Latihan LDA & LSA

Nama : Alya Fauziyah

NIM : 21110030

Kelas : S1-SD02A

```
!pip install regex matplotlib sastrawi xlsxwriter nltk
```

```
Requirement already satisfied: regex in /usr/local/lib/python3.10/dist-packages (2023.12.25)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (3.7.1)
Requirement already satisfied: sastrawi in /usr/local/lib/python3.10/dist-packages (1.0.1)
Requirement already satisfied: xlsxwriter in /usr/local/lib/python3.10/dist-packages (3.1.9)
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (4.47.2)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.4.5)
Requirement already satisfied: numpy>=1.20 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.23.5)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (23.2)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (2.8.2)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.3.2)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.7->matplotlib) (1.16.0)
```

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

```
data = pd.read_csv('dataset.csv')
data
```

	conversation_id_str	created_at	favorite_count	full_text	id_str	
0	1753950154116432276	Sun Feb 04 01:14:38 +0000 2024	13653	PEMBUNUHAN DI SIANG HARI Sebuah kasus pembunuh...	1753950154116432276	https://pt
1	1753816569476743417	Sat Feb 03 16:23:49 +0000 2024	203	Loe tau ga knp gerbong tetangga Jor²an keluari...	1753816569476743417	https://
2	1753750790475518100	Sat Feb 03 12:03:35 +0000 2024	3736	kasus yg sama tapi di India https://t.co/KscaV...	1753751082814312785	https://
3	1752994269181469088	Thu Feb 01 09:56:17 +0000 2024	2585	Yang menjegal Anies akan terjungkal 1. Diroast...	1752994269181469088	https://
4	1754030008522432851	Sun Feb 04 06:31:56 +0000 2024	16	Ya Allah baru tau kalo ternyata dia yang jadi ...	1754030008522432851	https://
...	...	...	...	...	...	
276	1753185420282892787	Thu Feb 01 22:35:51 +0000 2024	97	Wait.. ini kader PDIP ? Ada nama Cak Imin yg j...	1753185420282892787	
277	1751059387756810249	Sat Jan 27 01:47:45 +0000 2024	0	Why the hell there have been lot of bad news a...	1751059387756810249	
278	1751137798474830263	Sat Jan 27 07:15:43 +0000 2024	17	@convefrt HATI HATI YA KALIAN SEMUA BANYAK BGT...	1751141921366712787	
279	1752343920489717869	Wed Jan 31 02:53:12 +0000 2024	0	@Fahrihamzah Saya pengen kasus ini d ungkap ht...	1752525408162193617	https://
280	1753369310184219011	Fri Feb 02 10:46:34 +0000 2024	5	Senator Arya Wedakarna Dipecat dari Anggota DP...	1753369310184219011	

