Text Mining

PreProcesing #1

| Outline

1. Text Preprocessing Overview
2. Parsing & Tokenization
3. Case Folding
4. Filtering/Stopword

| #1 Preprocessing Overview

Karakteristik Dokumen Teks

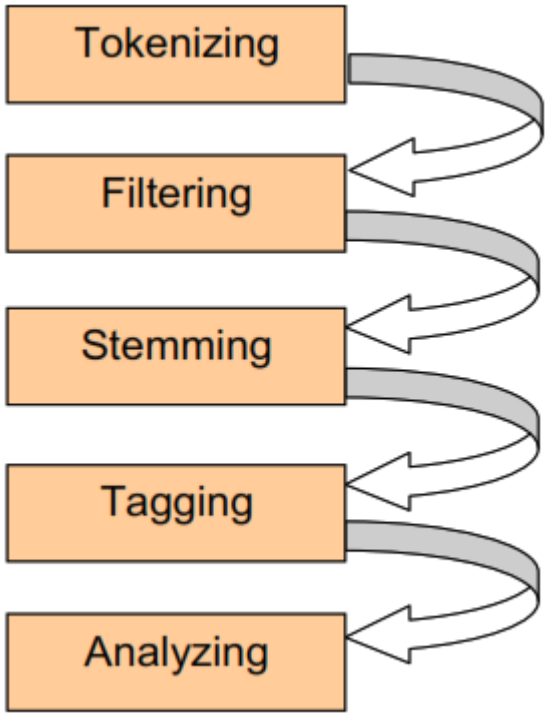
Menurut Loretta Auvila dan Duane Searsmith dari University of Illinois, karakteristik dokumen teks:

- Database berukuran besar
- Memiliki dimensi yang tinggi, yaitu satu kata merupakan satu dimensi
- Mengandung kumpulan kata yang saling terkait (frase) dan antara kumpulan kata satu dengan lain dapat memiliki arti yang berbeda
- Banyak mengandung kata ataupun arti bias (ambiguity)
- Dokumen e-mail, chat merupakan dokumen yang tidak memiliki struktrur bahasa yang baku, terdapat banyak istilah *slank* “r u there?”, “hellooo boooss, whatzzzuuppp?”, “Siap boscuu”.

| #1 Preprocessing Overview (con't)

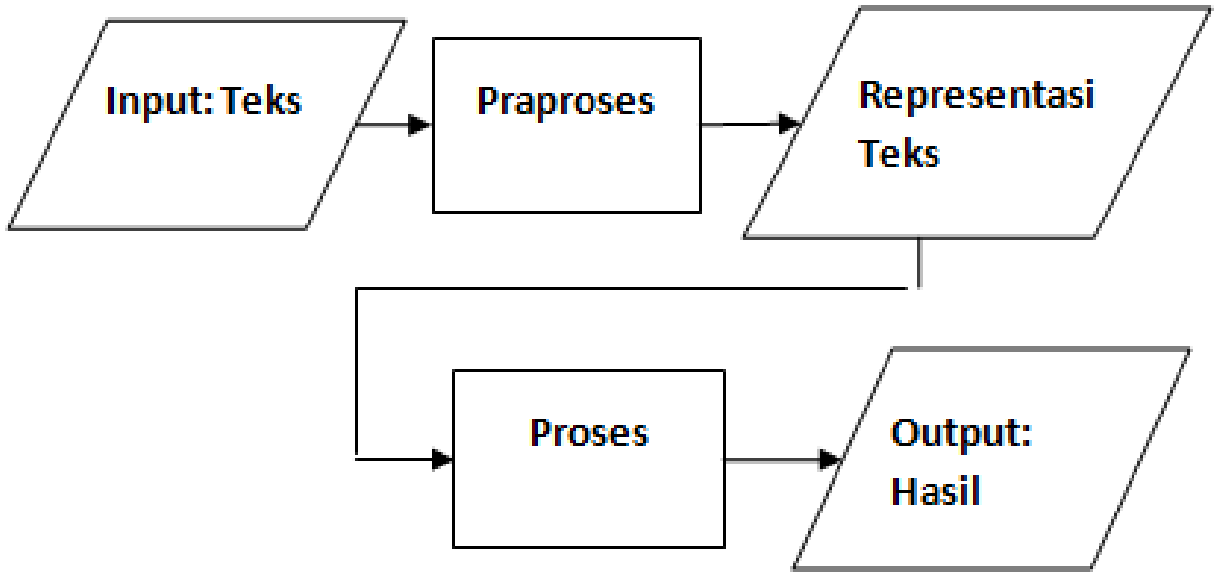
- Berdasarkan ketidakteraturan struktur data teks,maka proses text mining memerlukan beberapa tahap awal pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur

Tahapan text preprocessing secara umum:



| #1 Preprocessing Overview (Representasi Teks)

- Teks masukan bisa mempunyai satuan yang berbeda, tergantung aplikasi/permasalahannya.



| #1 Preprocessing Overview (Representasi Teks)

- Teks: tidak terstruktur atau semi terstruktur.
- Contoh teks

Harga Minyak Dunia Melonjak

REPUBLIKA.CO.ID, NEW YORK -- Harga minyak mentah melonjak setelah Badan Energi Internasional (IEA) mengatakan surplus minyak mentah global mulai menyusut, meskipun data AS menunjukkan peningkatan besar dalam persediaan domestiknya karena Badai Harvey. Harga minyak mentah AS, West Texas Intermediate (WTI) untuk pengiriman Oktober naik 1,07 dolar AS atau 2,2 persen menjadi menetap di 49,30 dolar AS per barel di New York Mercantile Exchange.

Sementara itu, harga patokan Eropa, minyak mentah Brent North Sea untuk pengiriman November naik 89 sen menjadi 55,16 dolar AS per barel di London ICE Futures Exchange. Harga bensin AS turun meski terjadi rekor penarikan dalam cadangan bahan bakar minyak

...

Sumber: <http://www.republika.co.id/berita/ekonomi/bisnis-global/17/09/14/ow8sj3382-harga-minyak-dunia-melonjak>

| #1 Preprocessing Overview (Representasi Teks)

- Bentuk terstruktur: matriks/tabel.
- Field, bisa beberapa jenis (biner, skor nilai, string, diskret).

	Field1	Field2	Field3
Obyek1			
Obyek2			
Obyek3			

| #1 Preprocessing Overview (Representasi Teks)

- Himpunan field: data isian field merepresentasikan masing2 obyek.
- Memungkinan adanya field yang tidak bisa diisi pada obyek tertentu. Misalnya “pekerjaan” untuk seorang bayi.

| #1 Preprocessing Overview (Representasi Teks)

- Memungkinan adanya field yang tidak bisa diisi pada obyek tertentu. Misalnya “pekerjaan” untuk seorang bayi.

	Field1	Field2	Field3
Obyek1	X		
Obyek2			
Obyek3		X	

| #1 Preprocessing Overview (Representasi Teks)

- Fitur atau Atribut, kadang disebut juga /Variabel/Prediktor.

Himpunan atribut

	Atribut1	Atribut2	Atribut3	
Teks1				
Teks2				
Teks3				

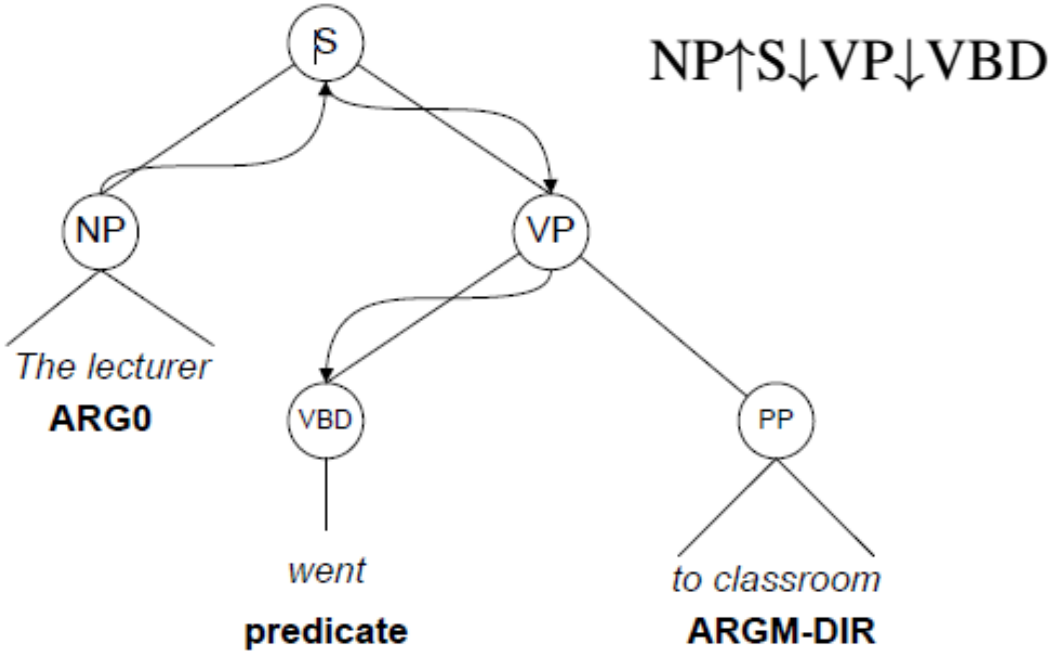
#1 Preprocessing Overview (Representasi Teks)

