In [40]:

```
import numpy as np
import pandas as pd
```

In [41]:

```
data = pd.read_csv("Kasus1.csv", delimiter=';',usecols=['text']) # read file csv nya dulu yang java heritage
data
```

Out[41]:

	text
0	Apa yg bisa diharapkan dari politisi gaya pela...
1	Kalau sy jadi @DennyJA_WORLD , Syahganda aka...
2	Di @PDemokrat banyak org2 yg punya akal sehat ...
3	Cypridophobia adalah fobia atau takut pada pel...
4	Yang RAMAI kerja waktu malam ni biasanya PELAC...
...	...
395	Letakan DiKepalak Sepatunya Klo Tidak Mau Koto...
396	Media pelacur lahir dari rahim yg tak jujur. h...
397	Jadi pelacur ko nanti dijawa nak. Jangan...
398	Kok jadi pelacur
399	/rlt/ no filter ini pake Samsung A50. Negara y...

400 rows × 1 columns

In [42]:

```
pip install Sastrawi
```

```
Requirement already satisfied: Sastrawi in c:\users\user\anaconda3\lib\site-packages (1.0.1)Note: you may need to restart t
he kernel to use updated packages.
```

In [43]:

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

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"}

processed_comments = []

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

    # Converting to Lowercase
    processed_comment = processed_comment.lower()

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

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

    #remove duplicate character
    pattern=re.compile(r"(.)\1{1,}",re.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()
    f = open("stopwords-tala.txt", "r")
    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_comment) if t not in stopwords]
    processed_comment = ' '.join(temp)

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

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

    processed_comments.append(processed_comment)
```

In [44]:

```
processed_comments
```

Out[44]:

```
['harap politis gaya lacur pramagtis warna politik prinsip dlm politik milik integritas politis model suka rakyat',
 'syahganda gugat syahganda mg mg lacur intlektual syahganda buka propaganda hasil survei deny bawaslu https politik rmol
co read syahganda buka propaganda hasil survei deny bawaslu',
 'pdemokrat org akal sehat maju bangsa orang begundal serta lacur politik grasak grusuk salah satu heran https twiter com
ferdinandhaean status',
 'cypridophobia fobia takut lacur tular sakit kelamin',
 'ramai kerja malam lacur siapa idea suh kerja malam tu takfahamtakapa',
 'lacur intlektual kpucurangrakyatmelawan kpungejarquickcounthttps twiter com suwandaben status',
 'lembaga lacur lsi denyja germo kejar target atas propinsi dlm opini publik simultan dg operasi ampo suara hina hidup me
rekahtps twiter com status',
 'bangun gada orang sukses kasur lacur selfreminder',
 'abai lacur agama hidup nkri hidup presiden joko widodo',
 'sat bersih jenis kotor bangsa masive awal laknat lacur agama biadab ancam selamat nkri',
 'sukses capai kasur lacur',
 'harap politis gaya lacur pramagtis warna politik prinsip dlm politik milik integritas politis model suka rakyat',
 'syahganda gugat syahganda mg mg lacur intlektual syahganda buka propaganda hasil survei deny bawaslu https politik rmol
co read syahganda buka propaganda hasil survei deny bawaslu',
 'pdemokrat org akal sehat maju bangsa orang begundal serta lacur politik grasak grusuk salah satu heran https twiter com
```

LDA

In [45]:

```
from sklearn.feature_extraction.text import CountVectorizer
tf_vectorizer = CountVectorizer(max_df=1.0, min_df=1)
tf = tf_vectorizer.fit_transform(processed_comments)
#hasil representasi
tf_terms = tf_vectorizer.get_feature_names()
print(tf_vectorizer.get_feature_names())
matrix = tf.toarray()
print(matrix)
```