281 rows × 15 columns

```
# Preprocessing
import re as reg
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', 'klo':'kalo', 'lg':
        'dlm':'dalam','dgn':'dengan', 'poto':'foto', 'g':'tidak', 'n':'dan', 'ad':'ada', 'brp': 'berapa', "abis": "habis", "ad": "ada", "adlh": "adalah", "afa
        "ahaha": "haha", "aj": "saja", "ajep-ajep": "dunia gemerlap", "ak": "saya", "akika": "aku", "akkoh": "aku", "akuwh": "aku", "alay": "norak", "alow": "
        "ancur": "hancur", "anjrit": "anjing", "anter": "antar", "ap2": "apa-apa", "apasih": "apa sih", "apes": "sial", "aps": "apa", "aq": "saya", "aquwh": "
        "aseekk": "asyik", "asekk": "asyik", "asem": "asam", "aspal": "asli tetapi palsu", "astul": "asal tulis", "ato": "atau", "au ah": "tidak mau tahu", "au
        "ayank": "sayang", "b4": "sebelum", "bakalan": "akan", "bandes": "bantuan desa", "bangedh": "banget", "banpol": "bantuan polisi", "banpur": "bantuan t
        "bcanda": "bercanda", "bdg": "bandung", "begajulan": "nakal", "beliin": "belikan", "bencong": "banci", "bentar": "sebentar", "ber3": "bertiga", "beres
        "bosan", "beud": "banget", "bg": "abang", "bgmn": "bagaimana", "bgt": "banget", "bijimane": "bagaimana", "bintal": "bimbingan mental", "bkl": "akan",
        "blegug": "bodoh", "blh": "boleh", "bln": "bulan", "blum": "belum", "bnci": "benci", "bnran": "yang benar", "bodor": "lucu", "bokap": "ayah", "boker":
        "bohong", "boljug": "boleh juga", "bonek": "bocah nekat", "boyeh": "boleh", "br": "baru", "brg": "bareng", "bro": "saudara laki-laki", "bru": "baru",
        "bt": "buat", "btw": "ngomong-ngomong", "buaya": "tidak setia", "bubbu": "tidur", "bubu": "tidur", "bumil": "ibu hamil", "bw": "bawa", "bwt": "buat",
        "cabal": "sabar", "cadas": "keren", "calo": "makelar", "can": "belum", "capcus": "pergi", "caper": "cari perhatian", "ce": "cewek", "cekal": "cegah tan
        "cengengesan": "tertawa", "cepet": "cepat", "cew": "cewek", "chuyunk": "sayang", "cimeng": "ganja", "cipika cipiki": "cium pipi kanan cium pipi kiri",
        "ckp": "cakep", "cmiiw": "correct me if i'm wrong", "cmpur": "campur", "cong": "banci", "conlok": "cinta lokasi", "cowwyy": "maaf", "cp": "siapa", "cpe
        "cucok": "cocok", "cuex": "cuek", "cumi": "Cuma miscall", "cups": "culun", "curanmor": "pencurian kendaraan bermotor", "curcol": "curahan hati colongan
        "d": "di", "dah": "deh", "dapet": "dapat", "de": "adik", "dek": "adik", "demen": "suka", "deyh": "deh", "dgn": "dengan", "diancurin": "dihancurkan",
        "dimintak": "diminta", "disono": "di sana", "dket": "dekat", "dkk": "dan kawan-kawan", "dll": "dan lain-lain", "dlu": "dulu", "dngn": "dengan", "dodol
        "dongs": "dong", "dpt": "dapat", "dri": "dari", "drmn": "darimana", "drtd": "dari tadi", "dst": "dan seterusnya", "dtg": "datang", "duh": "aduh", "dur
        "egp": "emang gue pikirin", "eke": "aku", "elu": "kamu", "emangnya": "memangnya", "emng": "memang", "endak": "tidak", "enggak": "tidak", "envy": "iri",
        "fifo": "first in first out", "folbek": "follow back", "fyi": "sebagai informasi", "gaada": "tidak ada uang", "gag": "tidak", "gaje": "tidak jelas",
        "gan": "juragan", "gaptek": "gagap teknologi", "gatek": "gagap teknologi", "gawe": "kerja", "gbs": "tidak bisa", "gebetan": "orang yang disuka", "geje
        "gepeng": "gelandangan dan pengemis", "ghiy": "lagi", "gile": "gila", "gimana": "bagaimana", "gino": "gigi nongol", "githu": "gitu", "gj": "tidak jelas
        "gn": "begini", "goblok": "bodoh", "golput": "golongan putih", "gowes": "mengayuh sepeda", "gpny": "tidak punya", "gr": "gede rasa", "gretongan": "grat
        "gua": "saya", "guoblok": "goblok", "gw": "saya", "ha": "tertawa", "haha": "tertawa", "hallow": "halo", "hankam": "pertahanan dan keamanan", "hehe": "h
        "hlm": "halaman", "hny": "hanya", "hoax": "isu bohong", "hr": "hari", "hrus": "harus", "hubdar": "perhubungan darat", "huff": "mengeluh", "hum": "rumah
        "ilfil": "tidak suka", "imho": "in my humble opinion", "imoetz": "imut", "item": "hitam", "itungan": "hitungan", "iye": "iya", "ja": "saja", "jadiin":
        "jayus": "tidak lucu", "jdi": "jadi", "jem": "jam", "jga": "juga", "jgnkan": "jangankan", "jir": "anjing", "jln": "jalan", "jomblo": "tidak punya paca
        "jutek": "galak", "k": "ke", "kab": "kabupaten", "kabor": "kabur", "kacrut": "kacau", "kadiv": "kepala divisi", "kagak": "tidak", "kalo": "kalau", "ka
        "kamtibmas": "keamanan dan ketertiban masyarakat", "kamuwh": "kamu", "kanwil": "kantor wilayah", "karna": "karena", "kasubbag": "kepala subbagian", "ka
        "kayanya": "kayaknya", "kbr": "kabar", "kdu": "harus", "kec": "kecamatan", "kejurnas": "kejuaraan nasional", "kekeuh": "keras kepala", "kel": "kelurah
        "kepengen": "mau", "kepingin": "mau", "kepsek": "kepala sekolah", "kesbang": "kesatuan bangsa", "kesra": "kesejahteraan rakyat", "ketrima": "diterima"
        "kibul": "bohong", "kimpoi": "kawin", "kl": "kalau", "klianz": "kalian", "kloter": "kelompok terbang", "klw": "kalau", "km": "kamu", "kmps": "kampus",
        "knp": "kenapa", "kodya": "kota madya", "komdis": "komisi disiplin", "komsov": "komunis sovyet", "kongkow": "kumpul bareng teman-teman", "kopdar": "ko
        "kpn": "kapan", "krenz": "keren", "krm": "kirim", "kt": "kita", "ktmu": "ketemu", "ktr": "kantor", "kuper": "kurang pergaulan", "kw": "imitasi", "kyk"
        "lam": "salam", "lamp": "lampiran", "lanud": "landasan udara", "latgab": "latihan gabungan", "lebay": "berlebihan", "leh": "boleh", "lelet": "lambat",
        "lgsg": "langsung", "liat": "lihat", "litbang": "penelitian dan pengembangan", "lmyn": "lumayan", "lo": "kamu", "loe": "kamu", "lola": "lambat berfiki
        "lp": "lupa", "luber": "langsung, umum, bebas, dan rahasia", "luchuw": "lucu", "lum": "belum", "luthu": "lucu", "lwn": "lawan", "maacih": "terima kasi
