python

NLP-Text Mining

Text Preprocessing

Sutriawan, S.Kom., M.Kom
UNIVERSITAS MUHAMMADIYAH BIMA

sutriawan@umbima.ac.id | Sutriawan1612@gmail.com

Pertemuan 1

A. Import data dari file Excel

Umumnya data *text* didalam bahasa pemrograman python berbentuk list, dictionary atau tuple. Data *text* tersebut dapat diambil dari DBMS *(Data Base Management System)* atau *file* dengan ekstensi json, xlsx, csv ataupun *file* lain yang sejenis. Berikut dijelaskan cara untuk melakukan *import* data komentar hotel ekstensi xlsx (Microsoft Excel), data tersebut lengkap dengan (Nama Hotel, Alamat Hotel, Komentar, Waktu Komentar, Nama Orang yang Memberikan Komentar) dapat dilihat pada Gambar di bawah ini.

	A	B	C	D	E
1	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Menu breakfast tidak mencerminkan hotel bintang 4	19 Desember 2016	Indrawati
2	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Telepon di kamar tidak bisa digunakan, kamar mandi yg terlalu kecil,	12 September 2016	Rahman
3	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Fasilitas yg lumayan lengkap, lokasi yg strategis, staff yg ramah	12 September 2016	Aris
4	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Breakfast banyak pilihan menu, kolam renang ok, cake slice di lobby	13 Agustus 2016	Indri
5	favehotel Premier	Jl. Cihampelas 129 Bandung	Tidak ada tissue dikamar tidur jg tidak disediakan teh celup, kopi dar	12 Februari 2017	Dwi
6	favehotel Premier	Jl. Cihampelas 129 Bandung	Area menuju kolam renang kurang bersih	14 Januari 2017	Joko
7	favehotel Premier	Jl. Cihampelas 129 Bandung	Fasilitas dikamar mandi rusak/copot	08 Januari 2017	Bagas
8	favehotel Premier	Jl. Cihampelas 129 Bandung	Tidak tersedia extra bed	03 Januari 2017	Estri
9	favehotel Premier	Jl. Cihampelas 129 Bandung	Lokasi dekat dengan pusat belanja dan tempat makan	27 Desember 2016	Yessi

Langkah *import* data xlsx kedalam Python List dengan menggunakan jupyter notebook :

1. *Import library* yang akan digunakan.

```
from openpyxl import load_workbook
import pandas as pd
```

   - `openpyxl` adalah *library* yang digunakan untuk *read* dan *write* file Excel (xlsx/xlsm/xltx/xltm)
   - `pandas` adalah *library data analysis*, digunakan untuk mengolah data secara terstruktur.

2. Inisialisasi *file* excel yang akan di *import*

```
wb = load_workbook(filename = 'komentar hotel.xlsx')
sheet_ranges = wb['Sheet1']

df = pd.DataFrame(sheet_ranges.values)
df.columns = ['Hotel', 'Alamat', 'Komentar', 'Tanggal','User']

df
```

   - `load_workbook` merupakan nama *function* dari *library* openpyxl yang digunakan untuk melakukan *import* data dari excel (kemudian disimpan dalam variabel wb).

- `Sheet_ranges` merupakan *variable* yang menampung data dari sheet mana yang akan diambil dalam file excel (pada contoh ini adalah Sheet1).
- `DataFrame` adalah *function* dari *library* pandas yang digunakan untuk melakukan parsing data terstruktur kedalam bentuk kolom dan baris, dengan demikian data yang telah diparsing akan menjadi sebuah *table* yang nampak seperti susunan pada *relational database*, dimana sebuah baris tunggal mewakili sebuah contoh tunggal dan kolom mewakili atribut tertentu.
  (Kemudian dimasukkan ke dalam variabel `df`).
- `Columns` *function* dari *library* pandas yang digunakan untuk memberikan header pada setiap kolom data.
- Hasil dari dataframe `df` dapat dilihat pada Tabel di bawah ini:

	Hotel	Alamat	Komentar	Tanggal	User
0	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Menu breakfast tidak mencerminkan hotel bintang 4	2016-12-19	Indrawati
1	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Telepon di kamar tidak bisa digunakan, kamar m...	2016-09-12	Rahman
2	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Fasilitas yg lumayan lengkap, lokasi yg strate...	2016-09-12	Aris
3	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Breakfast banyak pilihan menu, kolam renang ok...	2016-08-13	Indri
4	favehotel Premier	Jl. Cihampelas 129 Bandung	Tidak ada tissue dikamar tidur jg tidak disedi...	2017-02-12	Dwi
5	favehotel Premier	Jl. Cihampelas 129 Bandung	Area menuju kolam renang kurang bersih	2017-01-14	Joko
6	favehotel Premier	Jl. Cihampelas 129 Bandung	Fasilitas dikamar mandi rusak/copot	2017-01-08	Bagas
7	favehotel Premier	Jl. Cihampelas 129 Bandung	Tidak tersedia extra bed	2017-01-03	Estri
8	favehotel Premier	Jl. Cihampelas 129 Bandung	Lokasi dekat dengan pusat belanja dan tempat m...	2016-12-27	Yessi

3. Mengambil data pada DataFrame

   Data dapat diambil berdasarkan indeks kolom tertentu dengan mendefinisikan urutan indeks kolom pada *variable* DataFrame (`df`).

   a. Menampilkan data pada kolom komentar.

```
df['Komentar']
```

```
0    Menu breakfast tidak mencerminkan hotel bintang 4
1    Telepon di kamar tidak bisa digunakan, kamar m...
2    Fasilitas yg lumayan lengkap, lokasi yg strate...
3    Breakfast banyak pilihan menu, kolam renang ok...
4    Tidak ada tissue dikamar tidur jg tidak disedi...
5               Area menuju kolam renang kurang bersih
6                  Fasilitas dikamar mandi rusak/copot
7                              Tidak tersedia extra bed
8    Lokasi dekat dengan pusat belanja dan tempat m...
Name: Komentar, dtype: object
```

b. Menampilkan data pada kolom komentar dengan bentuk tabel.

```
df[['Komentar']]
```

	Komentar
0	Menu breakfast tidak mencerminkan hotel bintang 4
1	Telepon di kamar tidak bisa digunakan, kamar m...
2	Fasilitas yg lumayan lengkap, lokasi yg strate...
3	Breakfast banyak pilihan menu, kolam renang ok...
4	Tidak ada tissue dikamar tidur jg tidak disedi...
5	Area menuju kolam renang kurang bersih
6	Fasilitas dikamar mandi rusak/copot
7	Tidak tersedia extra bed
8	Lokasi dekat dengan pusat belanja dan tempat m...

c. Menampilkan data dengan jumlah tertentu

```
df[:3]
```

	Hotel	Alamat	Komentar	Tanggal	User
0	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Menu breakfast tidak mencerminkan hotel bintang 4	2016-12-19	Indrawati
1	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Telepon di kamar tidak bisa digunakan, kamar m...	2016-09-12	Rahman
2	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Fasilitas yg lumayan lengkap, lokasi yg strate...	2016-09-12	Aris

d. Menampilkan data secara Ascending atau Descending

```
df[::-1]
```

	Hotel	Alamat	Komentar	Tanggal	User
8	favehotel Premier	Jl. Cihampelas 129 Bandung	Lokasi dekat dengan pusat belanja dan tempat m...	2016-12-27	Yessi
7	favehotel Premier	Jl. Cihampelas 129 Bandung	Tidak tersedia extra bed	2017-01-03	Estri
6	favehotel Premier	Jl. Cihampelas 129 Bandung	Fasilitas dikamar mandi rusak/copot	2017-01-08	Bagas
5	favehotel Premier	Jl. Cihampelas 129 Bandung	Area menuju kolam renang kurang bersih	2017-01-14	Joko
4	favehotel Premier	Jl. Cihampelas 129 Bandung	Tidak ada tissue dikamar tidur jg tidak disedi...	2017-02-12	Dwi
3	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Breakfast banyak pilihan menu, kolam renang ok...	2016-08-13	Indri
2	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Fasilitas yg lumayan lengkap, lokasi yg strate...	2016-09-12	Aris
1	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Telepon di kamar tidak bisa digunakan, kamar m...	2016-09-12	Rahman
0	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Menu breakfast tidak mencerminkan hotel bintang 4	2016-12-19	Indrawati

e. Menampilkan data secara Ascending atau Descending berdasarkan kolom *komentar*.

```
df.sort_values(['Komentar'], ascending=[0])
```

