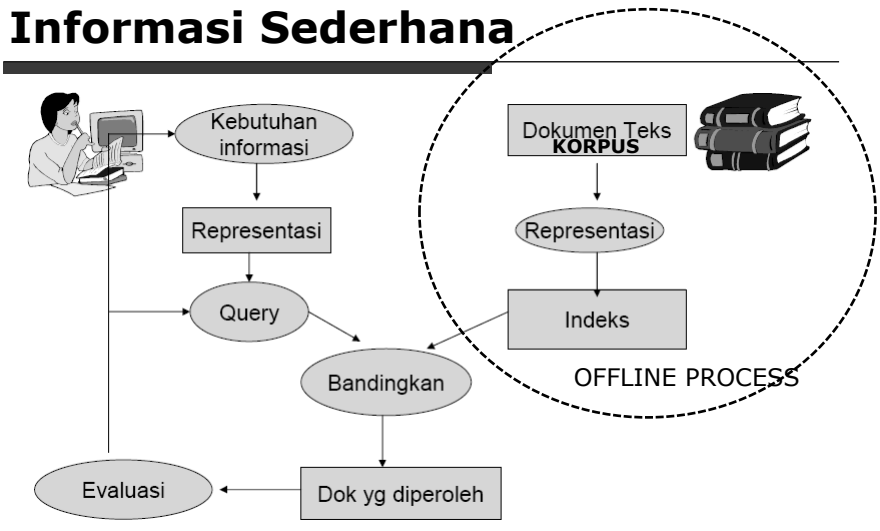
KOM341
Temu Kembali Informasi

KULIAH #2
- Pemrosesan Teks
- Java

Proses Perolehan Informasi Sederhana



JULIO ADISANTOSO - ILKOM IPB

Pengertian TEKS

- Teks ≈ Korpus ≈ Koleksi dokumen yang bisa dibaca oleh mesin
- Contoh:
  - Kumpulan artikel surat kabar yang diperoleh dari Internet
  - Kumpulan skripsi mahasiswa yang telah dikumpulkan secara digital oleh perpustakaan

Korpus

- Korpus adalah teks alami yang dipilih dengan cara tertentu.
- Masalah pada perancangan korpus
  - Ukuran
  - Jenis
  - Bahasa
- Media: teks, audio, video (multimedia)
- Isu pada korpus:
  - Tokenisasi pada korpus
  - Anotasi pada korpus

Contoh Korpus
Free text

Sekurangnya 17 ribu ayam ras milik peternak di wilayah kabupaten Kotawaringin Timur (Kotim) , Kalimantan Tengah mati dan kuat dugaan akibat terserang virus avian influenza (AI) atau yang lagi ramai disebut penyakit flu burung. Kasubdin Produksi Peternakan Dinas Pertanian Kotim Drh. Mawardi di Sampit, Selasa mengatakan sebanyak 17 ribu ekor ayam ras yang mati diduga terserang flu burung itu sejak Desember 2003.

Dari hasil diagnosa Balai Penyelidikan dan Pengujian Veteriner (BPPV) regional V Banjar Baru Kalimantan Selatan yang diterima Disnak Kotim, Senin (26/1) menyebutkan ayam yang mati terserang panyakit itu hanya ada dua kemungkinan yaitu terserang virus AI dan VVND atau tetelo. "Namun kasus kematian masal ayas ras di Kotim kemungkinan besar akibat akibat serangan virus avian influenza yang bila menular kepada manusia namanya menjadi flu burung," ucapnya.

Contoh Korpus
XML Format

```
<DOC>
<DOCNO>DOC01</DOCNO>
<TITLE>Flu Burung Menyerang Kalimantan Tengah</TITLE>
<AUTHOR>Ark, Ant</AUTHOR>
<DATE> 7 Februari 2003 </DATE>
<TEXT>
<P>Sekurangnya 17 ribu ayam ras milik peternak di wilayah kabupaten
Kotawaringin Timur (Kotim) , Kalimantan Tengah mati dan kuat dugaan akibat
terserang virus avian influenza (AI) atau yang lagi ramai disebut penyakit flu
burung.  Kasubdin Produksi Peternakan Dinas Pertanian Kotim Drh. Mawardi di
Sampit, Selasa mengatakan sebanyak 17 ribu ekor ayam ras yang mati diduga
terserang flu burung itu sejak Desember 2003.</p>

<P>Dari hasil diagnosa Balai Penyelidikan dan Pengujian Veteriner (BPPV) regional
V Banjar Baru Kalimantan Selatan yang diterima Disnak Kotim, Senin (26/1)
menyebutkan ayam yang mati terserang panyakit itu hanya ada dua kemungkinan
yaitu terserang virus AI dan VVND atau tetelo.  "Namun kasus kematian masal ayas
ras di Kotim kemungkinan besar akibat akibat serangan virus avian influenza yang
bila menular kepada manusia namanya menjadi flu burung," ucapnya.</P>
</TEXT>
</DOC>
```