        "kptsan":"keputusan", "krik": "garing", "krn": "karena", "ktauan": "ketahuan", "ktny": "katanya", "kudu": "harus", "kuq": "kok", "ky": "seperti", "kyk
        "lambreta": "lambat", "lansia": "lanjut usia", "lapas": "lembaga pemasyarakatan", "lbur": "libur", "lekong": "laki-laki", "lg": "lagi", "lgkp": "lengka
        "linmas": "perlindungan masyarakat", "lmyan": "lumayan", "lngkp": "lengkap", "loch": "loh", "lol": "tertawa", "lom": "belum", "loupz": "cinta", "lowh"
        "luchu": "lucu", "luff": "cinta", "luph": "cinta", "lw": "kamu", "lwt": "lewat", "maaciw": "terima kasih", "mabes": "markas besar", "macem-macem": "ma
        "maen": "main", "mahatma": "maju sehat bersama", "mak": "ibu", "makasih": "terima kasih", "malah": "bahkan", "malu2in": "memalukan", "mamz": "makan",
        "markus": "makelar kasus", "mba": "mbak", "mending": "lebih baik", "mgkn": "mungkin", "mhn": "mohon", "miker": "minuman keras", "milis": "mailing list
        "mnt": "minta", "moge": "motor gede", "mokat": "mati", "mosok": "masa", "msh": "masih", "mskpn": "meskipun", "msng2": "masing-masing", "muahal": "maha
        "mumet": "pusing", "muna": "munafik", "munaslub": "musyawarah nasional luar biasa", "musda": "musyawarah daerah", "muup": "maaf", "muuv": "maaf", "nal
        "naon": "apa", "napol": "narapidana politik", "naq": "anak", "narsis": "bangga pada diri sendiri", "nax": "anak", "ndak": "tidak", "ndut": "gendut",
        "nelfon": "menelepon", "ngabis2in": "menghabiskan", "ngakak": "tertawa", "ngambek": "marah", "ngampus": "pergi ke kampus", "ngantri": "mengantri", "nga
        "ngaruh": "berpengaruh", "ngawur": "berbicara sembarangan", "ngeceng": "kumpul bareng-bareng", "ngeh": "sadar", "ngekos": "tinggal di kos", "ngelamar"
        "ngemeng": "bicara terus-terusan", "ngerti": "mengerti", "nggak": "tidak", "ngikut": "ikut", "nginep": "menginap", "ngisi": "mengisi", "ngmg": "bicara
        "ngomongin": "membicarakan", "ngumpul": "berkumpul", "ni": "ini", "nyasar": "tersesat", "nyariin": "mencari", "nyiapin": "mempersiapkan", "nyiram": "m
        "ok": "ok", "priksa": "periksa", "pro": "profesional", "psn": "pesan", "psti": "pasti", "puanas": "panas", "qmo": "kamu", "qt": "kita", "rame": "ramai
        "red": "redaksi", "reg": "register", "rejeki": "rezeki", "renstra": "rencana strategis", "reskrim": "reserse kriminal", "sni": "sini", "somse": "sombo
        "sosbud": "sosial-budaya", "sospol": "sosial-politik", "sowry": "maaf", "spd": "sepeda", "sprti": "seperti", "spy": "supaya", "stelah": "setelah", "su
        "sumbangin": "sumbangkan", "sy": "saya", "syp": "siapa", "tabanas": "tabungan pembangunan nasional", "tar": "nanti", "taun": "tahun", "tawh": "tahu",
        "tekor": "rugi", "telkom": "telekomunikasi", "telp": "telepon", "temen2": "teman-teman", "tengok": "menjenguk", "terbitin": "terbitkan", "tgl": "tangg
        "thd": "terhadap", "thx": "terima kasih", "tipi": "TV", "tkg": "tukang", "tll": "terlalu", "tlpn": "telepon", "tman": "teman", "tmbh": "tambah", "tmn2
        "tnda": "tanda", "tnh": "tanah", "togel": "toto gelap", "tp": "tapi", "tq": "terima kasih", "trgntg": "tergantung", "trims": "terima kasih", "cb": "co
        "reklamuk": "reklamasi", "sma": "sama", "tren": "trend", "ngehe": "kesal", "mz": "mas", "analisise": "analisis", "sadaar": "sadar", "sept": "september
        "zonk": "bodoh", "rights": "benar", "simiskin": "miskin", "ngumpet": "sembunyi", "hardcore": "keras", "akhirx": "akhirnya", "solve": "solusi", "watuk"
        "masy": "masyarakat", "still": "masih", "tauk": "tahu", "mbual": "bual", "tioghoa": "tionghoa", "ngentotin": "senggama", "kentot": "senggama", "faktak
        "rubahnn": "rubah", "trlalu": "terlalu", "nyela": "cela", "heters": "pembenci", "nyembah": "sembah", "most": "paling", "ikon": "lambang", "light": "te
        "setting": "atur", "seting": "akting", "next": "lanjut", "waspadalah": "waspada", "gantengsaya": "ganteng", "parte": "partai", "nyerang": "serang", "n
        "jentelmen": "berani", "buangbuang": "buang", "tsangka": "tersangka", "kurng": "kurang", "ista": "nista", "less": "kurang", "koar": "teriak", "paranoi
        "tahi": "kotoran", "tirani": "tiran", "tilep": "tilap", "happy": "bahagia", "tak": "tidak", "penertiban": "tertib", "uasai": "kuasa", "mnolak": "tolak
        "taik": "tahi", "wkwkkw": "tertawa", "ahoknccc": "ahok", "istaa": "nista", "benarjujur": "jujur", "mgkin": "mungkin", 'ga':'tidak', 'cwe':'perempuan',
        'ny':'nya', 'htm':'harga tiket masuk', 'cm':'cuma', 'slalu':'selalu', 'tingi':'tinggi','neng':'senang'}
processed_comments = []

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

    # Converting to Lowercase
    processed_comment = processed_comment.lower()

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

    # remove all single characters
    processed_comment = reg.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', 'coba',
                     'ada','bersih', 'salur', 'baru', 'purwokerto', 'batas', 'hotel', 'coba', 'putus', 'ada','sama', 'suka', 'bilang',
                     'com', 'kamu', 'http', 'https', 'htps', 'htp', 'gak', 'jadi', 'lebih', 'kalau', 'banyak', 'jangan', 'iya', 'kok',
                     'apa', 'paling', 'semua', 'lah', 'ihwm', 'od', 'pakai', 'hayo', 'no', 'ihwjm', 'od'] #menambahkan stopword
    stopwords = factory.get_stop_words() + more_stopword
    temp = [t for t in reg.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 = reg.sub(r'\s+', ' ', processed_comment, flags=reg.I)

    processed_comments.append(processed_comment)

# Output data Preprocessing
processed_comments
```

```
     'buat tanya nak rektor kasus ks iinat angota tim satgas investigasi kasus ks ui periu jaun rekayasa emang siapa meiki trus kaiian mau orang angota
     satgas punya integritas co jdaoc',
