Text Mining

PreProcesing #2

| Outline

1. Stemming/Lemmatization
2. POS Tagging
3. Term Weighting (TF-IDF)

| #1 Stemming/Lemmatization

- **Stemming** adalah proses pengubahan bentuk kata menjadi kata dasar atau tahap mencari root kata dari tiap kata hasil filtering.
- **Lemmatization** adalah proses yang bertujuan untuk melakukan normalisasi pada teks dengan berdasarkan pada bentuk dasar yang merupakan bentuk lemmanya.
- Tujuan dari stemming dan lemmatization adalah untuk mengurangi bentuk infleksi dan kadang-kadang terkait bentuk Word ke bentuk dasar yang umum.

| #1 Stemming/Lemmatization (cont)

- **Lemmatisasi** terkait erat dengan **stemming**. Perbedaannya adalah bahwa stemmer beroperasi pada satu kata tunggal tanpa pengetahuan konteks, dan oleh karena itu tidak dapat membedakan antara kata-kata yang memiliki arti berbeda tergantung pada bagian dari dokumen. Namun, stemmer biasanya lebih mudah diimplementasikan dan dijalankan lebih cepat, dan akurasi yang berkurang mungkin tidak masalah untuk beberapa aplikasi.
- Contohnya:
  - Kata "lebih baik" memiliki "baik" sebagai lemma-nya. Tautan ini tidak terjawab oleh stemming, karena memerlukan pencarian kamus.
- Lemmatization : Stemming berdasarkan **kamus.**

#1 Stemming/Lemmatization (cont)

- Implementasi proses **stemming** sangat beragam , tergantung dengan **bahasa** dari dokumen.
- Beberapa metode untuk Stemming :
  - Porter Stemmer (English & Indonesia)
  - Stemming Arifin-Setiono (Indonesia)
  - Stemming Nazief-Adriani (Indonesia)

Hasil Token	Hasil Filtering	Hasil Stemming
they	-	-
are	-	-
applied	applied	apply
to	-	-
the	-	-
words	words	word
in	-	-
the	-	-
texts	texts	text

Hasil Token	Hasil Filtering	Hasil Stemming
namanya	namanya	nama
adalah	-	-
santiago	santiago	santiago
santiago	santiago	santiago
sudah	-	-
memutuskan	memutuskan	putus
untuk	-	-
mencari	mencari	cari
sang	-	-
alkemis	alkemis	alkemis

| #2 POS Tagging

- Part-of-speech (POS) tagging atau secara singkat dapat ditulis sebagai tagging merupakan proses pemberian penanda POS atau kelas sintaktik pada tiap kata di dalam corpus, seperti kata benda, kata kerja, kata sifat, dll.

```
2019-10-02 15:21:09,495 loading file resources/taggers/example-universal-pos/best-model.pt
saya <PRON> dan <CCONJ> dia <PRON> kemarin <ADV> pergi <VERB> ke <ADP> pasar <NOUN> bersama <ADP> untuk <ADP> membeli <VERB> jeruk <NOUN>
```

Keterangan label

ADJ : kata sifat
ADP : preposisi
ADV : keterangan
AUX : kata bantu
CCONJ : kata penghubung
INTJ : kata seru
NOUN : kata benda
NUM : angka
PART : partikel
PRON : kata ganti
PUNCT : tanda baca
SYM : simbol
VERB : kata kerja
X : lainnya

| #3 Term Weighting (TF-IDF)

- Term dapat berupa kata, frase atau unit hasil indexing lainnya dalam suatu dokumen yang dapat digunakan untuk mengetahui konteks dari dokumen tersebut. Karena setiap kata memiliki tingkat kepentingan yang berbeda dalam dokumen, maka untuk setiap kata tersebut diberikan sebuah indikator, yaitu term weight (Zafikri, 2008).
- Metode TF-IDF merupakan metode pembobotan term yang banyak digunakan sebagai metode pembanding terhadap metode pembobotan baru. Pada metode ini, perhitungan bobot term t dalam sebuah dokumen dilakukan dengan mengalikan nilai Term Frequency dengan Inverse Document Frequency.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

| #3 Term Weighting (TF-IDF) (simulasi)

Contoh Kasus : 1 paragraf yang terdiri dari **3 kalimat**

> Pelayanan hotel memuaskan. Menu sarapan bervariasi dan lengkap. Fasilitas hotel lengkap, pelayanannya bagus, kamarnya luas, hotel ini cocok untuk tempat menginap bersama keluarga.