```
['abai', 'abal', 'abdu', 'abg', 'abng', 'abu', 'abud', 'ada', 'adab', 'adik', 'adil', 'adrenalin', 'aduh', 'ae', 'agama',
'agenda', 'agt', 'ah', 'ahahaha', 'ahoax', 'ahok', 'air', 'ajah', 'ajak', 'ajakin', 'ajar', 'akal', 'akbar', 'akhirat',
'akomentar', 'aku', 'akubacadotcom', 'akun', 'akwkwk', 'ala', 'alah', 'alam', 'alas', 'alhamdulilah', 'ali', 'alias', 'al
im', 'alir', 'alohu', 'alus', 'aman', 'ambil', 'amca', 'amerika', 'amien', 'amin', 'ampas', 'amplop', 'ampo', 'ampum', 'a
mpun', 'an', 'anak', 'analogi', 'anarkis', 'anatomi', 'ancam', 'and', 'anda', 'andai', 'andik', 'andre', 'ane', 'aneh',
'angap', 'anget', 'angin', 'angkat', 'angkatanya', 'angota', 'aniesbaswedan', 'anime', 'anjeng', 'anjing', 'anti', 'anuar
musa', 'anus', 'anying', 'apa', 'apan', 'aparat', 'apresiasi', 'apung', 'arab', 'arah', 'are', 'arek', 'argentina', 'argu
men', 'arow', 'arti', 'arus', 'as', 'asa', 'asal', 'asli', 'asyik', 'atas', 'atlet', 'atu', 'atur', 'audience', 'australi
a', 'av', 'avatar', 'awal', 'ayah', 'ayam', 'ayo', 'ayu', 'azab', 'bab', 'babak', 'babi', 'baca', 'bacot', 'bada', 'bada
n', 'badut', 'bae', 'bagak', 'bagus', 'bahan', 'bahas', 'bahasa', 'baim', 'baja', 'bajet', 'bajing', 'bal', 'balipuspanew
s', 'balok', 'bambu', 'banci', 'bandar', 'banding', 'bang', 'banga', 'banget', 'bangkiang', 'bangkit', 'bangsa', 'bangsa
t', 'bangun', 'bani', 'bantal', 'banten', 'bantu', 'banyak', 'banyakin', 'banyam', 'bapa', 'bapaknyarico', 'barang', 'bar
angkali', 'baris', 'bas', 'bat', 'batang', 'batas', 'bau', 'bawa', 'bawaslu', 'bayangkn', 'bayar', 'be', 'bebas', 'becom
e', 'beda', 'bedebah', 'bego', 'begundal', 'bejad', 'bekas', 'bela', 'belagu', 'belah', 'belanda', 'belanja', 'beli', 'be
liau', 'beludru', 'benar', 'bencana', 'benci', 'benda', 'benibenana', 'berakir', 'berani', 'berat', 'berengsek', 'berfiki
r', 'berita', 'berjengot', 'berkonsekwensi', 'bersih', 'bersosmed', 'besar', 'betina', 'bgtupun', 'bhahahahaha', 'bhre',
'biadab', 'biang', 'biar', 'biarin', 'biasa', 'biaya', 'bicara', 'bielal', 'bijak', 'bikin', 'bilal', 'bilang', 'bin', 'b
inal', 'binatang', 'bing', 'biologi', 'birah', 'birokrasi', 'bisnis', 'blacklist', 'blbi', 'bledag', 'bledug', 'block',
'blogspot', 'blow', 'blowjob', 'bnget', 'bnl', 'bodo', 'bodoh', 'bohong', 'bojes', 'bold', 'bong', 'bowo', 'boyf', 'brart
```

In [46]:

```
# panggil class 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).fit(tf)
lda
```

Out[46]:

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

In [47]:

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

```
(400, 10)
```

Out[47]:

```
array([[0.00588256, 0.00588279, 0.00588328, ..., 0.00588236, 0.0058827 ,
        0.00588253],
       [0.0037039 , 0.00370397, 0.00370391, ..., 0.00370374, 0.96666464,
        0.00370386],
       [0.00454598, 0.00454641, 0.00454595, ..., 0.00454562, 0.00454585,
        0.00454574],
       ...,
       [0.02000242, 0.02000153, 0.02000101, ..., 0.02000006, 0.02000044,
        0.02000217],
       [0.05001691, 0.05001395, 0.0500139 , ..., 0.0500008 , 0.05000612,
        0.54987628],
       [0.00526458, 0.00526548, 0.00526366, ..., 0.00526317, 0.95262346,
        0.00526355]])
```

In [48]:

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

```
[[2.1        0.1        0.1        ... 0.1        0.1        0.1       ]
 [0.1        0.1        1.09999143 ... 0.1        0.1        0.1       ]
 [0.1        0.1        1.09999925 ... 0.1        0.1        0.1       ]
 ...
 [0.1        0.1        0.1        ... 0.1        0.1        0.1       ]
 [0.1        0.1        0.1        ... 0.1        0.1        0.1       ]
 [0.1        0.1        1.09996271 ... 0.1        0.1        0.10000178]]
```

In [49]:

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

Out[49]:

```
array([4, 8, 3, 1, 0, 3, 1, 9, 0, 6, 9, 4, 8, 3, 0, 3, 1, 0, 6, 1, 1, 7,
       6, 8, 1, 9, 5, 6, 9, 9, 4, 5, 6, 0, 0, 8, 5, 9, 6, 1, 4, 9, 3, 4,
       5, 5, 4, 9, 5, 4, 5, 6, 9, 4, 4, 0, 9, 9, 5, 9, 0, 0, 0, 2, 9, 2,
       4, 9, 5, 6, 4, 6, 9, 5, 5, 6, 8, 2, 4, 9, 6, 0, 4, 6, 8, 5, 9, 0,
       1, 9, 9, 5, 6, 9, 1, 6, 5, 6, 4, 5, 5, 4, 2, 6, 8, 4, 2, 9, 6, 4,
       5, 9, 6, 6, 2, 8, 8, 6, 1, 5, 0, 6, 6, 2, 5, 0, 6, 4, 5, 1, 2, 1,
       0, 0, 0, 9, 0, 4, 1, 1, 6, 4, 0, 5, 1, 2, 1, 0, 0, 0, 4, 2, 6, 9,
       0, 3, 1, 8, 6, 4, 2, 2, 6, 4, 0, 9, 0, 9, 1, 5, 9, 4, 4, 0, 2, 9,
       1, 9, 5, 2, 1, 4, 5, 4, 9, 1, 4, 0, 6, 3, 5, 6, 3, 9, 6, 3, 8, 4,
       4, 6, 5, 4, 0, 4, 4, 0, 6, 9, 0, 4, 6, 9, 1, 9, 8, 2, 2, 6, 9, 1,
       9, 9, 4, 2, 5, 5, 3, 5, 6, 5, 6, 1, 6, 5, 0, 5, 3, 4, 6, 4, 9, 2,
       9, 5, 9, 1, 6, 6, 2, 5, 4, 6, 6, 4, 8, 0, 5, 6, 4, 9, 9, 7, 1, 4,
       1, 4, 0, 2, 9, 3, 9, 1, 9, 8, 6, 6, 1, 6, 9, 6, 1, 9, 6, 3, 9, 2,
       1, 6, 9, 7, 9, 6, 9, 7, 8, 9, 1, 9, 6, 8, 6, 8, 4, 4, 2, 8, 2, 4,
       2, 4, 0, 6, 7, 4, 3, 6, 1, 2, 1, 6, 6, 9, 3, 8, 8, 9, 4, 2, 0, 1,
       3, 2, 2, 9, 8, 9, 0, 0, 2, 6, 1, 1, 6, 4, 4, 0, 4, 2, 4, 8, 5, 2,
       6, 1, 1, 8, 3, 4, 5, 8, 1, 1, 1, 1, 3, 0, 9, 1, 9, 1, 1, 5, 2, 5,
       9, 7, 1, 5, 8, 5, 1, 4, 0, 6, 9, 1, 9, 6, 9, 9, 9, 9, 6, 6, 6, 6,
       1, 4, 9, 8], dtype=int64)
```

In [50]:

```
#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]
 # store the words most relevant to the topic
 topic_words[topic] = [tf_vectorizer.get_feature_names()[i]+' '+str(comp[i]) for i in word_idx]
```

In [51]:

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