     'dugan pelangaran gibran sat bagi bagi susu cfd jakarta bak hilang tel bumi jajar pemprov dki kompak tutup kasus co honzbgobma',
     'selesai ketangkep kasus korupsi masuk politik filsuf co yedm',
     'henti beri tahu orang nyata el gemoy dalang balik kasus culi aktivis sharing caring demokrasi tetap jaga bukti deret rekam jejak pak wowo laen fahri
     nyala kades malaysia nganjuk jumatberkah bobo co yb wxh',
     'yes im talking about bang melki sini marah disgust thought you were an aly bang pernah aksi isu ks bareng nanganin kasus bareng ngobrol dulu nyata
     yah cuma buat branding diri amp alat politik bang',
     'mirip kasus gubernur jatim khofifah kantor geledah kpk esok langsung deklarasi dukung aman dong sandera sandera asalbukan co zwk lwcxvt',
     'beliau tolak vaksin covid buka dpr dikpkan mantan menkes siti fadilah supari tolak pandemi palsu lawan who dikpkan tarik mungkin tuju kpk era jokowi
     co bmcncepa',
     'tachixtus bangwin baca hambat sebab kasus nga lanjut sat prabowo ngejabat mantan danjen kopasus jabat pangkostrad kan mana libat beri informasi tsb
     bawahanya co fjymtucj',
     'salut pak ahok bukti orang orang tersandra kasus hukum sikap tentu pilih dukung pak ganjarpranowo pak mohmahfudmd angkat topi pak co imitasi
     srwtniv',
     'dips gara gara kasus anjem kemarin biar ulang so here is list of anjem jastip percaya thread',
     'aiman witjaksono aju lindung hukum dewan pers kasus polisi netral co mvjpv jdy tempometro',
     'sekarang hilang prabowo kuasa kasus mustahil pecah misan depan istana kelak orang tua korban terus menang cerdas pilih gaes asalbukanprabowo co
     cujamcln',
     'truth be told the situation was chaotic even my mom dare not to step outa our house what do you expect from this korban lapor diangap kasus sebut
     isap jempol belaka are you for real',
     'joki nulis tangan tangan ku kuat buat nulis kasus waktu hari',
     'jhonisinaga ganjarpranowo mohmahfudmd sebut tikus das kasus denysiregar telkomsel ktp angkat harun mas dulu co szyqrlfpn',
     'wkwkw au xodiac aku bakaln bikin friendship drama misteri sih kalian harap dibikinin bucin cewek aku bias bikin tema horor drama misteri angst baca
     webtonku endingnya mecahin satu kasus besar',
     'tkn prabowo konflik penting koalisi hampir tersandra kasus korupsi tim orang orang masalah pait pait asalbukanprabowogibran co xfnuso hf',
     'bupati pemkabsidoarjo gus muhdlor hilang sat ot rupa sibuk kampanye mirip buk kokipah yak punya kasus korupsi rapat paslon asalbukan co vfrzry has',
     'joki tugas manajemen soal kasus ai chatgpt jokitugas zonaba',
     'kejaksan agung tetap abdul had aviciena tertawa mantan general manager antam sangka kasus dugan korupsi jual emas logam mulia butik co mgs na',
     'bagi tanda tanda kasus kanjuruhan selesai terus kejar adil saudara lawan generasi tua acau lurus generasi muda sesat',
     'ned joki tugas character building studi kasus ned jokitugas joki',
     'mahfud gk mungkin nyelesain cuma omong kampanye sendiri kasus negara gk mungkin menyelesaikanya berani taruh deh soal kasus sangkut tingi negara
     konteks gk kategori salah',
     'cuma ikutin alur cerita akting mecahin kasus emang sendiri trust isue le seungi psikopat drakor mouse gara emg drakornya plot twist',
     'pak mahfud memang pernah beri contoh skenario pic restorative justice tahun tahun jadi kasus pemerkosan beneran pic amp beliau justru terima kasus
     cuma diresolve cara damai co lklipvn',
     'joki kasus posisi in rush dong sebut fe langsung kisar berapa zonauang',
     'kasus seleb bawah inget rafi dugan cuci uang bangun beach club kawasan lindung nagita kasus plagiarisme produk zh sebut aksi kamis dagang rachel
     kabur karantina covid marshel tebar benci rohingya cok nyabu alshad bajing inti co askmqlxc',
     'fans seunghan kaya dongkrak stereotype sifat fans lama peduli masa lalu seunghan dulu seunghan pacar ngerokok peduli tetep dukung seunghan biasa
     kasus kaya gin fans mungkin ningalin',
     'revisapratama who is paying you pilih percaya praduga salah kasus melki benar tangan pihak wenang polisi kejaksan adil dapat putus inkracht bukan
     rektor tolol in the mean time go ef yourself co tn',
     'kompastv presiden akhir jabat kasus korupsi besar proyek mangkrak ungkap kalah ungkap korupsi presiden nepotisme rusak demokrasi proyek mangkrak uda
     almarhum pak aku aku klaim sendiri',
     'lanjut nyata kasus protocol breach bukan kasus curi intelejen data strategis laku copy dokumen pekerjan usb mungkin kerja rumah tempat mana usah
     besar larang',
     'mel baik terlalu gegabah narasi korban bagus preseden kasus ks',
     'nomin prompt au ceritanye si kemal kena kasus gara gara manta paksa bocah cuti sebentar balik rumah intinye nyeritain si kemal lama cuti timnas
     ketemu orang suka lama si prana tetap semangat king jovan co zi mlru',
     'jokowi tukar guling kasus kayak si khofifah co yompgxornx',
     'blesing disguise lama anies dki diroasting giring malah populer larang monas bahkan bikin sendiri sirkuit formula background jis karya anies lapor
     kpk kasus formula bahkan dukung luas koalisi bubar bahkan cak imin paslon gameover co mimiwvewx',
     'kalian loh aku sifat keras aku mevalidasi ketidakadilan kasus the same way kalian harus mevalidasi rasa takut trauma kaum chindo kristen sat pilih
     kepala daerah mevalidasi marah kaum muslim btp chery picking',
     'fahrihamzah manusia macam ko bejad moral amp mental rela jilat pantat rezim sandera kasus kan kena kutuk alah maha tahu co rgreyvjge',
     'aneh banget nge cap orang peduli politik enga cuma base doang korban ui fes emang tone deaf aware masalah even lingkup ui sendiri anak ui tu peduli
     politik bawa bawa kasus akseyna jahat banget',
     'investor asal china dpo kasus palsu dokumen pabrik nikel hasil bayar co gydorwvwyl',
```

