahal": "mahal", "muker": "musyawarah kerja", "mumet": "pusing", "muna": "munafik",
"munaslub": "musyawarah nasional luar biasa", "musda": "musyawarah daerah", "muup": "maaf", "muuv": "maaf",
"nal": "kenal", "nangis": "menangis", "naon": "apa", "napol": "narapidana politik", "naq": "anak",
"narsis": "bangga pada diri sendiri", "nax": "anak", "ndak": "tidak", "ndut": "gendut",
"nekolim": "neokolonialisme", "nelfon": "menelepon", "ngabis2in": "menghabiskan", "ngakak": "tertawa",
"ngambek": "marah", "ngampus": "pergi ke kampus", "ngantri": "mengantri", "ngapain": "sedang apa",
"ngaruh": "berpengaruh", "ngawur": "berbicara sembarangan", "ngeceng": "kumpul bareng-bareng",
"ngeh": "sadar", "ngekos": "tinggal di kos", "ngelamar": "melamar", "ngeliat": "melihat",
"ngemeng": "bicara terus-terusan", "ngerti": "mengerti", "nggak": "tidak", "ngikut": "ikut",
"nginep": "menginap", "ngisi": "mengisi", "ngmg": "bicara", "ngocol": "lucu", "ngomongin": "membicarakan",
"ngumpul": "berkumpul", "ni": "ini", "nyasar": "tersesat", "nyariin": "mencari", "nyiapin": "mempersiapkan",
"nyiram": "menyiram", "nyok": "ayo", "o/": "oleh", "ok": "ok", "priksa": "periksa", "pro": "profesional",
"psn": "pesan", "psti": "pasti", "puanas": "panas", "qmo": "kamu", "qt": "kita", "rame": "ramai",
"raskin": "rakyat miskin", "red": "redaksi", "reg": "register", "rejeki": "rezeki", "renstra": "rencana strategis",
"reskrim": "reserse kriminal", "sni": "sini", "somse": "sombong sekali", "sorry": "maaf", "sosbud": "sosial-budaya",
"sospol": "sosial-politik", "sowry": "maaf", "spd": "sepeda", "sprti": "seperti", "spy": "supaya",
"stelah": "setelah", "subbag": "subbagian", "sumbangin": "sumbangkan", "sy": "saya", "syp": "siapa",
"tabanas": "tabungan pembangunan nasional", "tar": "nanti", "taun": "tahun", "tawh": "tahu", "tdi": "tadi",
"te2p": "tetap", "tekor": "rugi", "telkom": "telekomunikasi", "telp": "telepon", "temen2": "teman-teman",
"tengok": "menjenguk", "terbitin": "terbitkan", "tgl": "tanggal", "thanks": "terima kasih",
"thd": "terhadap", "thx": "terima kasih", "tipi": "TV", "tkg": "tukang", "tll": "terlalu", "tlpn": "telepon",
```

```
        "tman": "teman", "tmbh": "tambah", "tmn2": "teman-teman", "tmph": "tumpah", "tnda": "tanda", "tnh": "tanah",
        "togel": "toto gelap", "tp": "tapi", "tq": "terima kasih", "trgntg": "tergantung", "trims": "terima kasih",
        "cb": "coba", "y": "ya", "munfik": "munafik", "reklamuk": "reklamasi", "sma": "sama", "tren": "trend",
        "ngehe": "kesal", "mz": "mas", "analisise": "analisis", "sadaar": "sadar", "sept": "september",
        "nmenarik": "menarik", "zonk": "bodoh", "rights": "benar", "simiskin": "miskin", "ngumpet": "sembunyi",
        "hardcore": "keras", "akhirx": "akhirnya", "solve": "solusi", "watuk": "batuk", "ngebully": "intimidasi",
        "masy": "masyarakat", "still": "masih", "tauk": "tahu", "mbual": "bual", "tioghoa": "tionghoa",
        "ngentotin": "senggama", "kentot": "senggama", "faktakta": "fakta", "sohib": "teman", "rubahnn": "rubah",
        "trlalu": "terlalu", "nyela": "cela", "heters": "pembenci", "nyembah": "sembah", "most": "paling",
        "ikon": "lambang", "light": "terang", "pndukung": "pendukung", "setting": "atur", "seting": "akting",
        "next": "lanjut", "waspadalah": "waspada", "gantengsaya": "ganteng", "parte": "partai", "nyerang": "serang",
        "nipu": "tipu", "ktipu": "tipu", "jentelmen": "berani", "buangbuang": "buang", "tsangka": "tersangka",
        "kurng": "kurang", "ista": "nista", "less": "kurang", "koar": "teriak", "paranoid": "takut",
        "problem": "masalah", "tahi": "kotoran", "tirani": "tiran", "tilep": "tilap", "happy": "bahagia",
        "tak": "tidak", "penertiban": "tertib", "uasai": "kuasa", "mnolak": "tolak", "trending": "trend",
        "taik": "tahi", "wkwkkw": "tertawa", "ahokncc": "ahok", "istaa": "nista", "benarjujur": "jujur",
        "mgkin": "mungkin"}

processeds = [] # wadah hasil

for text in data['tweet_akhir']:

    # Menghapus hashtag
    processed = re.sub(r'#\w+', '', text)

    # Menghapus mention
    processed = re.sub(r'@\w+', '', text)

    # Menghapus email
    processed = re.sub(r'[\w\.-]+@[\w\.-]+', '', text)

    # Menghapus angka
    processed = re.sub(r'\d+', '', text)

    # Menghapus alamat web
    processed = re.sub(r'http\S+', '', text)
    processed = re.sub(r'www\S+', '', text)

    # Cek stopword apa bukan
    words = processed.split()
    rfrm=[slangs[word] if word in slangs else word for word in words]
    processed= " ".join(rfrm)

    # hapus stopword
```

```
    factory = StopWordRemoverFactory()
    f = open("stopwords-tala.txt", "r") # stopword tambahan disimpan di txt
    more_stopword = [] #menambahkan stopword
    for line in f:
        stripped_line = line.strip()
        line_list = stripped_line.split()
        more_stopword.append(line_list[0])
    f.close()
    stopwords = factory.get_stop_words() + more_stopword
    temp = [t for t in re.findall(r'\b[a-z]+-?[a-z]+\b',processed) if t not in stopwords]
    processed = ' '.join(temp)

    # stemming
    stemmer = StemmerFactory().create_stemmer()
    processed = stemmer.stem(processed)

    #Subtitusing multiple spaces with single space
    processed = re.sub(r'\s+', ' ', processed, flags=re.I)

    processeds.append(processed)
```

In [72]:

```
processeds
```

Out[72]:

