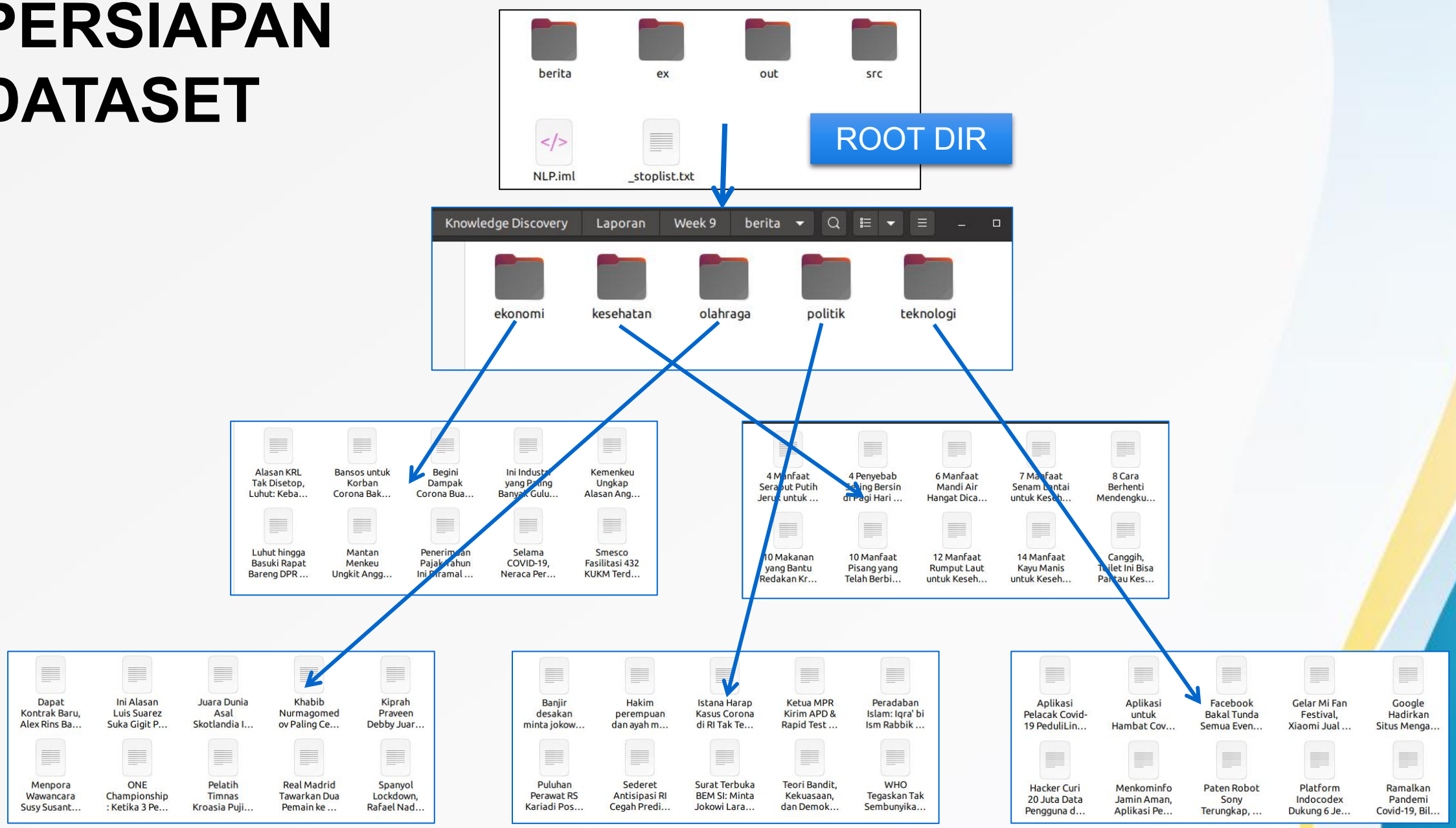
TEXT MINING AND CLASSIFICATION

ARTIKEL BERITA

EKONOMI, KESEHATAN, OLAHRAGA, POLITIK, TEKNOLOGI

Ahmad Rifa'i (2020191009)