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

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

```
processed_comments = ''.join(processed_comments)
```

```
type(processed_comments)
```

```
     str
```

```
from nltk.tokenize import word_tokenize
Docs = processed_comments
Sentences = nltk.sent_tokenize(Docs)
print ("Kalimat Awal: ", Sentences)
hasil_tokenizing = nltk.word_tokenize(Docs)
print("Setelah Tokenizing: ", hasil_tokenizing)
```

```
Kalimat Awal:  ['bunuh siang hari buah kasus bunuh butuh kolaborasi multidisiplin ilmu pengungkapanya cermat scientific logis co tyq bv cuotahu gerbong tet
Setelah Tokenizing:  ['bunuh', 'siang', 'hari', 'buah', 'kasus', 'bunuh', 'butuh', 'kolaborasi', 'multidisiplin', 'ilmu', 'pengungkapanya', 'cermat', 'sci
```

```
len(hasil_tokenizing)
```

```
5849
```

```
# Ekstraksi fitur / representasi dokumen menggunakan TF
from sklearn.feature_extraction.text import CountVectorizer
tf_vectorizer = CountVectorizer(max_df=1.0, min_df=1)

tf = tf_vectorizer.fit_transform(hasil_tokenizing)

# hasil representasi
tf_terms = tf_vectorizer.get_feature_names_out()
print(tf_vectorizer.get_feature_names_out())
matrix = tf.toarray()
print(matrix)
```

```
['abah' 'abai' 'abal' ... 'zul' 'zwk' 'zyb']
[[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]]
```