```
['rilis informasi kait prediksi',
 'cepat tangan tanggal covid',
 '',
 'bas ekosistem',
 'tangan sebar dasar provinsi status',
 'wilayah',
 'data uji tanggal',
 'kerah enam helikopter tangan darurat bencana banjir bandan',
 'sebar terap tanggal',
 'sukses korban',
 'bencana nasional nyawa',
 'kena bencana ajang citra wessss payah jal',
 'monggo waspada',
 'kemarin nurut mudik larang nurut pr',
 '',
 'kerah enam helikopter tangan darurat bencana banjir bandang tanah longsor landa kabupate',
 'akibat banjir anginkencang wilayah terak',
 'data baru angka korban dampak alami ubah',
 'ludah sembarang masa pandemi langsung dipenj',
 '',
```

```
'lipsinc',
'kunjung kanan tinjau jembatan runtuh akibat banjir bandang',
'bantu dana huni ribu rupiah keluarga ungsi rumah',
'distribusi bantu ringan beban warga dampak bencana',
'tetap status tanggap darurat tangan bencana banjir bandang longsor gelombang pasang',
'tanam pohon memperkua',
'dampak korban jiwa luka pengungs',
'data uji tanggal',
'cepat tangan tanggal',
'siap langkah antisipasi tular ungsi dampak banjir longsor',
'bagi donk data salindia',
'viralkan biar pusat perintah',
'terima ormas ludah muka seorg wanitaapalg saat pandemi covidjgnkan ludahkecipratan bolehmakan',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'kiri',
'cepat',
'',
'bencana sekitar link kasih',
'tropis picu bencana angin kencang tanah longsor banjir gelombang pasang wi',
'',
'',
'maaf harga sembilan bahan pokok sana lbh mahal drpd msih trbatas pasok',
'ludah sembarang masa pandemi lang',
'',
'alam hadir masyarakat siap survivor han',
'yes udan',
'harap inisiasi kerjasama ahli rancang mitigasi',
'sebar jadi bencana alam periode',
'sharing akun serta link berita',
'peta arah jauh',
'vaksinasi usia pondok gede or po',
'perintah kelola pandemi covid',
'kuat lemah',
'serius komitmen perintah upaya selenggara tanggulang bencana',
```

```
'lapor warga hilang akibat banjir bandang',
'kunjung jajar sangkut harap sege',
'bantu saudara bangsa tanah air kena bencana suku agama beda',
'mohon bantu cepat baik akses jalan mudah bantu masuk supply',
'',
'data jiwa men',
'',
'',
'korban dampak bencanaalam awat maksima',
'saudara bantu tolong nasib derita saudara daerah bencana',
'berat',
'tekan gonaku',
'',
'tanah pasca banjir tanah longsor',
'',
'',
'',
'',
'',
'',
'nomor tep hubung memgetahui kondisi terkin',
'',
'',
'ala traktor iya',
'data orang tewas ribu',
'dg tanam bibit poh',
'darah',
'',
'respon bantu ribet',
'',
'wes monggo sak kerso panjenengan',
'duka musibah saudara',
'manut wae awake sing ngatur mulih ketemu tuwo',
'',
'bencana nasional',
'',
'alaikumsala',
'tinggal edukasi mitigasi bencana drop siklonseroja',
'kewaspadaanya gala',
'ko hujan salah program kesiapsiagaan waspada bencana mana',
'',
'kesiapsiagaan waspada bencana wajib gala bentuk sekolah sungai kualitas',
'',
'awan potensi awan hujan rendah',
'',
```

```
'tanggal iya',
'',
'min website cek radar trouble gmn',
'rilis bibit siklon tropis dampak cuaca ekstrem satu potensi curah',
'malam ga larang',
'',
'sebar jadi bencana alam periode',
'dukacita',
'',
'status lembaga tingkat tri tri tanggulang bencana',
'ayo rachelvennya duta bencana idn doi responsive diba',
'peluang gak sukarelawan bencana support admi',
'sebar jadi bencana alam periode',
'',
'',
'update an cuaca esok rabu rilis iya min',
'deteksi bibit siklon tropis',
'mohon perhati saudara timur pandemi',
'korban tinggal org di-update an kurang jiwa bikin bingung',
'lingkung msih hujan ringan sekal',
'',
'sobatkriskres damping unjung',
'',
'',
'kalimat sumber',
'tuju banget pagi bikin utas',
'roto roto-roto banyu',
'malam sobatkriskes kembang tanggal pkl',
'',
'sumber data om',
'periode inidiplomasi',
'distribusi bantu logistik jalur laut',
'',
'',
'tebel songot',
'',
'mohon ditindaklanjuti gempa rusak berat smpai mnerima sdnkan rusak ri',
'jurang mabok sak polole iya longsorr awake tertawa',
'informasi longsoooor ndan',
'aplikasi windy',
'mhon ditindaklanjuti trkait bantu gempa rumah rusak berat',
'lapor warga tinggal dunia akibat ban',
'',
'teman teman tenaga pikir butuh saudara saudara dampak bencana',
```

```
 '',
 'be',
 'mudik wisata',
 '',
 '',
 'wilayah sumber',
 'santun tanggal',
 'santun tanggal',
 'sebar dasar provinsi status',
 'duka dalam bencana moga mentemen dib',
 'sebar dasar provinsi status',
 'pmi jateng nara sumber bentuk klaster logistik prov jateng selenggara',
 'pngolahan bulk vaksin menjd vaksin jadisesuai kaidah ditetapkn dida',
 'twitter kdatangan vaksin covid tahap',
 'detail sebar hadeh ajar hitung dah orang hilang',
 'data beda data daerah min',
 'gak iya isi benernya',
 '',
 'siklon tropis dahsyat global warming ga mitigasi',
 '',
 '',
 'pr',
 '',
 'prayfor',
 'prayfo',
 'malam sobatkriskes kembang tanggal pkl',
 '',
 'for bpbd prayforntt nusatenggaratimur siapuntukselamat',
 'ketik takut tangkap kang bakso bawa',
 'ada kluster',
 'duka dalam bencana moga mentemen',
 '',
 'lapor orang tinggal dunia akibat bencana alam banjir tanah longsor',
 'tanah pasca banjir tanah longsor',
 '',
 'cari selamat mangsa bencana ih',
 '',
 'yah law kma',
 '',
 'peta sebar',
 'mudik pilih kepala daerah gak',
 'pusat tutup mata prayf',
 'larang lahdiatur orang tua kampung sakit nunggu anak-ana',
 'bernamadotcom cari selamat mangsa bencana',
```