PERSIAPAN DATASET



MEMBACA DATA

```
// parent directory
String parentDir = "berita/";
// get child directory (category of news)
ArrayList<String> kategoriDir = FileHelper.getDir(parentDir);
// get all file
ArrayList<List_berita> list_beritas = new ArrayList<>();
for(String s : kategoriDir){
    list_beritas.add(new List_berita(s, FileHelper.getFile( parentDir: parentDir+s+"/")));
}

for(List_berita lb : list_beritas){
    System.out.println("\nKATEGORI\t: "+lb.getKategori());
    System.out.println(Arrays.toString(lb.getFiles().toArray(new String[0])));
}
```

```
public class FileHelper {
    public static ArrayList<String> getDir(String parentDir){
        ArrayList<String> out = new ArrayList<>();
        File folder = new File(parentDir);
        File[] listOfFiles = folder.listFiles();
        for (int i = 0; i < listOfFiles.length; i++) {
            if (listOfFiles[i].isDirectory()) {
                out.add(listOfFiles[i].getName());
            }
        }
        return out;
    }

    public static ArrayList<String> getFile(String parentDir){
        ArrayList<String> out = new ArrayList<>();
        File folder = new File(parentDir);
        File[] listOfFiles = folder.listFiles();
        for (int i = 0; i < listOfFiles.length; i++) {
            if (listOfFiles[i].isFile()) {
                out.add(listOfFiles[i].getName());
            }
        }
        return out;
    }
}
```

```
KATEGORI    : ekonomi
[Alasan KRL Tak Disetop, Luhut: Kebanyakan Ruginya.txt, Bansos untuk Korban Corona Bakal Mengalir S

KATEGORI    : kesehatan
[10 Makanan yang Bantu Redakan Kram Otot, Mudah Ditemukan.txt, 10 Manfaat Pisang yang Telah Berbint

KATEGORI    : olahraga
[Dapat Kontrak Baru, Alex Rins Bakal Lebih Ganas di MotoGP.txt, Ini Alasan Luis Suarez Suka Gigit P

KATEGORI    : politik
[Banjir desakan minta jokowi larang mudik lebaran gegara korona.txt, Hakim perempuan dan ayah magno

KATEGORI    : teknologi
[Aplikasi Pelacak Covid-19 PeduliLindungi Bahaya Bagi Pengguna? Ini Penjelasannya.txt, Aplikasi unt
```

PEMISAHAN DATA

Data training = 70% dan Data testing = 30%

```
Pemisahan_data pemisahanData = new Pemisahan_data(list_beritas,  num_test: 3);
```

```
public Pemisahan_data(List<List_berita> datas, int num_test) {
    this.datas = datas;
    this.num_test = num_test;
    this.proses();
}

private void proses(){
    // testing
    for(List_berita lb : datas){
        List<String> tmpS = new ArrayList<>();
        for(int i = 0; i < num_test; i++){
            tmpS.add(lb.getFiles().get(i));
        }
        testing.add(new List_berita(lb.getKategori(), tmpS));
    }

    // training
    for(List_berita lb : datas){
        List<String> tmpS = new ArrayList<>();
        for(int i = num_test; i < lb.getFiles().size(); i++){
            tmpS.add(lb.getFiles().get(i));
        }
        training.add(new List_berita(lb.getKategori(), tmpS));
    }
}
```

```
-----------------------------
KATEGORI    : ekonomi
data train  : 7
        - Ini Industri yang Paling Banyak Gulung Tikar Gegara Corona.txt
        - Kemenkeu Ungkap Alasan Anggaran Lawan Corona Tak Sebesar Malaysia.txt
        - Luhut hingga Basuki Rapat Bareng DPR Bahas Anggaran Corona, Hasilnya.txt
        - Mantan Menkeu Ungkit Anggaran Corona RI: Infrastruktur Bisa Ditunda.txt
        - Penerimaan Pajak Tahun Ini Diramal Anjlok 8,5%.txt
        - Selama COVID-19, Neraca Perdagangan Hasil Perikanan Meningkat.txt
        - Smesco Fasilitasi 432 KUKM Terdampak COVID-19 Lewat Pelatihan Online.txt

data test   : 3
        - Alasan KRL Tak Disetop, Luhut: Kebanyakan Ruginya.txt
        - Bansos untuk Korban Corona Bakal Mengalir Sampai 2022.txt
        - Begini Dampak Corona Buat Penerimaan Pajak.txt
```

```
-----------------------------
KATEGORI    : kesehatan
data train  : 7
        - 14 Manfaat Kayu Manis untuk Kesehatan dan Kecantikan, Cegah Penuaan Dini.txt
        - 4 Manfaat Serabut Putih Jeruk untuk Kesehatan, Jangan Dibuang.txt
        - 4 Penyebab Sering Bersin di Pagi Hari dan Cara Mencegahnya.txt
        - 6 Manfaat Mandi Air Hangat Dicampur Garam Secara Rutin, Bisa Sembuhkan Luka.txt
        - 7 Manfaat Senam Lantai untuk Kesehatan Tubuh Anak.txt
        - 8 Cara Berhenti Mendengkur Agar Tidur Tidak Mengganggu Orang Lain.txt
        - Canggih, Toilet Ini Bisa Pantau Kesehatan Lewat Kotoran.txt

data test   : 3
        - 10 Makanan yang Bantu Redakan Kram Otot, Mudah Ditemukan.txt
        - 10 Manfaat Pisang yang Telah Berbintik Cokelat bagi Kesehatan.txt
        - 12 Manfaat Rumput Laut untuk Kesehatan dan Kecantikan.txt
```