Melacak Teks

- Operasi dasar dalam string matching
- Contoh:
  - Dapatkan semua baris yang dimulai dengan kata Flu.
  - Dapatkan semua baris yang dimulai dengan kata Huruf Besar.
  - Dapatkan semua baris yang memiliki kata terdiri dari huruf besar semua.
  - Hitung banyaknya kata Flu pada dokumen tersebut.
  - dsb.

Statistik Teks

- Jumlah Kata
  - Seberapa besar korpus yang ada (N)
- Jenis kata
  - Berapa jumlah kata yang unik?
  - Berapa besar perbendaharaan kata pada korpus?
- Token kata
  - Berapa jumlah kata pada korpus?
  - Berapa frekuensi dari setiap jenis kata?
  - Kata apa yang paling sering muncul pada korpus?

Prosedur Menghitung Frekuensi Kata

- Tokenisasi : mendapatkan kata
- Ubah menjadi huruf kecil
- Urutkan menurut abjad
- Hitung frekuensi kemunculan kata
- Urutkan menurut frekuensinya
- Hitung frekuensi dari frekuensi kemunculan kata

Fenomena Frekuensi Kata

- Sejumlah kata merupakan kata yang sangat umum (frekuensi sangat besar), misalnya "the", "of"
- Kebanyakan kata sangat jarang muncul (frekuensi sangat kecil).
- Setengah dari kata-kata pada korpus hanya muncul sekali.

Contoh

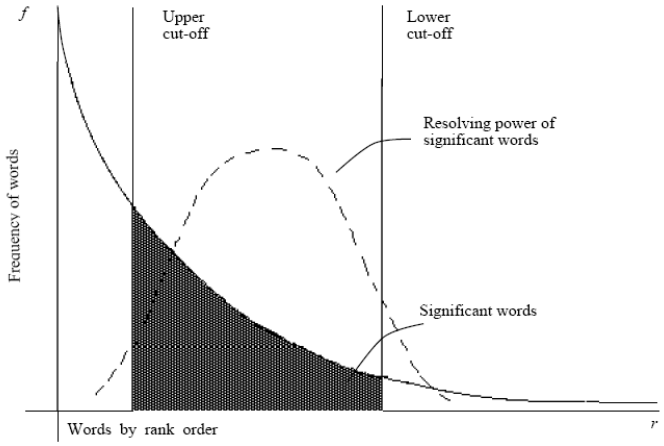
Kata	Frekuensi Kata (f)	Peringkat (r)	f * r
name	21	400	8400
comes	16	500	8000
group	13	600	7800
science	11	700	7700
family	10	800	8000
begin	9	900	8100
broke	4	2000	8000
seems	2	3000	6000
could	2	4000	8000

Hukum Zipf

- Menjelaskan adanya hubungan antara frekuensi dan urutan/rank (George Kingsley Zipf).
- Urutan/Rank:
  - hitung berapa kali kata muncul pada semua teks di dalam korpus (f).
  - urutkan sesuai dengan frekuensi kemunculan kata membentuk rank (r).
- Terdapat konstanta c sehingga f * r = c

Luhn's Ideas

kata-kata yang paling umum dan paling tidak umum adalah tidak signifikan untuk indexing



Stopwords

- STOPWORDS:
  - Terdapat kata yang merupakan bagian terbesar dari teks yang tidak perlu digunakan sebagai penciri dokumen.
  - Terdapat banyak ragam kata yang hanya muncul sedikit sekali di dalam suatu teks.
  - Contoh: to, in, form, yang, dan
- Kata-kata dengan frekuensi cukup (di bagian tengah) adalah yang paling baik digunakan sebagai penciri dokumen.

Tokenisasi

- Pengertian : suatu tahap pemrosesan di mana teks input dibagi menjadi unit-unit kecil yang disebut token, yang dapat berupa suatu kata, suatu angka, atau suatu tanda baca.
- Konsekuensinya:
  - Perlu mengenali unit secara otomatis
  - Apakah suatu kata itu?
  - Kalimat?
  - Paragraf?