```
'cari selamat mangsa bencana',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'',
'lereng gunung',
'',
'',
'daerah wisata',
'wilayah',
'',
'sintas banjir bandang banjirbandang',
'tim pagi merespon jadi banjir bandang',
'ubah judul kali luas cakup bencana sadar iya perintah',
'',
'rasa salah rasa panjat doa ulur tangan un',
'',
'',
'baju orange rompi unjung posko ungsi dapur',
'data uji tanggal',
'memange mudik tekan naik covid pariwisata ngak ruwet mudik dil',
'kondangan gapapa syawal hajat kondangan',
'libur buka sedia',
'menengo wis',
'',
'kondangan ketua gak hand sanitizer klau gak slh or bagiw masker',
'data rekap tinggal yes bikin',
'pers',
'tinggal orang luka orang hilang orang rumah timbun',
'dapur layan butuh pangan ungsi dampak banjir bandang',
'korban tinggal simpang-siur laman resmi korban tinggal te',
'org sulsel kakak daerah',
'terang pers konferens',
'',
'temennya org keluarga daerah',
'list nama korban tinggal proses cari',
'terang pers',
```

```
'iya',
'sobatkriskres damping me',
'do smga sabar',
'testing lapor antigen negatif langsung konfirm',
'',
'data orang tewas rib',
'status bencana bencana',
'',
'dgn hari korban hidup',
'lihat listnya temenku cari keluarga',
'cepat tangan tanggal',
'korban tinggal kayak ketuker media korban tin',
'pilih ungsi',
'pagi ajar bareng warga merapipemdes balerantepokdarwis cerita tangguh',
'sebar jadi bencana alam periode',
'',
'',
'kirain mbak sndri ngalami bay',
'orang don liat',
'duga',
'warga dampak bencana tambah data',
'isu bohong nih video lokasi cc',
'chek dusun',
'saudara masyarakat susah milu',
'lapar bencana nasional',
'',
'dampak',
'testing lapor antigen negatif langsung kon',
'data warga ungsi or',
'',
'tetiba sales perintah hahahaha',
'',
'gtu deh tunggu aksi iya om ayo',
'hujan jam min',
'yah untuk giat',
'emang',
'tanggal ht',
'perintah gembar gembor',
'',
'tinjau ulang kait',
'iya om demo om anti pand',
'',
'tim sehat',
'',
```

```
'tuh contoh jabat petinggi negar',
'serah deh om om plus om',
'',
'',
'',
'',
'',
'emang covid ndk tinggal',
'pikir balik tuju tetap status bencana nasional',
'cat lumpuh total',
'',
'status bencana ga tuh',
'tinggal beneran',
'virus biasaaa',
'masuk keb',
'sholat tarawih boleh wkwk gejala merokethati hati yakkematian protap co',
'gak',
'tugas informasi jujur masyaraka',
'nyalahin nakes',
'banjir bandang',
'sosok tanggung bencana banjir bandang',
'paham sedih alami saudara akibat dampak timbul bencana',
'nama pribadi rakyat dukacita dalam korban tinggal dunia mus',
'telinga hidung',
'',
'baju orange gelar konferensi pers kait pe',
'gerak bantu',
'disalahin',
'atensi',
'adil sulit lawan',
'potensi barat',
'konstitusi',
'',
'',
'bhw kunjung lokasi titik bencana',
'liput investigasi ada masker palsu',
'kab lembata camat rusak parah longsor',
'',
'',
'peran perintah daerah informasi',
'desa pulau',
'ajak masyarakat terap',
'indonesia doank kendali bangkok ampe ratus',
'tanggal puku',
'',
```

```
 '',
 '',
'tanggal',
'cuaca min',
'periode inidiplomasi',
'rangkai siklon tropis pesan terjemah bahasa awam konsumsi orang kampung',
'serta rombong lanjut jalan',
'sebar jadi bencana alam periode',
'',
'maaf orang awam sdh swab vaksin pakai masker jaga jarak',
'',
'kirim bantu ringan beban warga dampak bencan',
'teman ubud hilang kontak papa mohon retweet te',
'tugas moga sllu',
'luncur alat deteksi tsunami karya anak bangsa',
'ga covid min',
'progo hujan min',
'timpa duka bala bencana jauh',
'',
'banten do',
'concern mimin mitiga',
'malam sobatkriskes kembang tanggal pkl',
'gak eneng turun bop',
'jalan status capai tingkat pulih hai',
'putih rompi membe',
'instruksi kerah helikopter helikopter jangkau distribusi logistik bebe',
'hapus cuit gambar yaa salah',
'langkah signifikan abang tinggal daya',
'rangkai siklon tropis pesan terjemah bahasa awam konsumsi orang kampung',
'korporat manajemen kontak',
'kolaborasi',
'sih dua libat rencana mitig',
'bilangnyamau wisata kampung halaman',
'bantu please mohon up',
'jalan status capai tingkat',
'duka bencana alam tel korban hadap cuaca ekstrem',
'batas sebar informasi cuaca hari',
'sosok tanggung bencana banjir bandang',
'nuwun le',
'ajak komponen masyarakat partisipasi aktif',
'',
'',
'ngeri banget ih angin',
'',
```

```
 '',
 'salah tweet mutual inti tugas',
 'barulbh sektr bwh hari sekt',
 'ta sen',
 'perintah menteri menteri sehat polri seger',
 'gak mobil bantu',
 'covid in ukur',
 'nyata resmi perintah daerah bencana',
 'alamiah multi',
 'santun tanggal',
 'daerah wisata',
 'santun tanggal',
 'tenda ramah perempuan anak bantu prempuan anak selamat ancam leceh',
 '',
 'terima lapora',
 'lihat update moga turun hari',
 'salah inovasi dukung',
 'galau bro tf iya kirim',
 'terbang salah anak usaha jalan misi manusia me',
 'lemah kuat hilang tular tahap',
 'anjng sinte juragan',
 'roboh terjang banjir',
 'mas alam urusin bansos',
 'update',
 'wilayah sumber',
 'juang pulih cepat informasi lokasi dampak',
 'banget deh ber ga bayar',
 'kemarin beneran cuman rekap data tinggal',
 'lambat tular',
 'perintah menteri menteri sehat polri un',
 'gainget ser pas galau mesennya gua',
 'udh kelar kerja sesuai sop kerja tarung nyawa gaji ntarin melulu',
 'kirim bantu',
 'om',
 'moga mati hari tekan digit digit',
 'kasih',
 'catat wil dampak banjir adl',
 'lapor warga tinggal dunia akibat banjir bandan',
 'ditangkep',
 'samping korban jiwa banjir bandang akibat jembatan putus puluh rumah warga timbun lumpur',
 'banjir bandang landa desa camat',
 'tambah konfirmasi positif drastis',
 'terjang saudara saudara sana',
 'informasi warga hilang akibat banjir bandang har',
```