MEMBUAT FORMAT DATA STANDARD

hasil penggabungan seluruh kata
dari seluruh kategori dokumen

No	title	word1	word2	word3	word4	word5	wordN	label
1	judul1	x	x	x	x	x	x	kategori1
2	judul2	x	x	x	x	x	x	kategori1
3	judul3	....	....	....	....	....	....	kategori1
4	judul4	....	....	....	....	....	....	kategori1
N	judulN	....	....	....	....	....	....	kategoriN

MEMBUAT FORMAT DATA STANDARD

- Mengambil seluruh kata hasil proses text mining pada semua artikel

```
List<Text_tagging> textTaggings = new ArrayList<>();

String txt;
String[] words, keywords;
Map m = null;
ArrayList<String> keyList, tmpKeyList;
ArrayList<Integer> valueList, tmpValueList;
for(List_berita lb : list_beritas){
    for(int i = 0; i < lb.getFiles().size(); i++){
        txt = textMiningLib.readFile( filename: parentDir+lb.getKategori()+"/"+lb.getFiles().get(i));
        words = textMiningLib.Tokenizing(txt);
        words = textMiningLib.Filtering(words);
        keywords = textMiningLib.StemmingTagging(words);
        // temp list
        List<Item_text_tagging> litt = new ArrayList<>();
        tmpKeyList = new ArrayList<>();
        tmpValueList = new ArrayList<>();

        m = textMiningLib.Scoring(keywords);
        //System.out.println(m[i]);
        keyList = new ArrayList<>(m.keySet());
        valueList = new ArrayList<>(m.values());

        for(int j = 0; j < keyList.size(); j++){
            if(valueList.get(j) >= threshold){
                litt.add(new Item_text_tagging(keyList.get(j), valueList.get(j)));
                tmpKeyList.add(keyList.get(j));
                tmpValueList.add(valueList.get(j));
            }
        }
        textTaggings.add(new Text_tagging(
                lb.getKategori(),
                lb.getFiles().get(i),
                tmpKeyList,
                tmpValueList,
                litt)
        );
    }
}

for(Text_tagging tt : textTaggings){
    System.out.println("----------------------------------\nJUDUL \t\t : " +tt.getJudul());
    System.out.println("KATEGORI \t : " +tt.getKategori());
    System.out.println(Arrays.toString(tt.getWords().toArray(new String[0])));
    System.out.println(Arrays.toString(tt.getValues().toArray(new Integer[0])));
}
```