	Hotel	Alamat	Komentar	Tanggal	User
7	favehotel Premier	Jl. Cihampelas 129 Bandung	Tidak tersedia extra bed	2017-01-03	Estri
4	favehotel Premier	Jl. Cihampelas 129 Bandung	Tidak ada tissue dikamar tidur jg tidak disedi...	2017-02-12	Dwi
1	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Telepon di kamar tidak bisa digunakan, kamar m...	2016-09-12	Rahman
0	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Menu breakfast tidak mencerminkan hotel bintang 4	2016-12-19	Indrawati
8	favehotel Premier	Jl. Cihampelas 129 Bandung	Lokasi dekat dengan pusat belanja dan tempat m...	2016-12-27	Yessi
2	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Fasilitas yg lumayan lengkap, lokasi yg strate...	2016-09-12	Aris
6	favehotel Premier	Jl. Cihampelas 129 Bandung	Fasilitas dikamar mandi rusak/copot	2017-01-08	Bagas
3	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Breakfast banyak pilihan menu, kolam renang ok...	2016-08-13	Indri
5	favehotel Premier	Jl. Cihampelas 129 Bandung	Area menuju kolam renang kurang bersih	2017-01-14	Joko

f. Menampilkan data beberapa kolom tertentu

```
df[['Hotel','Komentar']]
```

	Hotel	Komentar
0	Discovery Hotel	Menu breakfast tidak mencerminkan hotel bintang 4
1	Discovery Hotel	Telepon di kamar tidak bisa digunakan, kamar m...
2	Discovery Hotel	Fasilitas yg lumayan lengkap, lokasi yg strate...
3	Discovery Hotel	Breakfast banyak pilihan menu, kolam renang ok...
4	favehotel Premier	Tidak ada tissue dikamar tidur jg tidak disedi...
5	favehotel Premier	Area menuju kolam renang kurang bersih
6	favehotel Premier	Fasilitas dikamar mandi rusak/copot
7	favehotel Premier	Tidak tersedia extra bed
8	favehotel Premier	Lokasi dekat dengan pusat belanja dan tempat m...

g. Memasukkan data beberapa kolom tertentu ke dalam list python

```
df[['Hotel','Komentar']].values.tolist()
```

```
[[u'Discovery Hotel', u'Menu breakfast tidak mencerminkan hotel bintang 4'],
 [u'Discovery Hotel',
  u'Telepon di kamar tidak bisa digunakan, kamar mandi yg terlalu kecil, kurangnya penanda arah ke fasilitas hotel..'],
 [u'Discovery Hotel',
  u'Fasilitas yg lumayan lengkap, lokasi yg strategis, staff yg ramah'],
 [u'Discovery Hotel',
  u'Breakfast banyak pilihan menu, kolam renang ok, cake slice di lobby enak & kalau malam disc 30% \u263a'],
 [u'favehotel Premier',
  u'Tidak ada tissue dikamar tidur jg tidak disediakan teh celup, kopi dan gula selain air mineral. Untuk sekelas dan sehar
ga itu sangat tidak sepadan sama sekali.'],
 [u'favehotel Premier', u'Area menuju kolam renang kurang bersih'],
 [u'favehotel Premier', u'Fasilitas dikamar mandi rusak/copot'],
 [u'favehotel Premier', u'Tidak tersedia extra bed'],
 [u'favehotel Premier', u'Lokasi dekat dengan pusat belanja dan tempat makan']]
```

h. Menampilkan data pada baris/*row* tertentu

```
df.iloc[1]
```

```
Hotel                                    Discovery Hotel
Alamat                          Jl. Lodan Timur No. 7 Jakarta
Komentar    Telepon di kamar tidak bisa digunakan, kamar m...
Tanggal                                  2016-09-12 00:00:00
User                                                  Rahman
Name: 1, dtype: object
```

```
df.iloc[[1]]
```

	Hotel	Alamat	Komentar	Tanggal	User
1	Discovery Hotel	Jl. Lodan Timur No. 7 Jakarta	Telepon di kamar tidak bisa digunakan, kamar m...	2016-09-12	Rahman

i. Menampilkan spesifik field (data pada kolom dan baris tertentu)

```
df[[3]].iloc[[1]]
```

	Tanggal
1	2016-09-12

```
df.loc[1,'Tanggal']
```

```
Timestamp('2016-09-12 00:00:00')
```

j. Memasukkan data pada kolom *komentar* ke dalam list python

```
list_komentar = []
list_komentar.append(df['Komentar'].values.tolist())

print list_komentar
```