Menggunakan LDA

```
from sklearn.decomposition import LatentDirichletAllocation as LDA

n_topics = 10 #untuk mendapatkan jumlah topik terbaik perlu trial
lda = LDA(n_components=n_topics, learning_method='batch', random_state=0)
lda.fit(tf)
lda
```

```
LatentDirichletAllocation(random_state=0)
```

```
vsm_topics = lda.transform(tf)
#tampilkan hasil
print(vsm_topics)
```

```
[[0.05       0.05       0.05       ... 0.05       0.05       0.05      ]
 [0.55       0.05       0.05       ... 0.05       0.05       0.05      ]
 [0.55       0.05       0.05       ... 0.05       0.05       0.05      ]
 ...
 [0.05       0.05       0.05       ... 0.05       0.05       0.54999999]
 [0.05       0.05       0.05       ... 0.05       0.05       0.05      ]
 [0.05       0.05       0.05       ... 0.05       0.55       0.05      ]]
```

```
len(vsm_topics)
```

```
5849
```

```
#Tampilkan nilai-nilai setiap fitur
print(lda.components_)
```

```
[[0.1        0.1        0.1        ... 0.1        0.1        0.1       ]
 [2.09999998 0.1        0.1        ... 2.09999998 0.1        0.1       ]
 [0.1        0.1        0.1        ... 0.1        0.1        0.1       ]
 ...
 [0.1        0.1        2.09999998 ... 0.1        0.1        0.1       ]
 [0.1        0.1        0.1        ... 0.1        0.1        1.09999998]
 [0.1        0.1        0.1        ... 0.1        0.1        0.1       ]]
```

```
#hasil label topic model untuk setiap dokumen
import numpy as np
topics = np.argmax(vsm_topics, axis=1)
topics
```

```
array([5, 0, 0, ..., 9, 6, 8])
```

```
#mencetak word fitur dengan nilai tertinggi pada setiap topik
n_top_words = 10 # jumlah fitur tertinggi yang kita tentukan
topic_words = {}
for topic, comp in enumerate(lda.components_):
 word_idx = np.argsort(comp)[::-1][:n_top_words] # susun indeks secara terbalik, sehingga yang paling tinggi muncul di awal.
 # store the words most relevant to the topic
 topic_words[topic] = [tf_vectorizer.get_feature_names_out()[i]+' '+str(comp[i]) for i in word_idx]


type(topic_words)
```

```
dict
```

```
topic_words
```

```
{0: ['kasus 281.0999999858961',
  'pak 23.09999998566645',
  'cuma 17.099999985577362',
  'the 12.099999985433392',
  'pernah 11.099999985388475',
  'kalian 11.099999985388475',
  'hilang 10.099999985334316',
  'selesai 10.099999985334316',
  'hari 9.099999985267738',
  'penistan 9.099999985267738'],
 1: ['buat 26.099999987148603',
  'banget 20.09999998708864',
  'korban 19.099999987074927',
  'salah 16.099999987023256',
  'terus 13.099999986947202',
  'bikin 13.099999986947202',
  'pilih 11.099999986872838',
  'masa 11.099999986872838',
  'jelas 10.099999986824182',
  'jokowi 9.099999986764367'],
 2: ['uang 20.099999988433783',
  'libat 20.099999988433783',
  'dukung 18.099999988407795',
  'lapor 12.099999988276604',
  'dulu 10.09999998819687',
  'bagaimana 10.099999988196869',
  'aman 9.099999988143285',
  'lalu 8.099999988075822',
  'kan 8.099999988075822',
  'sekali 8.099999988075822'],
 3: ['korupsi 27.099999988151417',
  'amp 20.099999988089294',
  'mau 19.099999988076632',
  'tahun 14.09999998798561',
  'deh 13.09999998795881',
  'presiden 11.099999987890214',
  'agama 10.099999987845328',
  'gara 10.099999987845326',
  'takut 9.099999987790143',
  'hak 7.0999999876305395'],
 4: ['co 188.09999998635143',
  'orang 47.09999998626417',
  'mungkin 14.099999985983432',
  'cuci 9.0999999857554',
  'menang 8.099999985674355',
  'seunghan 7.099999985569198',
  'beri 7.099999985569197',
  'pelangaran 7.099999985569197',
  'ketua 7.099999985569197',
  'terima 7.099999985569197'],
 5: ['prabowo 25.099999987630767',
  'bahkan 21.099999987592717',
  'tahu 20.099999987580784',
  'kpk 17.099999987536425',
  'ungkap 15.099999987496796',
  'ahok 12.099999987412005',
  'melki 12.099999987412003',
  'kena 10.099999987326385',
  'fahri 10.099999987326385'.
```

```
for topic, words in topic_words.items():
 print('Topic: %d' % topic)
 print(' %s' % ', '.join(words))
```

```
Topic: 0
 kasus 281.0999999858961, pak 23.09999998566645, cuma 17.099999985577362, the 12.099999985433392, pernah 11.099999985388475, kalian 11.099999985388475, hi
Topic: 1
 buat 26.099999987148603, banget 20.09999998708864, korban 19.099999987074927, salah 16.099999987023256, terus 13.099999986947202, bikin 13.09999998694720
Topic: 2
 uang 20.099999988433783, libat 20.099999988433783, dukung 18.099999988407795, lapor 12.099999988276604, dulu 10.09999998819687, bagaimana 10.099999988196
Topic: 3
 korupsi 27.099999988151417, amp 20.099999988089294, mau 19.099999988076632, tahun 14.09999998798561, deh 13.09999998795881, presiden 11.099999987890214,
Topic: 4
 co 188.09999998635143, orang 47.09999998626417, mungkin 14.099999985983432, cuci 9.0999999857554, menang 8.099999985674355, seunghan 7.099999985569198, b
Topic: 5
 prabowo 25.099999987630767, bahkan 21.099999987592717, tahu 20.099999987580784, kpk 17.099999987536425, ungkap 15.099999987496796, ahok 12.09999998741200
Topic: 6
 benar 21.099999987871307, tim 16.099999987798157, lama 15.09999998777755, jadi 14.099999987753947, ks 13.099999987726639, jabat 11.099999987656716, langs
Topic: 7
 laku 17.099999987470568, lihat 15.099999987430738, anak 14.099999987406466, mana 13.099999987378379, masuk 11.099999987306473, ui 10.099999987259428, bia
Topic: 8
 si 24.099999988668806, anies 18.099999988605088, ham 16.099999988573053, satu 16.099999988573053, sebut 16.099999988573046, negara 16.099999988573046, ny
Topic: 9
 aku 18.099999988506447, hukum 16.099999988474128, sat 11.099999988340524, soal 11.099999988340524, dalang 10.099999988297311, kampanye 10.09999998829731,
```