Kata

- Karakter alfanumerik yang saling terhubung yang dipisahkan oleh whitespace.
- Whitespace: spasi, tab, newline
- Masalah:
  - B2B, amazon.com, Micro$oft
  - isn't, Jum'at
  - pro-aktif, out-of-date
  - tanda sambung pada akhir baris

Segmentasi kata

- Proses tokenisasi sederhana, tetapi tidak terlepas dari kesalahan.
- Contoh:
  - Kata majemuk: Jurusan Surabaya-Jakarta
  - Frase: tusuk jarum, keras kepala, sistem informasi
  - Nomor telpon (0251) 8356653 +62 251 8625584
- Menjadi topik dari ekstraksi informasi

Kalimat

- Satu atau lebih string kata yang diakhiri dengan suatu tanda berhenti sepenuhnya, tanda tanya atau tanda seru.
- Contoh:
  - Akhir dari baris.
  - Akhir dari suatu cerita!
  - Apakah kamu sudah punya pacar?
  - Ia sering mengunjungi friendster.com.
  - Dr. Iwan pergi ke Surabaya.
  - Dia mengatakan “Bohong!”.

Batas kalimat

- Hipotesakan bahwa batas kalimat sesudah . ? !
- Pindahkan batas kalimat sesudah tanda petik, bukan setelah titik.
  - "Jangan ambil buku itu. Buku itu milik pak Budi. ", kata ibu kepada Ani.
- Jangan gunakan titik jika:
  - Sebelumnya adalah singkatan yg umum yg biasanya bukan akhir kalimat, tp biasanya diikuti oleh nama dengan huruf besar: Prof.
  - Didahului oleh singkatan yang umum dan tidak diikuti oleh kata dengan huruf besar: Jr.
- Jangan kenali sebagai batas jika ada ! atau ? yang diikuti oleh huruf kecil

Pemrosesan Teks Otomatis

- Menghitung kata pada teks
- Mengurutkan kata
- Menghitung berbagai nilai statistik kata
  - Input : berkas teks (bisa berupa korpus)
  - Output : daftar kata beserta frekuensinya
- Pertanyaannya:
  - Bagaimana program komputernya?
  - Bahasa pemrograman apa yang digunakan?

Java

- Free for download → http://java.sun.com
- Unit terkecil program Java adalah Class yang terdiri dari methods (C:procedure) dan instance (C: data)
- Contoh:

```
public class Hello {
   public static void main(String[] args) {
      // menampilkan string ke layar
      System.out.println("Hello world!");
   }
}
```


Program Java

- Program Java harus disimpan dengan nama *.java
- Nama File seharusnya sama dengan nama class public nya
- Program yang berada pada satu folder dianggap sebagai satu package
- Berisi komentar secukupnya untuk memperjelas kode program

Standard input

- Menggunakan kelas BufferedReader yang berada di java.io

```
import java.io.*;
```

- Menyimpan input keyboard ke dalam buffer

```
BufferedReader dataIn = new
BufferedReader(new InputStreamReader(System.in));
```

- Menyimpan input ke dalam variabel sementara bertipe String

```
try {
   String temp = dataIn.readLine();
} catch(IOException e){
   System.out.println("Error input");
}
```

- Contoh (stdin.java) → menghitung rata-rata dari n bilangan riil.

Standard output

- Tanpa format

```
System.out.print(rataan);
System.out.println(rataan);
```

- Dengan format

```
System.out.format("%.2f \n", rataan);
```

Array dalam Java

- Sama dengan program dalam C
- Mendeklarasikan variabel array

```
int []usia; atau int usia[];
```

- Membuat objek array (dalam Java disebut sebagai instantiation)

```
int usia[];
usia = new int[100];
```

atau bisa juga ditulis sekaligus menjadi

```
int usia[] = new int[100];
```

- Dapat juga langsung didefinisikan seperti dalam C

```
boolean hasil[]={ true, false, true };
int[] nilai = {100, 90, 80, 75};
String hari[] = {"Senin","Selasa","Rabu"};
```


Mendefinisikan Class dalam Java

- Definisi class

```
<modifier> class <name> {
   <attributeDeclaration>*
   <constructorDeclaration>*
   <methodDeclaration>*
}
```

- Contoh:

```
public class Lingkaran {
   //area penulisan kode selanjutnya
}
```

Contoh class Lingkaran

```
public class Lingkaran {
   // Instance variables
   private double x;
   private double y;
   private double r;

   // Instance methods
   public void set(double x, double y, double r) {
      this.x=x; this.y=y; this.r=r; }
   public double luas() {
      double phi=3.14;
      return phi*r*r; }

   // main routine
   public static void main(String[] args) {
      // penulisan kode program utama
   }
}
```


Tokenisasi dengan Java

- Dapat menggunakan dua metode:
  - Class StringTokenizer
  - Method split
- Lihat file contoh