	Atribut1	Atribut2	Atribut3	Kategori
Berita1				
Berita2				
Berita3				

	Atribut1	Atribut2	Atribut3	Kategori
setelah				
IAE				
mengatakan				

| #1 Preprocessing Overview (Representasi Teks)

- Atribut bisa berupa, al:
  - teks atau non teks.
  - data dari pemrosesan morfologis, sintaksis, dan semantik
    - Contoh unt pelabelan peran semantik, salah satu atributnya adalah jalur dari teks ke predikatnya. Unt teks *The lecturer*, jalurnya:

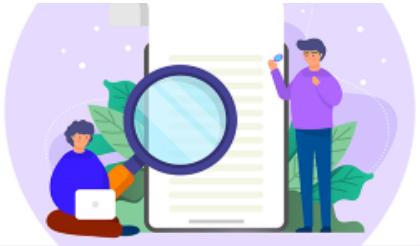


visit :
[1]http://www.bowdoin.edu/~allen/nlp/nlp6.html
[2]https://socs.binus.ac.id/2013/06/22/natural-language-processing/

| #1 Preprocessing Overview (Representasi Teks)

- Salah satu jenis atribut yang penting adalah kata yang ada dalam koleksi teks.
- Salah satu efeknya, matriks lebih banyak berisi nol (matriks yang jarang / *sparse*).

	Kata1	Kata2	Kata3		Kategori
Berita1					
Berita2					
Berita3					



#1 Preprocessing Overview (Representasi Teks)

- Pada teks panjang (misal paper di jurnal) umumnya tataran semantis (misalnya sinonim) tdk diperhatikan.
- Pada teks pendek (misal twit), tataran semantis biasanya perlu.

| #2 Parsing & Tokenization

Parsing Dokumen berurusan dengan **pengenalan dan "pemecahan"** struktur dokumen menjadi komponen-komponen terpisah. Pada langkah preprocessing ini, kita menentukan mana yang dijadikan **satu unit dokumen**;

- Contoh lain, buku dengan **100 halaman** bisa dipisah menjadi **100 dokumen**; masing-masing halaman menjadi 1 dokumen
- **Satu *tweet*** bisa dijadikan sebagai **1 dokumen**. Begitu juga dengan sebuah komentar pada forum atau review produk.

| #2 Parsing & Tokenization (cont)

- Tokenisasi adalah proses **pemotongan** string input berdasarkan tiap kata penyusunnya.
- Pada prinsipnya proses ini adalah **memisahkan** setiap kata yang menyusun suatu dokumen.

Pada proses ini dilakukan **penghilangan angka, tanda baca dan karakter selain huruf alfabet**, karena karakter-karakter tersebut dianggap sebagai pemisah kata (delimiter) dan tidak memiliki pengaruh terhadap pemrosesan teks.

Manajemen pengetahuan adalah sebuah konsep baru di dunia bisnis.

[Teks Input]

manajemen
pengetahuan
adalah
sebuah
konsep
baru
di
dunia
bisnis

[Hasil Token]

| #3 Case Folding

Pada tahapan ini juga dilakukan proses ***case folding***, dimana semua huruf diubah menjadi huruf kecil.

| #4 Filtering/StopWord

- Disebut juga **Filtering**
- **Filtering** adalah tahap **pemilihan** kata-kata penting dari hasil token, yaitu kata-kata apa saja yang akan digunakan untuk mewakili dokumen.
- Bisa menggunakan algoritma **stop list** (membuang kata yang kurang penting) atau **word list** (menyimpan kata penting)

manajemen
pengetahuan
adalah
sebuah
konsep
baru
di
dunia
bisnis

→

manajemen
pengetahuan
konsep
baru
dunia
bisnis

TASK

Text PreProcessing

(Tokenizing, Case Folding, & StopWords)

visit :

[1] https://medium.com/@ksnugroho/dasar-text-preprocessing-dengan-python-a4fa52608ffe

[2] https://devtrik.com/python/text-preprocessing-dengan-python-nltk/

TASK

1. Tugas dibuat perkelompok
2. Buat/pilih sebuah paragraf dalam bahasa Indonesia yang terdiri **minimal 15 kalimat** (*sesuai ketentuan*) lalu lakukan
   (1) Tokenizing,
   (2) Case Folding,
   (3) & StopWords
3. Buat laporan berisikan code, input, dan ouput di tiap tahapan
4. Dikumpulkan di google classroom

"You can dig out facts, but if you don't go deep enough, the facts won't tell you much."

(Jeff Catlin, CEO Lexalytics Inc)

Thank you

</ti>
teknik informatika

UNIVERSITAS ISLAM NEGERI
ALAUDDIN
MAKASSAR