OUTPUT

dengan **threshold = 5**

```
----------------------------------
JUDUL        : Ini Industri yang Paling Banyak Gulung Tikar Gegara Corona.txt
KATEGORI     : ekonomi
[industri, tikar, ktor, gulung, alat]
[12, 5, 5, 5, 5]
----------------------------------
JUDUL        : Kemenkeu Ungkap Alasan Anggaran Lawan Corona Tak Sebesar Malaysia.txt
KATEGORI     : ekonomi
[pdb, negara, anggaran, tanggulang]
[15, 10, 6, 5]
----------------------------------
JUDUL        : Luhut hingga Basuki Rapat Bareng DPR Bahas Anggaran Corona, Hasilnya.txt
KATEGORI     : ekonomi
[menteri, komisi, dpr]
[16, 5, 5]
```

```
JUDUL        : Surat Terbuka BEM SI: Minta Jokowi Larang Mudik hingga Kritik Pembebasan Napi.txt
KATEGORI     : politik
[perintah, bem, rakyat, bijak, indonesia, masyarakat, selamat, kritik, pusat, poin, hati, tulis, surat, remy, penting, oligarki, honorer, guru, buka]
[23, 19, 17, 12, 9, 8, 7, 7, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5]
----------------------------------
JUDUL        : Teori Bandit, Kekuasaan, dan Demokrasi di Masa Pandemi.txt
KATEGORI     : politik
[bandit, kuasa, demokrasi, sistem, tahap, kritik, evolusi, banditisme, anarki, check, balance, anti, and, tetap, teori, rule, orang, main, law, ekonomi, aturan, adab]
[23, 21, 19, 15, 7, 7, 7, 7, 7, 6, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5]
----------------------------------
JUDUL        : WHO Tegaskan Tak Sembunyikan Apapun, Sudah Ingatkan Soal Corona Sejak Awal.txt
KATEGORI     : politik
[who, dros]
[8, 6]
----------------------------------
JUDUL        : Gelar Mi Fan Festival, Xiaomi Jual 50.000 Smartphone dalam Sepekan.txt
KATEGORI     : teknologi
[xiaomi, platform, mff, jual]
[11, 5, 5, 5]
```

PROSES PENGGABUNGAN SELURUH LIST WORD

```
Stream combined = null;
int i = 0;
for(Text_tagging tt : textTaggings){
    System.out.println("-----------------------------------\nJUDUL \t\t : " +tt.getJudul());
    System.out.println("KATEGORI \t : " +tt.getKategori());
    System.out.println(Arrays.toString(tt.getWords().toArray(new String[0])));
    System.out.println(Arrays.toString(tt.getValues().toArray(new Integer[0])));
    if(i == 0) {
        combined = tt.getWords().stream();
    } else {
        combined = Stream.concat(combined, tt.getWords().stream());
    }
    i++;
}
System.out.println("\n\n-------------------\n HASIL PENGGABUNGAN \n-------------------\n");
//System.out.println(combined.map(s -> s).collect(Collectors.toList()));
List<String> lstr = (List<String>) combined.map(s -> s.toString()).collect(Collectors.toList());
Map<String, Long> wordsAllArticles = lstr.stream().collect(Collectors.groupingBy(s -> s, Collectors.counting()));
System.out.println("TOTAL KATA \t"+ wordsAllArticles.keySet().size());
System.out.println(wordsAllArticles.keySet());
```

VARIABEL : wordsAllArticles
TOTAL KATA : 371

```
-------------------
 HASIL PENGGABUNGAN
-------------------

TOTAL KATA  371
[xiaomi, sebar, ring, surat, tanding, rins, putih, robot, muncul, stres, bangun, kanan, lapa, barat, iode, serat, protein, masuk, unduh, banditisme, chiellini, tuga
```

Format data akhir

Setelah proses penggabungan seluruh kata hasil proses text mining

No	title	word1	word2	word3	word4	....	word371	label
1	judul1	x	x	x	x	x	x	kategori1
2	judul2	x	x	x	x	x	x	kategori1
3	judul3	....	....	....	....	....	....	kategori1
4	judul4	....	....	....	....	....	....	kategori1
N	judulN	....	....	....	....	....	....	kategoriN

Total feature (words atau kata-kata) dari hasil proses text mining, dengan threshold = 5, didapatkan sebanyak **371 feature**

MEMBUAT DATA TRAINING DAN DATA TESTING

No	Title	word1	word2	word3	word4	...	word371	label
1	Judul Berita1	0	0	0	10	...	5	ekonomi
2	Judul Berita2	0	0	0	5	...	3	ekonomi
3	Judul Berita3	1	2	5	1	....	1	politik
4	Judul Berita4	1	1	5	1	....	1	politik
5	Judul Berita5	3	3	6	1	....	1	kesehatan
6	Judul Berita6	4	4	9	1	....	1	olahraga
7	Judul Berita7	5	5	1	1	....	1	teknologi
8	Judul Berita8	5	5	4	4	...	1	ekonomi
...	...	....	....	....	....	....	....	...
35	Judul BeritaN	10	10	5	1	...	2	olahraga