```
'ber',
'',
'tambah konfirmasi positif drastis yait',
'malam sobatkriskes kembang tanggal pkl',
'',
'sekitar ungsi akibat air laut',
'lereng gunung',
'',
'',
'daerah wisata',
'sama dn sebab jatuhx korban jiwa',
'wilayah',
'informasi persentase pasien tinggal covi',
'',
'jadwal kunjung kabupaten seberang pulau alam',
'buka sekolah tutup',
'sana',
'perintah daerah tetap tanggap darurat perintah pusat turun lapang',
'kasih kakak muslim pe',
'dampak upaya mitigasi',
'',
'',
'',
'',
'',
'',
'',
'',
'data banjir longsor tel korban tinggal jiwa luka berat jiwa luka',
'updatenya iya pa min',
'citra radar mimin',
'',
'sakebon',
'instruksi mitigasi',
'alam silih ganti peduli sosial lupa kondisi',
'rencana mitigasi lokal',
'iring rombong',
'letak geografis bencana barangkali bentuk',
'oke kelihatanya depan cuaca',
'maaf moga normal iya rezeki hilang ganti kali kali lipat',
'kena phk maksud',
'auk ah',
'gtu iya',
'',
'pandemi pandemi palsu',
```

```
'mbak',
'gak hilang sumber hasil dr pandemi',
'udahan iya pandemi',
'gtuuu iya posisi baiqlaaah something new',
'ganti',
'kak',
'perintah tangkap dg hadir jokowi prabowo',
'informasi sederhana mudah erti masyarakat pakai bahasa ilmiah bikin ruwet',
'saintifik emang banget miskomunik',
'prediksi tropical cyclone saya',
'hmmmm perintah laku fearmongering',
'jabarnews beritalokal',
'gak nakes lakuin fearmongering kritisi nak',
'otoritas seberang ingat larang layar faktor cuaca',
'iya iya dooong mending ramein kesana banding nakes rakyat',
'ajar artis indo sih tertawa maju beda serius beda pola pikir iya',
'iya',
'gak kuasa bijak mbak coba langsung colek',
'kasih kakak',
'larang wisata buka',
'acara',
'duka om gerak cepat',
'',
'paham sedih alami saudara akibat dampak timbul bencana',
'bencana cuaca prediksi',
'semangat sehat',
'nama pribadi rakyat dukacita dalam korban tinggal dunia da',
'data banjir longsor tel korban tinggal jiwa luka berat jiwa',
'jam rumah ku puncak nya parah',
'jam rumah ku puncak nya parah',
'pdate',
'',
'',
'',
'',
'',
'kerja dinas emg kytu gel kerja pokok',
'lengkap daerah bancana mana da',
'lapor warga tinggal dunia akibat banjir bandang',
'tugas pasti siyap aman tolong korban upaya men',
'ngak sibuk saksi nikah',
'prokes hilang',
'cepat',
'jadi sabtu ko lapor tanggap duluan kontrol',
```

```
 '',
 '',
'lelah nya beliau serta rombong semangat masyarakat manusia',
'minggu april kak',
'',
'logistik cepat cc',
'',
'semangat coba ditweet hasil',
'kasih kak video kendala jaring gagal',
'minggu landa banjir dmana longsor padi sawah rendam jagu',
'',
'cpt tangan manusia',
'kasih kak video kendala jaring gagal ditweet kak',
'tetap status tanggap darurat',
'rencana mitigasi',
'banget ekslusif',
'dampak upaya mitigasi',
'safe berkat',
'keluarga',
'segitu kondangan twit jam mala',
'tambah personil latih alat le',
'moga sana baik',
'moga cepat bikin',
'kasih reportase video like pe',
'lapor banjir bandang wilayah setemp',
'bencana',
'kasih kakak musli',
'informasi grafis distribusi gempa susul lombok catat update',
'hadir sperti acara ramai kmaren rakyat',
'mohon tanggulang saudara kondisi',
'bhw bencana gembel jalan selesai asyi',
'bantu rt nya',
'terima lapor perintah pelaksanaaan tangan damp',
'daerah kemarin siang cma personilnya dgan alat batas',
'perintah daerah tetap tanggap darurat perintah pusat turun lapang',
'simanis jembatan iya abang',
'gerak sob mobilisasi segera',
'jadwal kunjung kabupaten seberang pulau jug',
'sana',
'siang min rutin update',
'dengar do',
'bencana henti perintah cepat ambil langkah bantu pulih',
'bencana henti perintah cepat ambil langkah bantu pulih',
'doang yes rekap banten',
```

```
'duka bencana alam tel korban hadap cuaca ekstrem',
'bantu',
'',
'selamat saudara',
'hanyut rumah warga video amatir warga',
'',
'kepala hormat parah gak',
'tindaklanjuti iya perihal lepas media coverage sa',
'segera tangan bencana',
'pagi tolak',
'pdate',
'pagi ngingetin gaji transfer',
'nyata kondisi teman sana sebar informasi bantu',
'',
'tangan cepat koordinasi sgera btindak cepat lancar',
'',
'jembatan rumput',
'sih nikah seleb gercep',
'landa saat memfo',
'data warga ungsi',
'maaf',
'terhormatkiranya aspirasi masyarakat daerah moga tangg',
'baik saudara moga lekas baik ada',
'batu bandang',
'temu pusat informasi wilayah pusat tenda ungsi',
'halo mohon bantu',
'duka',
'prihatin',
'gercep',
'',
'letak geografis bencana barangkali bentuk',
'',
'',
'',
'data baru korban warga dampak akibat banjir stake holder ter',
'dibantuuuuu',
'teman ubud hilang kontak papa mohon retweet',
'',
'org org pinter republik si kusnandi vaksin',
'temu pusat informasi wilayah pusat tenda pengun',
'pusat tutup mata',
'kirim bantua',
'satu doa timpa bencana alam moga tinggal ampun dosa te',
'satu doa timpa bencana alam moga tinggal ampun dosa dib',
```

```
'ini sdh',
'nih kondangan ape',
'masyarakat ambil jurus meteorologi klimatologi baca eksklusif',
'bantu',
'dmna butuh bantu',
'selat',
'badai lekas bantu saudara',
'baik timpa musibah bencana alam kekua',
'turun jam wilayah sebab bendung',
'cepat gerak garda sdh bubar',
'',
'',
'rekam bibit badai seh',
'denga',
'lokasi bencana tanah gerak',
'terhormatkiranya aspirasi masyarakat daerah semo',
'kirim bantu korban banjir makan oba',
'tambah informasi bibit siklon selatan jawa gbr te',
'dekat bantu perahu karet',
'jurai datu sriwijaya hubung kerabat',
'',
'manusia',
'kirim bantu korban banjir makan',
'dmna butuh bantu',
'putus hubung jalan',
'silih ganti',
'prihatin musibah cuaca ekstrem wilayah tangan hadir menolo',
'tempat antisipasi kali',
'peta sebar',
'',
'kasih iya',
'prihatin musibah cuaca ekstrem wilayah tangan hadir',
'',
'hadir lokasi bencana sm sperti mreka hadir acara nikah kmar',
'unsur kait pasti sua',
'imbau waspada potensi cuaca ekstrem picu bencana hidrometeorologi pek',
'kemarin sore bang kakak perempuan tawar panas',
'pagi minum kopi',
'pagi bang',
'bhw bencana gembel jalan selesai',
'bencana jadi takut masayarakat',
'sana lindung tolong',
'bantu please mohon up',
'situation report flash update badai siklon tropis',
```