Interpretasi hasil :

Misal topic 0, didalamnya terdapat kata kasus, pak, cuma, the, pernah, dsb dengan probabilitas masing - masing. Di mana kita dapat melihat bahwa kata kasus memiliki nilai 281.09. Artinya bahwa kata "kasus" dalam topik tersebut memiliki distrusi probalitas paling tinggi.

Menggunakan LSA

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

```
df = pd.read_csv('dataset.csv')
```

```
df.head()
```

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen
0	1753950154116432276	Sun Feb 04 01:14:38 +0000 2024	13653	PEMBUNUHAN DI SIANG HARI Sebuah kasus pembunuh...	1753950154116432276	https://pbs.twimg.com/media/GFdIXW7aUAAjGuW.jpg	
1	1753816569476743417	Sat Feb 03 16:23:49 +0000 2024	203	Loe tau ga knp gerbong tetangga Jor²an keluari...	1753816569476743417	https://pbs.twimg.com/ext_tw_video_thumb/17538...	
2	1753750790475518100	Sat Feb 03 12:03:35 +0000 2024	3736	kasus yg sama tapi di India https://t.co/KscaV...	1753751082814312785	https://pbs.twimg.com/ext_tw_video_thumb/17537...	dxrkc
3	1752994269181469088	Thu Feb 01 09:56:17 +0000 2024	2585	Yang menjegal Anies akan terjungkal 1. Diroast...	1752994269181469088	https://pbs.twimg.com/ext_tw_video_thumb/17529...	
4	1754030008522432851	Sun Feb 04 06:31:56 +0000 2024	16	Ya Allah baru tau kalo ternyata dia yang jadi ...	1754030008522432851	https://pbs.twimg.com/ext_tw_video_thumb/17540...	

```
df = df['full_text']
```

```
df = pd.DataFrame(df)
```

```
df.head()
```

	full_text
0	PEMBUNUHAN DI SIANG HARI Sebuah kasus pembunuh...
1	Loe tau ga knp gerbong tetangga Jor²an keluari...
2	kasus yg sama tapi di India https://t.co/KscaV...
3	Yang menjegal Anies akan terjungkal 1. Diroast...
4	Ya Allah baru tau kalo ternyata dia yang jadi ...

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

```
  (0, 78)       0.11101321081239572
  (0, 134)      0.11101321081239572
  (0, 84)       0.17782595561800688
  (0, 82)       0.2952016139310396
  (0, 395)      0.19178005258049574
  (0, 170)      0.08269502508164485
  (0, 129)      0.32080852145064936
  (0, 91)       0.16105803105574315
  (0, 276)      0.8271095563148579
  (1, 251)      0.19157623756026299
  (1, 9)        0.13941084785366942
  (1, 261)      0.19437761260102351
  (1, 17)       0.19157623756026299
  (1, 328)      0.18398029858206205
  (1, 364)      0.24331282909268118
  (1, 25)       0.37783717497083386
  (1, 196)      0.19157623756026299
  (1, 188)      0.24331282909268118
  (1, 150)      0.26074640351049094
  (1, 236)      0.20382420605627072
  (1, 235)      0.5214928070209819
  (1, 165)      0.23029152061356048
  (1, 115)      0.19157623756026299
  (1, 352)      0.20382420605627072
  (1, 78)       0.07309744210927126
  :     :
  (278, 174)    0.16330943004900883
  (278, 73)     0.2291944083222713
```

```
  (278, 68)     0.18045987373524858
  (278, 162)    0.21692867164252955
  (278, 92)     0.18309869675724108
  (278, 394)    0.15501728040312862
  (278, 396)    0.11153215562152466
  (278, 328)    0.1733046951684099
  (278, 150)    0.24561638569408586
  (278, 352)    0.19199714410056512
  (278, 170)    0.0512915600938876 34
  (279, 386)    0.5829418446092439
  (279, 111)    0.4900117367665208
  (279, 312)    0.5215454385933068
  (279, 141)    0.2731280936992009
  (279, 78)     0.16956462374575065
  (279, 134)    0.16956462374575065
  (279, 170)    0.1263106499758028
  (280, 28)     0.5520461369931775
  (280, 103)    0.46555482899630485
  (280, 86)     0.3354956187082457
  (280, 249)    0.5520461369931775
  (280, 78)     0.15476018076269282
  (280, 134)    0.15476018076269282
  (280, 170)    0.11528264912036723
(281, 400)
```

```
# topic modeling using LSA
from sklearn.decomposition import TruncatedSVD
lsa_model = TruncatedSVD(n_components=5, n_iter=10, random_state=42)

lsa_top=lsa_model.fit_transform(vektor_dt)
print(lsa_top)
```

```
[[ 0.24489392 -0.03299483  0.04468685  0.11017428 -0.0706721 ]
 [ 0.2574406   0.06202824 -0.00622217 -0.07019275 -0.05792675]
 [ 0.43916557 -0.03676021 -0.01696484 -0.20779744 -0.04608295]
 ...
 [ 0.1863004   0.09574083 -0.17533301 -0.10586382  0.0529373 ]
 [ 0.29529912 -0.13863177  0.37550745  0.20018226  0.1873339 ]
 [ 0.21061474 -0.0542782   0.16932642 -0.08459686 -0.07850294]]
```

```
len(lsa_top)
```

```
281
```

```
lsa_model.get_params()
```

```
{'algorithm': 'randomized',
 'n_components': 5,
 'n_iter': 10,
 'n_oversamples': 10,
 'power_iteration_normalizer': 'auto',
 'random_state': 42,
 'tol': 0.0}
```