Apabila pada data training tidak terdapat “kata” pada sekumpulan “kata” yang telah ada dari seluruh dokumen hasil proses text mining, maka nilai untuk kata tersebut akan **di nol kan**

MEMBUAT DATA TRAINING DAN TESTING

```java
public class DataMaker {
    public static void run(Set<String> allWords, List<Text_tagging> text_taggings, String filename){
        FileWriter writer = null;
        try {  writer = new FileWriter(filename); } catch (IOException e) { e.printStackTrace(); }
        StringBuffer sb = new StringBuffer();
        // add header
        sb.append("title;");
        sb.append(allWords.stream().collect(Collectors.joining( delimiter: ";")));
        sb.append(";label\n");
        for(Text_tagging tg : text_taggings){
            ArrayList<String> v = new ArrayList<>();
            v.add(tg.getJudul());
            for(String s : allWords){
                boolean isThere = false;
                for(String tgs : tg.getWords()){
                    if(s.equalsIgnoreCase(tgs)){
                        isThere = true;
                        break;
                    }
                }
                int idx = Helper.getIndex(tg.getWords(), s);
                if(idx == 999999){
                    v.add("0"); // data not found
                } else {
                    v.add(String.valueOf(tg.getValues().get(idx)));
                }
            }
            v.add(tg.getKategori());
            // add body
            String collect = v.stream().collect(Collectors.joining( delimiter: ";"));
            sb.append(collect);
            sb.append("\n");
        }
        try {
            writer.write(sb.toString());
            writer.close();
        } catch (IOException e) {
            e.printStackTrace();
        }
    }
}
```

Data Training

```java
// MEMBUAT DATA TRAINING
List<Text_tagging> textTaggingsTraining = TextMining.stemmingTagging(
        pemisahanData.getTraining(),
        parentDir,
        threshold
);
DataMaker.run(wordsAllArticles.keySet(), textTaggingsTraining, filename: "training.csv");
```

Data Testing

```java
// MEMBUAT DATA TESTING
List<Text_tagging> textTaggingsTesting = TextMining.stemmingTagging(
        pemisahanData.getTesting(),
        parentDir,
        threshold
);
DataMaker.run(wordsAllArticles.keySet(), textTaggingsTesting, filename: "testing.csv");
```