```
'dorong hidup',
'waspada moga lindung',
'yoh terang yukk',
'',
'semenjak krisis situasi gambar',
'cepat tangan tanggal',
'lho andal media tolak ukur baca lapor situasi',
'situation report flash update badai siklon tropis',
'lapor banjir bandang korban tinggal hilan',
'aksi timpa',
'presiden habis kenyang',
'masuk kategori bencana nasional',
'',
'donasi',
'data mati banten jiwa',
'bencana tangan umtuk',
'kabupaten kota jalan',
'',
'tinggal status cari',
'biar daerah dampak bencana',
'banten kayak data hari setor ga masuk akal positif',
'bawa jalur hukum cuman',
'hasil evak mas',
'kabupaten ga informasi korban tinggal hilang',
'lapor banjir bandang korban tinggal',
'',
'salur bantu tolong suadara juang lewat musibah tangan',
'ingat usaha koordinasi lain upaya mitigas',
'tinggal',
'darurat',
'bag selatan harap status kait cuaca maupu',
'data uji tanggal',
'baca orang tinggal daerah antisipatif kurang korban jiwa',
'',
'',
'kab kota landa bencana korban jiwa orang',
'duka musibah banjir bandang doa moga',
'for',
'teman area malam-subuh kati bantu ju',
'duka musibah flores',
'serah tetap status darurat bencana daerah',
'',
'duka musibah saudar',
'',
```

```
'periode inidiplomasi',
'wainnailaihi rojiun ketemu',
'kasih abg',
'',
'maksud datang kenal nomor tetap',
'iya bebas ga masker',
'nilai milik komitmen kurang resiko bencana ajar peng',
'ajar langsung jadi standar pelay',
'bicara bangun komitmen kepala',
'definisi atur sich',
'sana sgr mdpt bantu evakuasi dg terima ksih',
'nyata status darurat bencana daer',
'malam sobatkriskes kembang tanggal pkl',
'',
'positif specimen tabel kolom',
'',
'jalan sebar dasar provinsi status',
'bantu presiden',
'tinggal orang luka orang hilang orang rumah tert',
'selamat malam maaf',
'',
'duka prihatin bencana timpa timur baik',
'gitu iya ikutin saran mas',
'',
'peta sebar',
'',
'tewas banjir bandang orang',
'bantu pusat daerah keteter',
'dampak bencana tangan',
'kir',
'atas',
'ingat usaha koordinasi lain upa',
'cebong gerak medsos bukti',
'pantes udan angine mrene',
'dampak menginventa',
'loh tengok engga kasi bu mensos arah tanggap darurat bencana',
'dampak',
'dr jogja td tenan',
'luka',
'hujan deras',
'',
'bag selatan harap status kait cuaca',
'adil sulit lawan',
'jam hujan deras plus angin',
```

```
'kepung pesta',
'santun tanggal',
'bantu pus',
'sakjane istilah ingat kuwi sing kepriye yo sak',
'tanggal ht',
'santun tanggal',
'dlm slalu waspada',
'update camat',
'saat turun banten',
'turu',
'informasi wilayah tangerang kabupaten tangsel camat',
'nih izin ramai',
'spesimen ambil hasil tracing suspect test man',
'inisiasi canang tanggal',
'himbauan',
'kota ancur-ancuran kena banjir lain mati orang',
'virus pindah banten',
'',
'moga surut banjir nya',
'salah tuh',
'malam sobatkriskes kembang tanggal pkl',
'kondisi mobil bawa banjir',
'banten sinkronisasi data data tingkat kota provinsi emang suka beda',
'bandang',
'tinggal suka gak update data tingkat kota data tingkat nasional',
'tiru malaysia sembuh korona',
'aneh emang data sembuh tinggal suka gak update kaya',
'',
'terapi khusus jimat',
'turun covid hari aktif',
'indonesia tiru negara tambah sembuh',
'bantu up moga menindaklanjuti',
'data mati keri',
'',
'moga kembang daerah',
'data pkembangan covid',
'deh kaya masuk akal banget sembuh langsung',
'updatetannya',
'terjawan kurang angka kematia',
'mati',
'setor data',
'hujan reda kerja bakti lokasi pondok',
'habis pres',
'pasien korona sana terapi khusus langsung sembuh singkat correct me if wrong',
```

```
'tinggal angka klarifikasi',
'min hujan extreme',
'',
'dasar lapor banjir bandan',
'aneh aktif an angka sembuh hari an',
'gaung',
'lonjak',
'malang selatan',
'korban',
'laut',
'malam pulang mancing pantai',
'libur yow',
'kab kota landa bencana korban jiwa orang',
'hitung situ adl tjd laut pulau dg',
'roboh terjang banjir',
'wilayah sumber',
'kangen tuh min',
'maturnuwun',
'min',
'',
'luka',
'update lambat',
'',
'banget tinggi so far aja',
'kondisi',
'',
'iya asbun',
'sore tlg update citra',
'paham',
'banget tinggi so far',
'ampungak tega lihat anak bang gk',
'bencana banjir tanah longso',
'dr nih badai',
'minggu landa banjir dmana longsor padi sawah rendam',
'sekitar ungsi akibat air laut',
'min badai min',
'cepat tangan tanggal',
'lambat update',
'',
'',
'',
'',
'',
'',
```

```
 '',
 '',
'update deras lho',
 '',
'webupdate citra radar telatgmn',
'update badai bantul',
'landa hujan badai hujan lebat henti serta angin kencang saudara',
'lapor bencana medsos biasa cc gak viral',
'landa hujan badai hujan lebat henti serta angin kencang sauda',
'to min butuh nih manual udh',
'iyaah',
'bandang terjang camat sebab or',
'landa hujan badai hujan lebat henti serta angin kencang masyar',
'rakyat',
'korban temu cepat',
'kabar',
'korban duka cita',
 '',
'darurat daerah panik',
 '',
'dikit president',
'bayar gaji bahagia',
 '',
'lapor bencana posting aja',
'tangan',
 '',
'kondisi mobil bawa banjir',
'genomic surveillance varian hati',
'bupati president daerah',
'perintah kirim dampak to',
'desa beda',
 '',
'darah',
'bencana kini trans kabupaten',
'dukung tangan tanggal',
 '',
 '',
'sebar virus corona kendali teriak patuh',
'rawan gabung bersih puing lumpur pascabanjir bandang',
'arah',
'bentuk implementasi',
'iya min',
'kupang hujan banjir kak',
'kar',
```