```
Topic: 0
 lacur 36.858302877433104, ahok 14.100063547015367, com 10.33124212517124, twiter 9.3115701625257, negeri 9.10005257926472
3, kerja 7.100019285676232, agama 6.100051317894472, presiden 6.100040681345306, tu 6.100037092189047, laki 6.1000116111161
855
Topic: 1
 lacur 34.17682018061304, status 20.95321637287863, twiter 20.72725394922268, com 20.633781250824796, htps 16.6273719501033
8, leceh 9.002413895605685, tutup 8.100028846486598, tusuk 8.100019253365828, ras 8.099903939505213, verbal 7.9899674768977
045
Topic: 2
 lacur 23.100425806660525, ras 11.788535040417313, orang 9.83517213284242, politik 6.100118994339808, iya 6.10002453701038,
suka 5.944528899123657, dasar 5.100115092764825, jaya 5.0999999995441625, gitu 5.09999999923475, maf 4.10008184079854
Topic: 3
 lacur 12.784085463794048, com 7.100041782956527, twiter 6.100040680309208, orang 5.100091145426324, status 5.1000283347639
49, ga 4.100110571668511, ras 4.099959526783606, gak 3.957519079901731, politik 3.1000609901631715, abdu 3.1000392953393807
Topic: 4
 lacur 43.823716325696914, ras 16.100238531340743, politik 16.055280738731454, gak 7.98385730774608, anjing 6.1000224106218
83, politis 5.100032068752101, khianat 4.100031667465745, kayak 3.9825075811719395, rakyat 3.755852239722175, yah 3.1000448
193071923
Topic: 5
 lacur 41.35840756780545, politik 10.100082105401173, com 10.10001285224174, twiter 9.100020097764915, htps 9.1000192583666
63, leceh 8.100048691293653, verbal 7.100032011533481, status 7.0999990467894145, nama 6.479139947659218, anjing 6.10002202
4478947
Topic: 6
 lacur 56.97048436460543, agama 10.10004876004643, ras 9.411608944711196, mesum 8.099994396168306, twiter 8.09995653113680
4, com 8.099930792909472, media 7.100021220351084, orang 6.364847741523899, jual 6.100053142647907, demokrasi 6.10002990096
1755
Topic: 7
 mesum 2.1001448956667024, org 2.100064838133289, skrg 2.100011354295569, duduk 2.100005646965668, iseng 2.099999999803583
4, gegara 2.0999999997168217, hasil 1.100040421259895, nya 1.1000272434582734, cina 1.100025188701681, parah 1.100023937424
0526
Topic: 8
 lacur 10.488518025497777, ras 8.100149221777906, syahganda 8.099999999695108, buka 5.100057243928322, orang 5.100054285997
409, emang 5.1000509453611125, kpop 5.09999999964672, gak 4.9446389126115395, htps 4.447129214571693, hasil 4.1001277366085
69
Topic: 9
 lacur 56.33942820903878, nya 10.8824002150346, orang 9.053454869108535, mesum 8.519297213112779, gak 7.100095266646303, ot
ak 7.040117200387565, iya 6.889353722015582, hina 5.81791365053057, beda 5.100014331414758, juang 5.100014231682571
```

LSA

In [52]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vektor = TfidfVectorizer(max_features = 1000)
```

In [53]:

```
#menghitung tf-idf dengan TfidfTransformer
vektor_dt = vektor.fit_transform(data['text'])
print(vektor_dt)
print(vektor_dt.shape)
```

```
  (0, 762)      0.15272779359201613
  (0, 242)      0.19192313773966022
  (0, 36)       0.14177393727307744
  (0, 110)      0.18255764560035231
  (0, 607)      0.19192313773966022
  (0, 601)      0.19192313773966022
  (0, 817)      0.14679231387154673
  (0, 190)      0.1752932055260729
  (0, 359)      0.19192313773966022
  (0, 578)      0.19192313773966022
  (0, 925)      0.29358462774309346
  (0, 802)      0.1496249699362145
  (0, 805)      0.12857401747832858
  (0, 134)      0.18255764560035231
  (0, 245)      0.1643393492071342
  (0, 742)      0.19192313773966022
  (0, 756)      0.13545318669050188
  (0, 733)      0.19192313773966022
  (0, 979)      0.18255764560035231
  (0, 389)      0.15999223366629556
  (0, 941)      0.24078494856409796
  (0, 737)      0.19192313773966022
  (0, 695)      0.04390090115685355
  (0, 295)      0.19192313773966022
  (0, 734)      0.3286786984142684
  :     :
  (397, 626)    0.5441070786149599
  (397, 382)    0.3955651449392686
  (397, 475)    0.4969607890072219
  (397, 623)    0.4427114345470066
  (397, 376)    0.2983654185284531
  (397, 695)    0.1244601946296941
  (398, 477)    0.8143305631184415
  (398, 376)    0.5356650091115894
  (398, 695)    0.22344738079620147
  (399, 792)    0.35686995865224147
  (399, 155)    0.2859150865896594
  (399, 712)    0.3339049400532245
  (399, 629)    0.2716809628530919
  (399, 674)    0.2657134797002912
  (399, 726)    0.2657134797002912
  (399, 358)    0.16915875560193464
  (399, 851)    0.23242200425197623
  (399, 552)    0.26031523455388494
  (399, 992)    0.1830994138898145
  (399, 187)    0.1642305438990429
  (399, 960)    0.16713572966240264
  (399, 923)    0.20739716938742805
  (399, 96)     0.24665612798854372
  (399, 217)    0.3403960541036873
  (399, 695)    0.07637811647778488
(400, 1000)
```

In [54]:

```
idf = vektor.idf_
dd = dict(zip(vektor.get_feature_names(), idf))
l = sorted(dd, key = (dd).get)
print(l[0], l[-1])
print(dd['pelacur'])
print(dd['wilde'])
```

```
pelacur wilde
1.3485145296633316
6.300814246746624
```

In [55]:

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

In [56]:

```
# jumlah dokumen * jumlah topik
print(lsa_top.shape)
```

```
(400, 10)
```

In [57]:

```
# top lsa
print(lsa_top)
```

```
[[ 1.48659449e-01 -9.07423341e-02 -9.32955306e-02 ... -1.12854271e-01
   5.71319799e-02 -1.41289657e-01]
 [ 1.17667791e-01 -6.98253277e-03 -9.08301957e-03 ...  1.49051278e-01
   4.12538022e-02  3.94359639e-03]
 [ 2.63552696e-01 -1.22096489e-01  1.58809497e-01 ... -1.89116206e-02
   8.87302284e-02 -8.13867092e-02]
 ...
 [ 1.43796548e-01 -6.33206813e-03 -9.86829494e-02 ...  4.92812579e-02
   8.67508362e-02  1.10058605e-01]
 [ 2.17107931e-01 -1.82523631e-04 -1.11976866e-01 ...  8.87682837e-02
  -2.13490196e-02  7.28173096e-02]
 [ 2.65659717e-01 -1.95614557e-01  3.84057286e-02 ...  1.87547711e-01
  -9.76421356e-02  2.07392673e-02]]
```

In [58]:

```
# 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  :  14.865944900587117
Topic  1  :  -9.074233411497108
Topic  2  :  -9.329553056265286
Topic  3  :  -1.5971514002380256
Topic  4  :  -0.6461870061277563
Topic  5  :  -5.2107829719236065
Topic  6  :  -18.1891684898388
Topic  7  :  -11.28542705524751
Topic  8  :  5.713197985199807
Topic  9  :  -14.128965680771529
```

In [59]:

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

```
(10, 1000)
[[ 0.00725228  0.00192549  0.00399414 ...  0.00162078  0.0292714
   0.00869898]
 [ 0.00018748 -0.00018537 -0.004088   ... -0.00342359  0.03029904
  -0.00146251]
 [-0.00159705 -0.00032499 -0.00182139 ... -0.0034003   0.00997559
  -0.00775767]
 ...
 [-0.01669732  0.0085105  -0.00188842 ... -0.01050738 -0.02939351
   0.00095881]
 [-0.01419099  0.00259324  0.006313   ...  0.00109504 -0.01150681
   0.0032488 ]
 [ 0.00805032  0.00020352 -0.0013523  ... -0.00040257  0.02123483
   0.02292166]]
```

In [60]:

```
# word / kata paling penting dalam setiap topik
vocab = vektor.get_feature_names()
for i, comp in enumerate(lsa_model.components_):
    vocab_comp = zip(vocab, comp)
    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:
pelacur, politik, com, twitter, yg, ini, status, di, dan, https,

Topic 1:
politik, pelacur, negri, la, yusrilnorakdankasarmainnya, tutup01tusuk02, jabatan, cinta, para, yim,

Topic 2:
twitter, com, status, https, politik, pic, tutup01tusuk02, kpucurangrakyatmelawan, kpungejarquickcounthttps, suwandaben,

Topic 3:
politik, rasis, mesum, aja, ini, gak, lu, yg, ya, otak,

Topic 4:
mesum, aku, otak, ini, juga, babi, izinkan, cabul, ulama, gak,

Topic 5:
aku, izinkan, menjadi, tuhan, politik, ini, dan, binatang, jalang, filosofi,

Topic 6:
rasis, pelacur, aja, aku, mereka, kamu, tu, mesum, gak, kau,

Topic 7:
di, ahok, mesum, kecuali, kasur, jadi, negeri, tutup01tusuk02, otak, pic,

Topic 8:
kau, tu, politik, kata, orang, yang, la, kecuali, hina, tidak,

Topic 9:
kau, ini, anjing, muka, eh, tu, jadi, la, nak, pki,
```

Kesimpulan

- Terdapat beberapa perbedaan hasil Topic Modelling menggunakan LDA dan LSA
- Topik pertama dari kedua metode LDA maupun LSA sama dan menghasilkan output "pelacur" sebagai kata teratas
- Sedangkan di metode LDA, kata teratas yang dominan dari 10 topik menghasilkan "pelacur