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

```
Topik- Topik:
Topic  0  :  24.48939205115884
Topic  1  :  -3.299483099895188
Topic  2  :  4.4686850962989855
Topic  3  :  11.01742761619714
Topic  4  :  -7.06721042520225
```

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

```
(5, 400)
[[ 1.85015241e-02  5.88822759e-02  2.43492109e-02 ...  1.56573564e-02
   6.32247140e-03  2.22711349e-02]
 [-1.30119150e-02 -2.89562226e-02 -6.58324558e-04 ... -9.65710952e-03
  -4.29626584e-03 -2.72892895e-02]
 [-3.32844689e-03  8.10170347e-03  2.46733987e-02 ... -2.77526211e-04
  -1.32632987e-02 -8.11958373e-02]
 [-1.18225738e-02 -1.02181715e-01  3.73236062e-02 ...  2.54792279e-02
   2.36031839e-03 -4.65885077e-02]
 [-2.42424169e-02 -6.54073501e-02 -1.16090003e-02 ...  2.10883943e-02
   6.58769776e-03  3.00338064e-01]]
```

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

```
Topic 0:
co, https, kasus, yg, ini, di, dan, yang, ada, bisa,

Topic 1:
tim, agama, dalang, penistaan, ternyata, ahok, prabowo, allah, tau, baru,

Topic 2:
co, https, fahrihamzah, hamzah, fahri, benur, kasus, saya, dugaan, takutnya,

Topic 3:
yang, ini, saya, korban, dan, dengan, tidak, buat, hari, fahrihamzah,

Topic 4:
joki, zonauang, tugas, fahrihamzah, jokitugas, hamzah, ini, fahri, benur, sebut,
```

```
for topic, words in topic_words.items():
 print('Topic: %d' % topic)
 print(' %s' % ', '.join(words))
```

```
Topic: 0
 kasus 281.0999999858961, pak 23.099999998566645, cuma 17.099999985577362, the 12.099999985433392, pernah 11.099999985388475, kalian 11.099999985388475, hi
Topic: 1
 buat 26.099999987148603, banget 20.099999998708864, korban 19.099999987074927, salah 16.099999987023256, terus 13.099999986947202, bikin 13.09999998694720
Topic: 2
 uang 20.099999988433783, libat 20.099999988433783, dukung 18.099999988407795, lapor 12.099999988276604, dulu 10.099999998819687, bagaimana 10.099999988196
Topic: 3
 korupsi 27.099999988151417, amp 20.099999988089294, mau 19.099999988076632, tahun 14.09999998798561, deh 13.09999998795881, presiden 11.099999987890214,
Topic: 4
 co 188.09999998635143, orang 47.099999998626417, mungkin 14.099999985983432, cuci 9.0999999857554, menang 8.099999985674355, seunghan 7.099999985569198, b
Topic: 5
 prabowo 25.099999987630767, bahkan 21.099999987592717, tahu 20.099999987580784, kpk 17.099999987536425, ungkap 15.099999987496796, ahok 12.09999998741200
Topic: 6
 benar 21.099999987871307, tim 16.099999987798157, lama 15.09999998777755, jadi 14.099999987753947, ks 13.099999987726639, jabat 11.099999987656716, langs
Topic: 7
 laku 17.099999987470568, lihat 15.099999987430738, anak 14.099999987406466, mana 13.099999987378379, masuk 11.099999987306473, ui 10.099999987259428, bia
Topic: 8
 si 24.099999988668806, anies 18.099999988605088, ham 16.099999988573053, satu 16.099999988573053, sebut 16.099999988573046, negara 16.099999988573046, ny
Topic: 9
 aku 18.099999988506447, hukum 16.099999988474128, sat 11.099999988340524, soal 11.099999988340524, dalang 10.099999988297311, kampanye 10.09999998829731,
```

Interpretasi hasil :

Misal topic 0, didalamnya terdapat kata kasus, pak, cuma, the, pernah, dsb dengan probabilitas masing - masing. Di mana kita dapat melihat bahwa kata kasus memiliki nilai 281.09. Artinya bahwa kata "kasus" dalam topik tersebut memiliki distrusi probalitas paling tinggi.