```
'vaksin aneh org islam tolak vaksin gil',
'',
'periode inidiplomasi',
'tanggal ht',
'rakyat guling kuasa tindak lanjut',
'giat cek lokasi tangan darurat bencana jembatan putus kaligintung batuagung',
'terima kunjung',
'sebar jadi bencana alam periode',
'data uji tanggal',
'malam sobatkriskes kembang tanggal pkl',
'',
'moga pulih',
'sebar dasar provinsi',
'mang bijak daerah lintas menteri lembaga unsur bersa',
'cancel layar cc wind speed gin walah hyuuung',
'holopiscom',
'',
'nonton rmh mngurangi mobilitas rumah',
'progres tangan jalan amblas kali bakung maret jam monggo suportnya',
'progres tangan jalan amblas kali bakung maret jam monggo suportnya',
'pandemi covid live ta',
'banner silah guna daerah intsnasinya',
'upaya kontinyu giat tangan darurat bencana jalan amblas dalam meter jalan',
'update kemarin spesimen kumulatif umum',
'update kemarin spesimen kumulatif umum salah seh',
'update kemarin update',
'',
'korban jiwa peristiwa jt',
'laksana giat cari korban tenggelam',
'giat cek lokasi tangan darurat bencana jembatan putus kaligintung batuagung',
'dlm cari korban tenggelam',
'laksana giat cari korban tenggelam',
'angka kemarin',
'gotong royong rawan wedi chah kali pemdes rawan sukorejo tanam ribu bibi',
'',
'santun tanggal',
'sebar dasar provinsi',
'santun tanggal',
'',
'rumah rendam banjir sentimeter',
'antisipasi january lobang kentut bicara',
'',
'',
'terjang camat april',
```

```
'mah baik tangan corona pakai konferensi pers',
'wilayah sumber',
'sdh dosis kabupaten diri gurah vaksinasi sm usia gmn',
'',
'',
'',
'',
'',
'',
'',
'ramai',
'',
'',
'',
'ngalahin jekarte',
'reage',
'',
'perhati tuh nya',
'klambu banjir',
'malam sobatkriskes kembang tanggal pkl',
'',
'',
'',
'min informasi hujan skrg dongg rumah deres bgt',
'',
'cuti po gak update prakira haria',
'',
'makan sesuai gol',
'',
'banjir nich min',
'update kondisi cuaca baru min maturnuwun',
'',
'',
'',
'',
'',
'',
'',
'',
'update min deres plus angin',
'web update situasi siklon kini bbrp informasi ditampilk',
'',
'angin mobat mabit panasnyaaa cetarrrrr tarrrrrr kipasss berasaaaaa belikan aceeee minnn',
'',
'lahan makam habis melulucepat habismasih lahan dasar',
'pasti turun turun tes buka donk data sebar distribusi tes pcr',
'minkalau surat antigen negatif genose lg',
'salur bantu masyarakat yng dampak banjir bandang puluh',
'darah',
'',
```

```
'teman kenal calon mahasiswa jurus',
'peta sebar',
'',
'suplai bantu mrk minggu dpn teduh',
'',
'pantau visual batas keluar awan panas pa',
'serah bantu',
'',
'kiri',
'rapat dorong bentuk kuat tangan mobilitas',
'tanggal',
'',
'kasih dikau koreksi rangkum',
'',
'estimasi maret april',
'ingat wafat raya pandu jalan member',
'sebar jadi bencana alam periode',
'cepat tangan tanggal',
'bantu tanggulang bantu',
'gunung dengar',
'salah kota tetap banguna',
'',
'cuaca meleset lelah',
'kurang',
'data uji tanggal',
'rumah rendam air cm unit rumah rusak unit kantor sarana pendidi',
'malam sobatkriskes kembang tanggal pkl',
'bandang malam picu tinggi intensitas hujan alih fungsi',
'adl sebab mati tinggi via',
'periode inidiplomasi',
'tanggal',
'peta sebar',
'vaksin agam',
'',
'santun tanggal',
'masuk wajib kali tes usap karantina',
'kiri arah',
'dasar https',
'gotong royong rawan wedi chah kali pemdes rawan sukorejo tanam ribu',
'sebar dasar provinsi status',
'malam sobatkriskes kembang tanggal pkl',
'',
'jalan tol macet gak tuh tugas patroli random cek',
'',
```

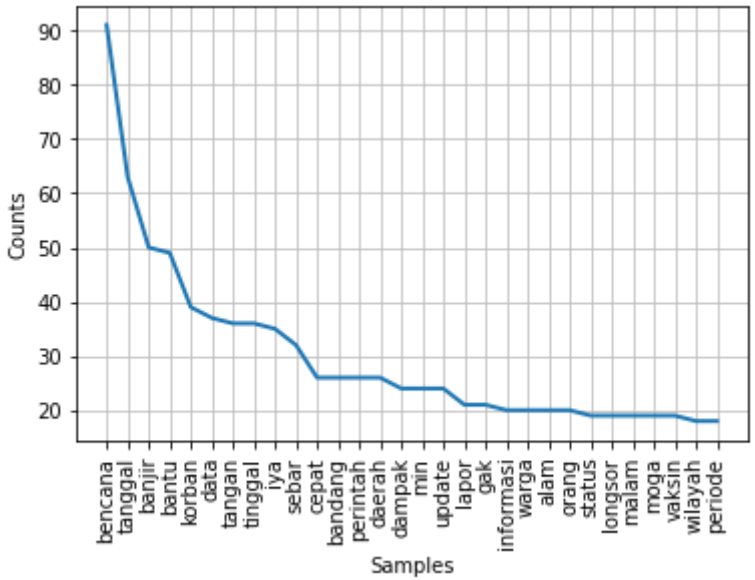
```
 '',
 'https',
 'min iya usia dahulu vaksinasi banding remaja usia',
 'dasar',
 'temu positif repatriasi masuk wilayah',
 'mohon ditindaklanjuti besok universitas laksana silaturahmi mahasiswa yan',
 '',
 '',
 '',
 '',
 '',
 '',
 '',
 ...]
```

Bag of Words dalam histogram

In [73]:

```
from nltk import FreqDist
import matplotlib.pyplot as plt

filtered = []
for string in processeds:
    filtered.extend(string.split())
fdist = FreqDist(filtered)
fdist.plot(30, cumulative=False)
plt.show()
```



In [74]:

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(processeds)

print('Jumlah fitur : ', len(vectorizer.get_feature_names()))
print(X.shape)
print(X.toarray())
```

```
Jumlah fitur :  1562
(1281, 1562)
[[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. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

In [75]:

```
pip install sklearn-pycrfsuite
```

```
Collecting sklearn-pycrfsuiteNote: you may need to restart the kernel to use updated packages.
  Using cached sklearn_pycrfsuite-0.4.0-py2.py3-none-any.whl
```