Term(t)	D1(dokumen 1)	D2	D3
Pelayanan	1	0	1
Hotel	1	0	2
Memuaskan	1	0	0
Menu	0	1	0
Sarapan	0	1	0
Bervariasi	0	1	0
Lengkap	0	1	1
Fasilitas	0	0	1
Bagus	0	0	1
Kamar	0	0	1
Luas	0	0	1
Cocok	0	0	1
Menginap	0	0	1
Keluarga	0	0	1

Term(t)	DF
Pelayanan	2
Hotel	2
Memuaskan	1
Menu	1
Sarapan	1
Bervariasi	1
Lengkap	2
Fasilitas	1
Bagus	1
Kamar	1
Luas	1
Cocok	1
Menginap	1
Keluarga	1

1. Menghitung Document Frequency (DF), yaitu banyaknya dokumen dimana suatu term (t) muncul. Contoh berdasarkan soal yang sama pada poin pertama

| #3 Term Weighting (TF-IDF) (simulasi)

2. Menghitung Invers Document Frequency (IDF)

Term(t)	DF	IDF
Pelayanan	2	log(3/2)=0,176
Hotel	2	log(3/2)=0.176
Memuaskan	1	log(3/1)=0,477
Menu	1	log(3/1)=0,477
Sarapan	1	log(3/1)=0,477
Bervariasi	1	log(3/1)=0,477
Lengkap	2	log(3/2)=0.176
Fasilitas	1	log(3/1)=0,477
Bagus	1	log(3/1)=0,477
Kamar	1	log(3/1)=0,477
Luas	1	log(3/1)=0,477
Cocok	1	log(3/1)=0,477
Menginap	1	log(3/1)=0.477
Keluarga	1	log(3/1)=0.477

3. Menghitung nilai TF x IDF

Term(t)	TF			IDF	TF-IDF		
	D1	D2	D3		D1	D2	D3
Pelayanan	1	0	1	0,176	0,176	0	0,176
Hotel	1	0	2	0,176	0,176	0	0.352
Memuaskan	1	0	0	0,477	0,477	0	0
Menu	0	1	0	0,477	0	0,477	0
Sarapan	0	1	0	0,477	0	0,477	0
Bervariasi	0	1	0	0,477	0	0,477	0
Lengkap	0	1	1	0.176	0	0,176	0,176
Fasilitas	0	0	1	0,477	0	0	0,477
Bagus	0	0	1	0,477	0	0	0,477
Kamar	0	0	1	0,477	0	0	0,477
Luas	0	0	1	0,477	0	0	0,477
Cocok	0	0	1	0,477	0	0	0,477
Menginap	0	0	1	0.477	0	0	0,477
Keluarga	0	0	1	0.477	0	0	0,477

| #3 Term Weighting (TF-IDF) (simulasi)

4. Hasil Akhir

dok\tf-idf	pelayanan	hotel	memuaskan	menu	sarapan	bervariasi	lengkap	fasilitas	bagus	kamar	luas	cocok	menginap	keluarga	KELAS
D1	0.176	0.176	0.477	0	0	0	0	0	0	0	0	0	0	0	
D2	0	0	0	0.477	0.477	0.477	0.176	0	0	0	0	0	0	0	
D3	0.176	0.352	0	0	0	0	0.176	0.477	0.477	0.477	0.477	0.477	0.477	0.477	

Thank you

</ti>
FST | UINAM
teknik informatika

UNIVERSITAS ISLAM NEGERI
ALAUDDIN
MAKASSAR