OUTPUT FILE DATA TRAINING

```
title;xiaomi;sebar;ring;surat;tanding;rins;putih;robot;muncul;stres;bangun;kanan;lapa;barat;iode;serat;protein;masuk;unduh;banditisme;chiellini;tugas;aturan;jerawat;sistem;gelar;suara;bill;pr
Ini Industri yang Paling Banyak Gulung Tikar Gegara Corona.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;12;0;0;0;0;0;
Kemenkeu Ungkap Alasan Anggaran Lawan Corona Tak Sebesar Malaysia.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
Luhut hingga Basuki Rapat Bareng DPR Bahas Anggaran Corona, Hasilnya.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
Mantan Menkeu Ungkit Anggaran Corona RI: Infrastruktur Bisa Ditunda.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
Penerimaan Pajak Tahun Ini Diramal Anjlok 8,5%.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;8;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
Selama COVID-19, Neraca Perdagangan Hasil Perikanan Meningkat.txt;0;0;0;0;0;0;0;0;0;0;0;0;0;0;8;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;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
Smesco Fasilitasi 432 KUKM Terdampak COVID-19 Lewat Pelatihan Online.txt;0;0;0;0;0;0;0;0;0;0;0;0;0;13;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;6;0;7;0;0;0;0;0;0;0;0;0;0;0;0;0;
14 Manfaat Kayu Manis untuk Kesehatan dan Kecantikan, Cegah Penuaan Dini.txt;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;8;0;0;0;0;0;54;0;0;0;0;0;0;0;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
4 Manfaat Serabut Putih Jeruk untuk Kesehatan, Jangan Dibuang.txt;0;0;0;0;0;0;21;0;0;0;0;0;0;0;0;8;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;5;9;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
4 Penyebab Sering Bersin di Pagi Hari dan Cara Mencegahnya.txt;0;0;0;0;0;0;0;0;0;0;5;0;0;0;0;0;0;7;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;5;0;0;0;0;0;0;0;0;0;0
6 Manfaat Mandi Air Hangat Dicampur Garam Secara Rutin, Bisa Sembuhkan Luka.txt;0;0;0;0;0;0;0;0;0;5;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;0;0;0;0;0;0;0;0;0;
7 Manfaat Senam Lantai untuk Kesehatan Tubuh Anak.txt;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;0;0;0;0;0;0;0;0;0;0;0;12;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
8 Cara Berhenti Mendengkur Agar Tidur Tidak Mengganggu Orang Lain.txt;0;0;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;7;0;0;0;0;0;0;0;13;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
Canggih, Toilet Ini Bisa Pantau Kesehatan Lewat Kotoran.txt;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;5;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;
Khabib Nurmagomedov Paling Cepat Kembali Bertarung Agustus.txt;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;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
Kiprah Praveen Debby Juarai All England 2016 Tanpa Didampingi Pelatih.txt;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;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
Menpora Wawancara Susy Susanti di Hari Kartini.txt;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;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
ONE Championship: Ketika 3 Petarung Indonesia Rindu Duel di Atas Oktagon.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;7;0;0;0
Pelatih Timnas Kroasia Puji Performa Marko Simic di Persija.txt;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;0;0;6;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
Real Madrid Tawarkan Dua Pemain ke Manchester United.txt;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;0;0;0;0;0;6;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
Spanyol Lockdown, Rafael Nadal Frustrasi Tak Bisa Latihan.txt;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;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;
Ketua MPR Kirim APD & Rapid Test COVID-19 ke Jateng.txt;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;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;
Peradaban Islam: Iqra' bi Ism Rabbik (2).txt;0;0;0;0;0;0;0;0;0;0;0;5;16;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;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;5;0;0;0;0;0;0;0;0;
Puluhan Perawat RS Kariadi Postif Corona, Istana: Kuncinya Orang Harus Jujur.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
Sederet Antisipasi RI Cegah Prediksi 106 Ribu Kasus Corona Terjadi.txt;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;6;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;8;0;0;0;0;0
Surat Terbuka BEM SI: Minta Jokowi Larang Mudik hingga Kritik Pembebasan Napi.txt;0;0;0;5;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
Teori Bandit, Kekuasaan, dan Demokrasi di Masa Pandemi.txt;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;7;0;0;5;0;15;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
WHO Tegaskan Tak Sembunyikan Apapun, Sudah Ingatkan Soal Corona Sejak Awal.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
Gelar Mi Fan Festival, Xiaomi Jual 50.000 Smartphone dalam Sepekan.txt;11;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
Google Hadirkan Situs Mengajar dari Rumah versi Bahasa Indonesia.txt;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;13;0;0;0;0;0;9;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;5;0;0;0;0;
Hacker Curi 20 Juta Data Pengguna dari Toko Aplikasi Pihak Ketiga Androi.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
Menkominfo Jamin Aman, Aplikasi PeduliLindungi Sudah Diunduh 1,9 Juta Kali.txt;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;5;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;0
Paten Robot Sony Terungkap, Bisa Ngobrol dan Jadi Teman Curhat Main Gim.txt;0;0;0;0;0;0;11;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;0;0;0;0;0;0;0;0;0;0;0;0;0
Platform Indocodex Dukung 6 Jenis Aset Kripto.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;6;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;9;
Ramalkan Pandemi Covid-19, Bill Gates Timbun Makanan di Basement.txt;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;7;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
```