```
Requirement already satisfied: six in c:\users\user\anaconda3\lib\site-packages (from sklearn-pycrfsuite) (1.16.0)
Requirement already satisfied: tqdm>=2.0 in c:\users\user\anaconda3\lib\site-packages (from sklearn-pycrfsuite) (4.62.3)
Requirement already satisfied: tabulate in c:\users\user\anaconda3\lib\site-packages (from sklearn-pycrfsuite) (0.9.0)
Collecting python-crfsuite-extension
  Using cached python-crfsuite-extension-0.9.7.tar.gz (485 kB)
Requirement already satisfied: colorama in c:\users\user\anaconda3\lib\site-packages (from tqdm>=2.0->sklearn-pycrfsuite) (0.4.4)
Building wheels for collected packages: python-crfsuite-extension
  Building wheel for python-crfsuite-extension (setup.py): started
  Building wheel for python-crfsuite-extension (setup.py): finished with status 'error'
  Running setup.py clean for python-crfsuite-extension
Failed to build python-crfsuite-extension
Installing collected packages: python-crfsuite-extension, sklearn-pycrfsuite
    Running setup.py install for python-crfsuite-extension: started
    Running setup.py install for python-crfsuite-extension: finished with status 'error'


  ERROR: Command errored out with exit status 1:
   command: 'C:\Users\USER\anaconda3\python.exe' -u -c 'import io, os, sys, setuptools, tokenize; sys.argv[0] = '"'"'C:\\Users\\US
ER\\AppData\\Local\\Temp\\pip-install-o7lavm5f\\python-crfsuite-extension_72ec53254bbc4dc5bd29134ce797beca\\setup.py'"'"'; __file_
_='"'"'C:\\Users\\USER\\AppData\\Local\\Temp\\pip-install-o7lavm5f\\python-crfsuite-extension_72ec53254bbc4dc5bd29134ce797beca\\se
tup.py'"'"';f = getattr(tokenize, '"'"'open'"'"', open)(__file__) if os.path.exists(__file__) else io.StringIO('"'"'from setuptool
s import setup; setup()'"'"');code = f.read().replace('"'"'\r\n'"'"', '"'"'\n'"'"');f.close();exec(compile(code, __file__, '"'"'ex
ec'"'"'))' bdist_wheel -d 'C:\Users\USER\AppData\Local\Temp\pip-wheel-etppiww_'
       cwd: C:\Users\USER\AppData\Local\Temp\pip-install-o7lavm5f\python-crfsuite-extension_72ec53254bbc4dc5bd29134ce797beca\
  Complete output (12 lines):
  running bdist_wheel
  running build
  running build_py
  creating build
  creating build\lib.win-amd64-3.9
  creating build\lib.win-amd64-3.9\pycrfsuite
  copying pycrfsuite\_dumpparser.py -> build\lib.win-amd64-3.9\pycrfsuite
  copying pycrfsuite\_logparser.py -> build\lib.win-amd64-3.9\pycrfsuite
  copying pycrfsuite\__init__.py -> build\lib.win-amd64-3.9\pycrfsuite
  running build_ext
  building 'pycrfsuite._pycrfsuite' extension
  error: Microsoft Visual C++ 14.0 or greater is required. Get it with "Microsoft C++ Build Tools": https://visualstudio.microsof
t.com/visual-cpp-build-tools/
  ----------------------------------------
  ERROR: Failed building wheel for python-crfsuite-extension
    ERROR: Command errored out with exit status 1:
     command: 'C:\Users\USER\anaconda3\python.exe' -u -c 'import io, os, sys, setuptools, tokenize; sys.argv[0] = '"'"'C:\\Users
\\USER\\AppData\\Local\\Temp\\pip-install-o7lavm5f\\python-crfsuite-extension_72ec53254bbc4dc5bd29134ce797beca\\setup.py'"'"'; __f
ile__='"'"'C:\\Users\\USER\\AppData\\Local\\Temp\\pip-install-o7lavm5f\\python-crfsuite-extension_72ec53254bbc4dc5bd29134ce797beca
```

```
\\setup.py'"'";f = getattr(tokenize, '"'"'open'"'", open)(__file__) if os.path.exists(__file__) else io.StringIO('"'"'from setup
tools import setup; setup()'"'");code = f.read().replace('"'"'\r\n'"'", '"'"'\n'"'");f.close();exec(compile(code, __file__,
'"'"'exec'"'"'))' install --record 'C:\Users\USER\AppData\Local\Temp\pip-record-4y6qetop\install-record.txt' --single-version-exte
rnally-managed --compile --install-headers 'C:\Users\USER\anaconda3\Include\python-crfsuite-extension'
         cwd: C:\Users\USER\AppData\Local\Temp\pip-install-o7lavm5f\python-crfsuite-extension_72ec53254bbc4dc5bd29134ce797beca\
    Complete output (12 lines):
    running install
    running build
    running build_py
    creating build
    creating build\lib.win-amd64-3.9
    creating build\lib.win-amd64-3.9\pycrfsuite
    copying pycrfsuite\_dumpparser.py -> build\lib.win-amd64-3.9\pycrfsuite
    copying pycrfsuite\_logparser.py -> build\lib.win-amd64-3.9\pycrfsuite
    copying pycrfsuite\__init__.py -> build\lib.win-amd64-3.9\pycrfsuite
    running build_ext
    building 'pycrfsuite._pycrfsuite' extension
    error: Microsoft Visual C++ 14.0 or greater is required. Get it with "Microsoft C++ Build Tools": https://visualstudio.microso
ft.com/visual-cpp-build-tools/
    ----------------------------------------
ERROR: Command errored out with exit status 1: 'C:\Users\USER\anaconda3\python.exe' -u -c 'import io, os, sys, setuptools, tokeniz
e; sys.argv[0] = '"'"'C:\\Users\\USER\\AppData\\Local\\Temp\\pip-install-o7lavm5f\\python-crfsuite-extension_72ec53254bbc4dc5bd291
34ce797beca\\setup.py'"'"'; __file__='"'"'C:\\Users\\USER\\AppData\\Local\\Temp\\pip-install-o7lavm5f\\python-crfsuite-extension_7
2ec53254bbc4dc5bd29134ce797beca\\setup.py'"'"';f = getattr(tokenize, '"'"'open'"'", open)(__file__) if os.path.exists(__file__) e
lse io.StringIO('"'"'from setuptools import setup; setup()'"'");code = f.read().replace('"'"'\r\n'"'", '"'"'\n'"'");f.close();e
xec(compile(code, __file__, '"'"'exec'"'"'))' install --record 'C:\Users\USER\AppData\Local\Temp\pip-record-4y6qetop\install-recor
d.txt' --single-version-externally-managed --compile --install-headers 'C:\Users\USER\anaconda3\Include\python-crfsuite-extension'
Check the logs for full command output.
```

In [76]:
```
pip install python-crfsuite
```