- Deklarasi list array dengan nama `list_komentar`
- Menambahkan isi `dataframe` indeks ke-2 ke dalam list `list_komentar`.
- Hasil print dari `list_komentar`

```
[[u'Menu breakfast tidak mencerminkan hotel bintang 4', u'Telepon di kamar tidak bisa digunakan, kamar mandi yg terlalu kec
il, kurangnya penanda arah ke fasilitas hotel..', u'Fasilitas yg lumayan lengkap, lokasi yg strategis, staff yg ramah', u'B
reakfast banyak pilihan menu, kolam renang ok, cake slice di lobby enak & kalau malam disc 30% \u263a', u'Tidak ada tissue
dikamar tidur jg tidak disediakan teh celup, kopi dan gula selain air mineral. Untuk sekelas dan seharga itu sangat tidak s
epadan sama sekali.', u'Area menuju kolam renang kurang bersih', u'Fasilitas dikamar mandi rusak/copot', u'Tidak tersedia e
xtra bed', u'Lokasi dekat dengan pusat belanja dan tempat makan']]
```

- Menampilkan `list_komentar` satu persatu.

```
for komentar in list_komentar:
    for komen in komentar:
        print komen
```

```
Menu breakfast tidak mencerminkan hotel bintang 4
Telepon di kamar tidak bisa digunakan, kamar mandi yg terlalu kecil, kurangnya penanda arah ke fasilitas hotel..
Fasilitas yg lumayan lengkap, lokasi yg strategis, staff yg ramah
Breakfast banyak pilihan menu, kolam renang ok, cake slice di lobby enak & kalau malam disc 30% ☺
Tidak ada tissue dikamar tidur jg tidak disediakan teh celup, kopi dan gula selain air mineral. Untuk sekelas dan seharga i
tu sangat tidak sepadan sama sekali.
Area menuju kolam renang kurang bersih
Fasilitas dikamar mandi rusak/copot
Tidak tersedia extra bed
Lokasi dekat dengan pusat belanja dan tempat makan
```

B. Mengubah penulisan kalimat menjadi *lower case*

Umumnya penggunakan huruf besar dan huruf kecil pada data teks komentar tidak konsisten, problem seperti itu akan berpengaruh pada proses analisis data terlebih pada penghitungan TF *(Term Frekuensi)*. Untuk itu perlu dilakukan proses konversi data teks menjadi keseluruhan huruf kecil. Berikut tahap konversi penulisan menjadi *lower case*.

```
kecil = ("PELAYANAN HOTELNYA SANGAT BAGUS").lower()
besar = ("pelayanan hotelnya sangat bagus").upper()

print kecil
print besar
```

```
pelayanan hotelnya sangat bagus
PELAYANAN HOTELNYA SANGAT BAGUS
```

C. Tokenisasi

Ada dua macam tokenisasi yang digunakan dalam *pre-processing*, diantaranya adalah proses pemecahan paragraf kedalam kalimat dan proses pemecahan kalimat kedalam kata. Proses pemecahan tersebut dapat dijelaskan sebagai berikut.

- **Memecah paragraf ke dalam kalimat.**

  Proses pemecahan paragraf ke dalam kalimat umumnya didasarkan pada penggunakan tanda baca, dengan contoh ? "tanda tanya", . "titik" dan tanda baca sejenis yang menunjukkan kalimat telah diakhiri. Berikut dijelaskan beberapa teknik pemecahan paragraf kedalam kalimat.

  1. Teknik pemecahan menggunakan split.

```
sentences = ("Pelayanan hotelnya bagus sekali. Menu makanannya juga enak.")
split_sentence= sentences.split(".")

print split_sentence
```

```
['Pelayanan hotelnya bagus sekali', ' Menu makanannya juga enak', '']
```

  - Dijelaskan split diatas melakukan pemecahan atau pemenggalan kalimat berdasarkan tanda baca . "titik".

  2. Teknik pemecahan menggunakan regular expression.

```
import re
```

```
sentences = ("Pelayanan hotelnya bagus sekali. Menu makanannya juga enak.")
split_sentence= re.split(r'(?<=[^a-z].[.?]) +(?=[a-z])', sentences)

print split_sentence
```

```
['Pelayanan hotelnya bagus sekali. Menu makanannya juga enak.']
```

  - `re` adalah class *reguler expression* yang umumnya digunakan untuk *string matching*.
  - `split` digunakan untuk *spliting* kata dari hasil pencarian *regular expression* yang dihasilkan.

  3. Teknik pemecahan menggunakan *function* sent_tokenize yang diimport dari class tokenize library nltk *(Natural Language Toolkit)*

```
from nltk import tokenize
```

```
sentences = ("Pelayanan hotelnya bagus sekali. Menu makanannya juga enak.")

split_sentence=tokenize.sent_tokenize(sentences)

print split_sentence
```

```
['Pelayanan hotelnya bagus sekali.', 'Menu makanannya juga enak.']
```

- **Memecah kalimat ke dalam kata**

  Proses pemecahan kalimat ke dalam kata umumnya didasarkan pada spasi antar kata. Berikut dijelaskan beberapa contoh teknik pemecahan kalimat kedalam kata.

  - Teknik pemecahan menggunakan split.

```
kalimat = ("Pelayanan hotelnya sangat bagus").split()

print kalimat
```

```
['Pelayanan', 'hotelnya', 'sangat', 'bagus']
```

  - Teknik pemecahan menggunakan function word_tokenize yang diimport dari class tokenize library nltk *(Natural Language Toolkit)*

```
from nltk.tokenize import word_tokenize
```

```
token = word_tokenize("Menu breakfast tidak mencerminkan hotel bintang 4")
print token
```

```
['Menu', 'breakfast', 'tidak', 'mencerminkan', 'hotel', 'bintang', '4']
```

D. Stopword Removal

Stopword removal merupakan metode *filtering* yang sering diterapkan pada proses *pre-processing* di dalam *text mining*. Penerapan *stopword removal* tersebut bertujuan untuk menghilangkan noise terhadap kata yang akan di proses ke dalam tahap analisis. Cara kerja *stopword removal* tersebut, dengan menghilangkan sejumlah kelas kata penghubung ataupun kata yang jumlahnya banyak namun tidak mempengaruhi konten dokumen secara keseluruhan untuk dapat dilakukan analisis data *text*. Berikut dijelaskan langkah dari penerapan *stopword removal*.

1. Definisi daftar kata stopword removal yang digunakan dalam proses *pre-processing*

```
data_stopword = ["yg","di"]
```

2. Proses *pre-processing* dapat dilakukan seperti gambar di bawah ini.

```
sentence = ("Fasilitas yg lumayan lengkap, lokasi yg strategis, staff yg ramah")

words_tokenize = []
words = sentence.split()
for word in words:
    check = word in data_stopword
    if not check :
        words_tokenize.append(word)

print words_tokenize
```

```
['Fasilitas', 'lumayan', 'lengkap,', 'lokasi', 'strategis,', 'staff', 'ramah']
```

   - Dapat dijelaskan *sentence* adalah kalimat yang akan dilakukan *pre-processing*
   - words_tokenize adalah definisi list.

E. Stemming

Stemming adalah metode *mapping* token ke bentuk kata dasar, namun bentuk kata dasar tersebut tidak berarti sama dengan akar kata (*root word*). Istilah *stemming* sering diartikan merubah kata berimbuhan menjadi kata dasar.

Misalnya :

- menahan => tahan
- berbalas-balasan => balas

Sastrawi merupakan Library stemming yang sering digunakan untuk pre-processing teks dalam bahasa indonesia. Sedangkan untuk kata bahasa inggris dapat menggunakan library stemming dari NLTK salah satunya adalah algoritma porter stemming. Berikut dijelaskan langkah – langkah stemming pada proses pre-processing.

- Stemming kata berbahasa Indonesia (Sastrawi Stemming)
  1. Sastrawi Stemming

     Untuk menginstall Sastrawi *stemming* dapat menggunakan pip dengan instruksi sebagai berikut: `pip install Sastrawi`

  2. Proses Stemming
     - *Import library stemming* sastrawi seperti berikut:

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
```

     - Proses pemanggilan *function* create_stemmer dari class StemmerFactory:

```
factory = StemmerFactory()
stemmer = factory.create_stemmer()

print stemmer.stem("berbalas-balasan")
```

```
balas
```

- Stemming kata berbahasa Inggris (Porter Stemming)
  - Import library porter sebagai berikut:

```
from nltk.stem.porter import PorterStemmer
```

  - Proses pemanggilan function stem dari class PorterStemer:

```
stemmer = PorterStemmer()

print stemmer.stem("meeting")
```

```
meet
```

Exercise

1. Import data dari excel

   Terdapat data review online shop pada file excel seperti dibawah ini.

	A	B	C	D	E
1	12/01/2016	Doki Store	Jakarta	Sudah sampai gan. Mantap barangnya..	
2	13/01/2016	Doki Store	Jakarta	kualitas bgs .. dpt bonus lagi dr penjual	
3	06/01/2017	Doki Store	Jakarta	Barang rapih packingnya n bagus. Cuma pengirimannya yang agak lama	
4	15/03/2016	Doki Store	Jakarta	Makasih barang sudah sampai dengan selamat sampai tujuan.	
5	05/07/2016	Doki Store	Jakarta	pihak penjual memperhatikan dan mengikuti perkembangan proses pengiriman dengan s	
6	17/01/2016	Icad Shop	Yogakarta	Respon cepat, pengiriman super cepat, packing aman dan rapi.	
7	28/06/2016	Icad Shop	Yogakarta	Puas belanja disini. Mudahan laris terus ya gan.	
8	19/05/2016	Icad Shop	Yogakarta	packing rapih, barang sesuai dgn yg diiklankan. nice and recommended seller	
9	07/08/2016	D'Stores	Surabaya	packing nya mantapp yaa aman banget .. seller juga Mau bertanggung jawab kalo ni bara	
10	21/09/2016	D'Stores	Surabaya	barangnya mulus, pengirimannya juga cepat. makasih ya sis	
11	13/03/2016	D'Stores	Surabaya	Produk diterima sesuai deskripsi. Berfungsi dengan baik	

   a. Buatlah *DataFrame* untuk menampung data online shop tersebut.

      Hasil:

	Tanggal	Shop	Kota	Review
0	2016-01-12	Doki Store	Jakarta	Sudah sampai gan. Mantap barangnya...
1	2016-01-13	Doki Store	Jakarta	kualitas bgs .. dpt bonus lagi dr penjual
2	2017-01-06	Doki Store	Jakarta	Barang rapih packingnya n bagus. Cuma pengirim...
3	2016-03-15	Doki Store	Jakarta	Makasih barang sudah sampai dengan selamat sam...
4	2016-07-05	Doki Store	Jakarta	pihak penjual memperhatikan dan mengikuti perk...
5	2016-01-17	Icad Shop	Yogakarta	Respon cepat, pengiriman super cepat, packing ...
6	2016-06-28	Icad Shop	Yogakarta	Puas belanja disini. Mudahan laris terus ya gan.
7	2016-05-19	Icad Shop	Yogakarta	packing rapih, barang sesuai dgn yg diiklankan...
8	2016-08-07	D'Stores	Surabaya	packing nya mantapp yaa aman banget .. seller ...
9	2016-09-21	D'Stores	Surabaya	barangnya mulus, pengirimannya juga cepat. mak...
10	2016-03-13	D'Stores	Surabaya	Produk diterima sesuai deskripsi. Berfungsi de...

   b. Ambil kolom *Shop* dan *Review*, kemudian sorting *Ascending* berdasarkan kolom *Shop*.

      Hasil:

	Shop	Review
8	D'Stores	packing nya mantapp yaa aman banget .. seller ...
9	D'Stores	barangnya mulus, pengirimannya juga cepat. mak...
10	D'Stores	Produk diterima sesuai deskripsi. Berfungsi de...
0	Doki Store	Sudah sampai gan. Mantap barangnya...
1	Doki Store	kualitas bgs .. dpt bonus lagi dr penjual
2	Doki Store	Barang rapih packingnya n bagus. Cuma pengirim...
3	Doki Store	Makasih barang sudah sampai dengan selamat sam...
4	Doki Store	pihak penjual memperhatikan dan mengikuti perk...
5	Icad Shop	Respon cepat, pengiriman super cepat, packing ...
6	Icad Shop	Puas belanja disini. Mudahan laris terus ya gan.
7	Icad Shop	packing rapih, barang sesuai dgn yg diiklankan...

c. Masukan data pada kolom *Review* ke dalam *list* python dengan nama *list_review*.

```
[[u'Sudah sampai gan. Mantap barangnya..', u'kualitas bgs .. dpt bonus lagi dr penjual', u'Barang rapih packingnya n bagus.
Cuma pengirimannya yang agak lama', u'Makasih barang sudah sampai dengan selamat sampai tujuan.', u'pihak penjual memperhat
ikan dan mengikuti perkembangan proses pengiriman dengan seksama,pelayanan yang baik dan aman menurut pemantauan saya,terim
akasih', u'Respon cepat, pengiriman super cepat, packing aman dan rapi.', u'Puas belanja disini. Mudahan laris terus ya gan
.', u'packing rapih, barang sesuai dgn yg diiklankan. nice and recommended seller', u'packing nya mantapp yaa aman banget .
. seller juga Mau bertanggung jawab kalo ni barang mengalami kerusakan.. Keren lah barang juga Ori thank you yaaa', u'baran
gnya mulus, pengirimannya juga cepat. makasih ya sis', u'Produk diterima sesuai deskripsi. Berfungsi dengan baik']]
```

2. Tokenization

a. Lakukan lowercase kalimat dalam *list_review*.

Hasil:

```
sudah sampai gan. mantap barangnya..
kualitas bgs. dpt bonus lagi dr penjual
barang rapih packingnya n bagus. cuma pengirimannya yang agak lama
makasih barang sudah sampai dengan selamat sampai tujuan.
pihak penjual memperhatikan dan mengikuti perkembangan proses pengiriman dengan seksama,pelayanan yang baik dan aman menuru
t pemantauan saya,terimakasih
respon cepat, pengiriman super cepat, packing aman dan rapi.
puas belanja disini. mudahan laris terus ya gan.
packing rapih, barang sesuai dgn yg diiklankan. nice and recommended seller
packing nya mantapp yaa aman banget. seller juga mau bertanggung jawab kalo ni barang mengalami kerusakan. keren lah barang
juga ori thank you yaaa
barangnya mulus, pengirimannya juga cepat. makasih ya sis
produk diterima sesuai deskripsi. berfungsi dengan baik
```

b. Lalukan pemecahan kalimat dalam *list_review* yang memiliki lebih dari satu kalimat dalam *review*.

```
1 sudah sampai gan. mantap barangnya..
2 kualitas bgs. dpt bonus lagi dr penjual
3 barang rapih packingnya n bagus. cuma pengirimannya yang agak lama
4 makasih barang sudah sampai dengan selamat sampai tujuan.
5 pihak penjual memperhatikan dan mengikuti perkembangan proses pengiriman dengan seksama,pelayanan yang baik dan aman menu
rut pemantauan saya,terimakasih
6 respon cepat, pengiriman super cepat, packing aman dan rapi.
7 puas belanja disini. mudahan laris terus ya gan.
8 packing rapih, barang sesuai dgn yg diiklankan. nice and recommended seller
9 packing nya mantapp yaa aman banget. seller juga mau bertanggung jawab kalo ni barang mengalami kerusakan. keren lah bara
ng juga ori thank you yaaa
10 barangnya mulus, pengirimannya juga cepat. makasih ya sis
11 produk diterima sesuai deskripsi. berfungsi dengan baik
```

Hasil:

```
1 sudah sampai gan.
2 mantap barangnya..
3 kualitas bgs.
4 dpt bonus lagi dr penjual
5 barang rapih packingnya n bagus.
6 cuma pengirimannya yang agak lama
7 makasih barang sudah sampai dengan selamat sampai tujuan.
8 pihak penjual memperhatikan dan mengikuti perkembangan proses pengiriman dengan seksama,pelayanan yang baik dan aman menu
rut pemantauan saya,terimakasih
9 respon cepat, pengiriman super cepat, packing aman dan rapi.
10 puas belanja disini.
11 mudahan laris terus ya gan.
12 packing rapih, barang sesuai dgn yg diiklankan.
13 nice and recommended seller
14 packing nya mantapp yaa aman banget.
15 seller juga mau bertanggung jawab kalo ni barang mengalami kerusakan.
16 keren lah barang juga ori thank you yaaa
17 barangnya mulus, pengirimannya juga cepat.
18 makasih ya sis
19 produk diterima sesuai deskripsi.
20 berfungsi dengan baik
```

c. Lakukan pemecahan kata pada *review*.

Hasil:

```
1 ['sudah', 'sampai', 'gan', '.']
2 ['mantap', 'barangnya..']
3 ['kualitas', 'bgs', '.']
4 ['dpt', 'bonus', 'lagi', 'dr', 'penjual']
5 ['barang', 'rapih', 'packingnya', 'n', 'bagus', '.']
6 ['cuma', 'pengirimannya', 'yang', 'agak', 'lama']
7 ['makasih', 'barang', 'sudah', 'sampai', 'dengan', 'selamat', 'sampai', 'tujuan', '.']
8 ['pihak', 'penjual', 'memperhatikan', 'dan', 'mengikuti', 'perkembangan', 'proses', 'pengiriman', 'dengan', 'seksama', ',
', 'pelayanan', 'yang', 'baik', 'dan', 'aman', 'menurut', 'pemantauan', 'saya', ',', 'terimakasih']
9 ['respon', 'cepat', ',', 'pengiriman', 'super', 'cepat', ',', 'packing', 'aman', 'dan', 'rapi', '.']
10 ['puas', 'belanja', 'disini', '.']
11 ['mudahan', 'laris', 'terus', 'ya', 'gan', '.']
12 ['packing', 'rapih', ',', 'barang', 'sesuai', 'dgn', 'yg', 'diiklankan', '.']
13 ['nice', 'and', 'recommended', 'seller']
14 ['packing', 'nya', 'mantapp', 'yaa', 'aman', 'banget', '.']
15 ['seller', 'juga', 'mau', 'bertanggung', 'jawab', 'kalo', 'ni', 'barang', 'mengalami', 'kerusakan', '.']
16 ['keren', 'lah', 'barang', 'juga', 'ori', 'thank', 'you', 'yaaa']
17 ['barangnya', 'mulus', ',', 'pengirimannya', 'juga', 'cepat', '.']
18 ['makasih', 'ya', 'sis']
19 ['produk', 'diterima', 'sesuai', 'deskripsi', '.']
20 ['berfungsi', 'dengan', 'baik']
```

3. Stopword

a. Buat kumpulan data stopword untuk diterapkan pada pre-processing data teks review online shop.

```
data_stopword = ["yg","yang", "lah", "juga", "."]
```

b. Lakukan proses stopword removal pada data teks review online shop yang telah diproses pada tahap tokenisasi.

Hasil:

```
['sudah', 'sampai', 'gan']
['mantap', 'barangnya..']
['kualitas', 'bgs']
['dpt', 'bonus', 'lagi', 'dr', 'penjual']
['barang', 'rapih', 'packingnya', 'n', 'bagus']
['cuma', 'pengirimannya', 'agak', 'lama']
['makasih', 'barang', 'sudah', 'sampai', 'dengan', 'selamat', 'sampai', 'tujuan']
['pihak', 'penjual', 'memperhatikan', 'dan', 'mengikuti', 'perkembangan', 'proses', 'pengiriman', 'dengan', 'seksama', ',',
'pelayanan', 'baik', 'dan', 'aman', 'menurut', 'pemantauan', 'saya', ',', 'terimakasih']
['respon', 'cepat', ',', 'pengiriman', 'super', 'cepat', ',', 'packing', 'aman', 'dan', 'rapi']
['puas', 'belanja', 'disini']
['mudahan', 'laris', 'terus', 'ya', 'gan']
['packing', 'rapih', ',', 'barang', 'sesuai', 'dgn', 'diiklankan']
['nice', 'and', 'recommended', 'seller']
['packing', 'nya', 'mantapp', 'yaa', 'aman', 'banget']
['seller', 'mau', 'bertanggung', 'jawab', 'kalo', 'ni', 'barang', 'mengalami', 'kerusakan']
['keren', 'barang', 'ori', 'thank', 'you', 'yaaa']
['barangnya', 'mulus', ',', 'pengirimannya', 'cepat']
['makasih', 'ya', 'sis']
['produk', 'diterima', 'sesuai', 'deskripsi']
['berfungsi', 'dengan', 'baik']
```

4. Stemming

   Lakukan proses stemming pada data review menggunakan sastrawi stemming.

   Hasil:

```
['sudah', 'sampai', 'gan']
['mantap', 'barang']
['kualitas', 'bgs']
['dpt', 'bonus', 'lagi', 'dr', 'jual']
['barang', 'rapih', 'packingnya', 'n', 'bagus']
['cuma', 'kirim', 'agak', 'lama']
['makasih', 'barang', 'sudah', 'sampai', 'dengan', 'selamat', 'sampai', 'tuju']
['pihak', 'jual', 'perhati', 'dan', 'ikut', 'kembang', 'proses', 'kirim', 'dengan', 'seksama', '', 'layan', 'baik', 'dan',
'aman', 'turut', 'pantau', 'saya', '', 'terimakasih']
['respon', 'cepat', '', 'kirim', 'super', 'cepat', '', 'packing', 'aman', 'dan', 'rapi']
['puas', 'belanja', 'sini']
['mudah', 'laris', 'terus', 'ya', 'gan']
['packing', 'rapih', '', 'barang', 'sesuai', 'dgn', 'iklan']
['nice', 'and', 'recommended', 'seller']
['packing', 'nya', 'mantapp', 'yaa', 'aman', 'banget']
['seller', 'mau', 'tanggung', 'jawab', 'kalo', 'ni', 'barang', 'alami', 'rusa']
['keren', 'barang', 'ori', 'thank', 'you', 'yaaa']
['barang', 'mulus', '', 'kirim', 'cepat']
['makasih', 'ya', 'sis']
['produk', 'terima', 'sesuai', 'deskripsi']
['fungsi', 'dengan', 'baik']
```