OUTPUT FILE DATA TRAINING

```
title;xiaomi;sebar;ring;surat;tanding;rins;putih;robot;muncul;stres;bangun;kanan;lapa;barat;iode;serat;protein;masuk;unduh;banditisme;chiellini;tugas;aturan;jerawat;sistem;gelar;suara;bill;pr
Alasan KRL Tak Disetop, Luhut: Kebanyakan Ruginya.txt;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;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;0;0;0
Bansos untuk Korban Corona Bakal Mengalir Sampai 2022.txt;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;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
Begini Dampak Corona Buat Penerimaan Pajak.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;9;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;9
10 Makanan yang Bantu Redakan Kram Otot, Mudah Ditemukan.txt;0;0;0;0;0;0;0;0;0;0;6;0;0;0;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;9;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
10 Manfaat Pisang yang Telah Berbintik Cokelat bagi Kesehatan.txt;0;0;0;0;0;0;0;0;0;0;6;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;10;0;0;0
12 Manfaat Rumput Laut untuk Kesehatan dan Kecantikan.txt;0;0;7;0;0;0;0;0;0;0;0;0;0;0;14;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;6;0;0;0;0;0;0;0;0;0;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;5;0;0
Dapat Kontrak Baru, Alex Rins Bakal Lebih Ganas di MotoGP.txt;0;0;0;0;0;8;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;6;0;0;0;0;0;0
Ini Alasan Luis Suarez Suka Gigit Pemain Lawan.txt;0;0;0;0;7;0;0;0;0;0;0;0;0;0;0;0;0;0;7;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0
Juara Dunia Asal Skotlandia Ingin Menjajal Ketangguhan Manny Pacquiao.txt;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;5;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;8;0;0
Banjir desakan minta jokowi larang mudik lebaran gegara korona.txt;0;9;0;7;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;0;0;0;8;0;0;0;0;0;8;0;0;0;0;0;0;10
Hakim perempuan dan ayah magnosis.txt;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;0;0;0;0;0;6;0;0;0;0;0;0;0;0;0;0;0;0;0;0;8;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
Istana Harap Kasus Corona di RI Tak Tembus 106 Ribu Seperti Prediksi.txt;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;7;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;0
Aplikasi Pelacak Covid-19 PeduliLindungi Bahaya Bagi Pengguna? Ini Penjelasannya.txt;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;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
Aplikasi untuk Hambat Covid-19 Diimbau Tidak Lacak Lokasi Pengguna.txt;0;6;0;0;0;0;0;0;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;9;0;15;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;5;0;0;0;5;0;0;0;0;0
Facebook Bakal Tunda Semua Event Offline hingga 2021.txt;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;12;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
```

PROSES KLASIFIKASI

Load data training

```
DATA = pd.read_csv("training.csv", delimiter=";")
DATA
```

title	xiaomi	sebar	ring	surat	tanding	rins	putih	robot	muncul	...	bluetooth	pelaku	bantal	matang	utara	konsums	bijak	udara	tidur	label
yang nyak Tikar G...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	ekonomi
nkeu asan aran rona T...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	ekonomi
ngga apat DPR ng...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	ekonomi
nkeu aran a RI: fra...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	ekonomi
naan n Ini njlok 5%...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	ekonomi

Proses Training dan Testing menggunakan dataset yang sama

```
X = DATA.iloc[:, 1:-1].values
y = DATA.iloc[:,-1].values
print(X)
print(y)
```

```
[[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]]
['ekonomi' 'ekonomi' 'ekonomi' 'ekonomi' 'ekonomi' 'ekonomi' 'ekonomi'
 'kesehatan' 'kesehatan' 'kesehatan' 'kesehatan' 'kesehatan' 'kesehatan'
 'kesehatan' 'olahraga' 'olahraga' 'olahraga' 'olahraga' 'olahraga'
 'olahraga' 'olahraga' 'politik' 'politik' 'politik' 'politik' 'politik'
 'politik' 'politik' 'teknologi' 'teknologi' 'teknologi' 'teknologi'
 'teknologi' 'teknologi' 'teknologi']
```

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit(X, y)
y_predict = knn.predict(X)
print("Akurasi : ", (y_predict == y).mean())
```

```
Akurasi :  0.37142857142857144
```

```
from sklearn.svm import SVC
svc = SVC()
svc.fit(X, y)
y_predict = svc.predict(X)
print("Akurasi : ", (y_predict == y).mean())
```

```
Akurasi :  0.8285714285714286
```

Didapatkan akurasi yang bagus diperoleh dengan menggunakan algoritma SVC. dengan nilai akurasi = 82,8%

Load data testing

DATA TESTING

```
TESTING = pd.read_csv("testing.csv", delimiter=";")
TESTING
```

	title	xiaomi	sebar	ring	surat	tanding	rins	putih	robot	muncul	...	bluetooth	pelaku	bantal	matang	utara	konsums	bijak	udara	ti
0	Alasan KRL Tak Disetop, Luhut: Kebanyakan Rugi...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
1	Bansos untuk Korban Corona Bakal Mengalir Samp...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
2	Begini Dampak Corona Buat Penerimaan Pajak.txt	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
3	10 Makanan yang Bantu Redakan Kram Otot, Mudah...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	

Proses testing data

```
X_test = TESTING.iloc[:, 1:-1].values
y_test = TESTING.iloc[:,-1].values
print(X_test)
print(y_test)
```

```
[[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 6 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
['ekonomi' 'ekonomi' 'ekonomi' 'kesehatan' 'kesehatan' 'kesehatan'
 'olahraga' 'olahraga' 'olahraga' 'politik' 'politik' 'politik'
 'teknologi' 'teknologi' 'teknologi']
```

KNN

```
y_predict = knn.predict(X_test)
print("Akurasi : ", (y_predict == y_test).mean())
TESTING['knn'] = y_predict
```

```
Akurasi :  0.3333333333333333
```

SVC

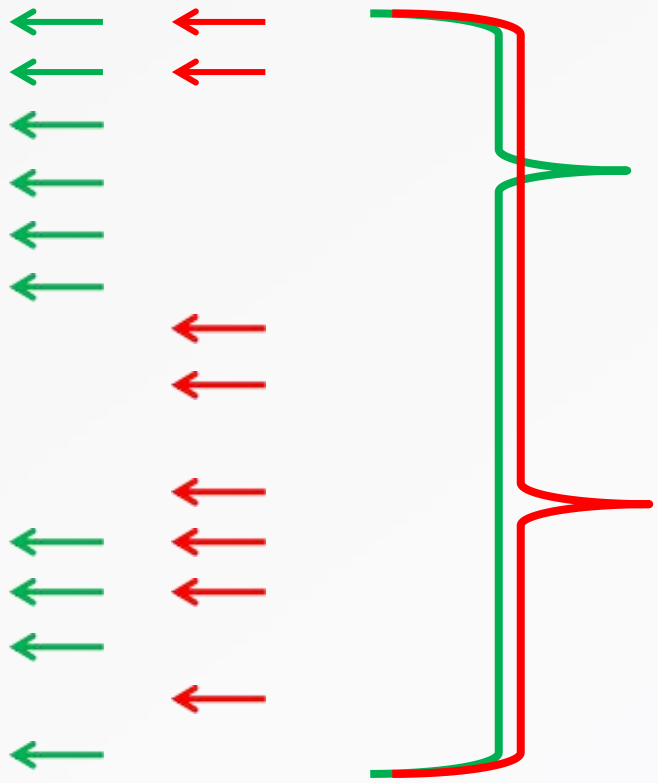
```
y_predict = svc.predict(X_test)
print("Akurasi : ", (y_predict == y_test).mean())
TESTING['svc'] = y_predict
```

```
Akurasi :  0.4666666666666667
```

Dari hasil proses testing didapatkan akurasi yang bagus diperoleh dengan menggunakan algoritma SVC. dengan nilai akurasi = 46,7%

Missclassified

	title	label	knn	svc
0	Alasan KRL Tak Disetop, Luhut: Kebanyakan Rugi...	ekonomi	olahraga	teknologi
1	Bansos untuk Korban Corona Bakal Mengalir Samp...	ekonomi	olahraga	teknologi
2	Begini Dampak Corona Buat Penerimaan Pajak.txt	ekonomi	olahraga	ekonomi
3	10 Makanan yang Bantu Redakan Kram Otot, Mudah...	kesehatan	olahraga	kesehatan
4	10 Manfaat Pisang yang Telah Berbintik Cokelat...	kesehatan	olahraga	kesehatan
5	12 Manfaat Rumput Laut untuk Kesehatan dan Kec...	kesehatan	olahraga	kesehatan
6	Dapat Kontrak Baru, Alex Rins Bakal Lebih Gana...	olahraga	olahraga	teknologi
7	Ini Alasan Luis Suarez Suka Gigit Pemain Lawan...	olahraga	olahraga	teknologi
8	Juara Dunia Asal Skotlandia Ingin Menjajal Ket...	olahraga	olahraga	olahraga
9	Banjir desakan minta jokowi larang mudik lebar...	politik	politik	kesehatan
10	Hakim perempuan dan ayah magnosis.txt	politik	olahraga	kesehatan
11	Istana Harap Kasus Corona di RI Tak Tembus 106...	politik	olahraga	teknologi
12	Aplikasi Pelacak Covid-19 PeduliLindungi Bahay...	teknologi	olahraga	teknologi
13	Aplikasi untuk Hambat Covid-19 Diimbau Tidak L...	teknologi	teknologi	kesehatan
14	Facebook Bakal Tunda Semua Event Offline hingg...	teknologi	olahraga	teknologi



KNN MissClassified

10

SVC MissClassified

8

TERIMAKASIH