```
Requirement already satisfied: python-crfsuite in c:\users\user\anaconda3\lib\site-packages (0.9.8)
Note: you may need to restart the kernel to use updated packages.
```

In [77]:
```
import pycrfsuite
```

In [78]:
```
tag = CRFTagger()
tag.set_model_file('all_indo_man_tag_corpus_model.crf.tagger')
```

In [79]:
```
#SAVE HASIL PREPROCESSING
import xlsxwriter
workbook = xlsxwriter.Workbook('hasil.xlsx', {'nan_inf_to_errors': True})
worksheet=workbook.add_worksheet()
row=0
col=0
x=data
worksheet.write(row, col, "hasil")
row+=1
for e in processeds:
    worksheet.write(row, col, e)
    row+=1
workbook.close()
```

In [80]:
```
datam=pd.read_excel("hasil.xlsx")
datam
```

Out[80]:

	hasil
0	rilis informasi kait prediksi
1	cepat tangan tanggal covid
2	NaN
3	bas ekosistem
4	tangan sebar dasar provinsi status
...	...
1276	sebar jadi bencana alam periode
1277	jiwa ungsi milik
1278	malam sobatkriskes kembang tanggal pkl
1279	jiwa ungsi milik
1280	santun tanggal

1281 rows × 1 columns

In [81]:
```
sans = datam['hasil']
sans
```

Out[81]:
```
0                rilis informasi kait prediksi
1                   cepat tangan tanggal covid
2                                          NaN
3                                bas ekosistem
4           tangan sebar dasar provinsi status
                        ...
1276           sebar jadi bencana alam periode
1277                           jiwa ungsi milik
1278    malam sobatkriskes kembang tanggal pkl
1279                           jiwa ungsi milik
1280                             santun tanggal
Name: hasil, Length: 1281, dtype: object
```

In [83]:
```
from nltk import word_tokenize
from nltk.tag import CRFTagger
```

In [84]:
```
token_word =sans.apply(word_tokenize)
token_word
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_7964/1557967397.py in <module>
----> 1 token_word =sans.apply(word_tokenize)
      2 token_word

~\anaconda3\lib\site-packages\pandas\core\series.py in apply(self, func, convert_dtype, args, **kwargs)
   4355         dtype: float64
   4356         """
-> 4357         return SeriesApply(self, func, convert_dtype, args, kwargs).apply()
   4358
   4359     def _reduce(

~\anaconda3\lib\site-packages\pandas\core\apply.py in apply(self)
   1041             return self.apply_str()
   1042
-> 1043         return self.apply_standard()
   1044
```

```
   1045     def agg(self):

~\anaconda3\lib\site-packages\pandas\core\apply.py in apply_standard(self)
   1096                 # List[Union[Callable[..., Any], str]]]]"; expected
   1097                 # "Callable[[Any], Any]"
-> 1098                 mapped = lib.map_infer(
   1099                     values,
   1100                     f,  # type: ignore[arg-type]

~\anaconda3\lib\site-packages\pandas\_libs\lib.pyx in pandas._libs.lib.map_infer()

~\anaconda3\lib\site-packages\nltk\tokenize\__init__.py in word_tokenize(text, language, preserve_line)
    127     :type preserve_line: bool
    128     """
--> 129     sentences = [text] if preserve_line else sent_tokenize(text, language)
    130     return [
    131         token for sent in sentences for token in _treebank_word_tokenizer.tokenize(sent)

~\anaconda3\lib\site-packages\nltk\tokenize\__init__.py in sent_tokenize(text, language)
    105     """
    106     tokenizer = load(f"tokenizers/punkt/{language}.pickle")
--> 107     return tokenizer.tokenize(text)
    108
    109

~\anaconda3\lib\site-packages\nltk\tokenize\punkt.py in tokenize(self, text, realign_boundaries)
   1275         Given a text, returns a list of the sentences in that text.
   1276         """
-> 1277         return list(self.sentences_from_text(text, realign_boundaries))
   1278
   1279     def debug_decisions(self, text):

~\anaconda3\lib\site-packages\nltk\tokenize\punkt.py in sentences_from_text(self, text, realign_boundaries)
   1332         follows the period.
   1333         """
-> 1334         return [text[s:e] for s, e in self.span_tokenize(text, realign_boundaries)]
   1335
   1336     def _slices_from_text(self, text):

~\anaconda3\lib\site-packages\nltk\tokenize\punkt.py in <listcomp>(.0)
   1332         follows the period.
   1333         """
-> 1334         return [text[s:e] for s, e in self.span_tokenize(text, realign_boundaries)]
   1335
```

```
   1336     def _slices_from_text(self, text):

~\anaconda3\lib\site-packages\nltk\tokenize\punkt.py in span_tokenize(self, text, realign_boundaries)
   1322         if realign_boundaries:
   1323             slices = self._realign_boundaries(text, slices)
-> 1324         for sentence in slices:
   1325             yield (sentence.start, sentence.stop)
   1326

~\anaconda3\lib\site-packages\nltk\tokenize\punkt.py in _realign_boundaries(self, text, slices)
   1363         """
   1364         realign = 0
-> 1365         for sentence1, sentence2 in _pair_iter(slices):
   1366             sentence1 = slice(sentence1.start + realign, sentence1.stop)
   1367             if not sentence2:

~\anaconda3\lib\site-packages\nltk\tokenize\punkt.py in _pair_iter(iterator)
    317     iterator = iter(iterator)
    318     try:
--> 319         prev = next(iterator)
    320     except StopIteration:
    321         return

~\anaconda3\lib\site-packages\nltk\tokenize\punkt.py in _slices_from_text(self, text)
   1336     def _slices_from_text(self, text):
   1337         last_break = 0
-> 1338         for match in self._lang_vars.period_context_re().finditer(text):
   1339             context = match.group() + match.group("after_tok")
   1340             if self.text_contains_sentbreak(context):

TypeError: expected string or bytes-like object
```

In [85]:

```
hasil_POS= tag.tag_sents(token_word)
hasil_POS
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_7964/581351199.py in <module>
----> 1 hasil_POS= tag.tag_sents(token_word)
      2 hasil_POS

NameError: name 'token_word' is not defined
```

In [ ]:
```
#mengubah nested list menjadi flat list
flatList = [element for innerList in hasil_POS for element in innerList]
flatList
```

In [ ]:
```
nltk.download('averaged_perceptron_tagger')
nltk.download('maxent_ne_chunker')
nltk.download('words')
```

In [ ]:
```
entities =[]
labels =[]

for i,j in flatList:
    if i and j != None:
        entities.append(i)
        labels.append(j)

entities_labels = list(set(zip(entities, labels)))
entities_df = pd.DataFrame(entities_labels)
entities_df.columns = ["Entities","Labels"]
entities_df
```

In [ ]: