TEMU KEMBALI INFORMASI

JULIO ADISANTOSO
Departemen Ilmu Komputer IPB

Pertemuan 1
PENDAHULUAN

Identitas Mata Kuliah

Nama Mata Kuliah	:	Temu Kembali Informasi
Kode Mata Kuliah	:	KOM431
Koordinator	:	Julio Adisantoso (JAS)
Semester	:	Ganjil 2016/2017
Beban SKS	:	3(3-0)

Deskripsi Mata Kuliah

Matakuliah ini menjelaskan pengantar temu kembali informasi, dasar-dasar temu kembali informasi: pemodelan, evaluasi, query, operasi teks dan multimedia, indexing and searching. Topik dalam temu kembali informasi: relevance feedback, query expansion, text classification, text clustering, summarization, cross-language, question answering, web search, semantic web, semantic search.

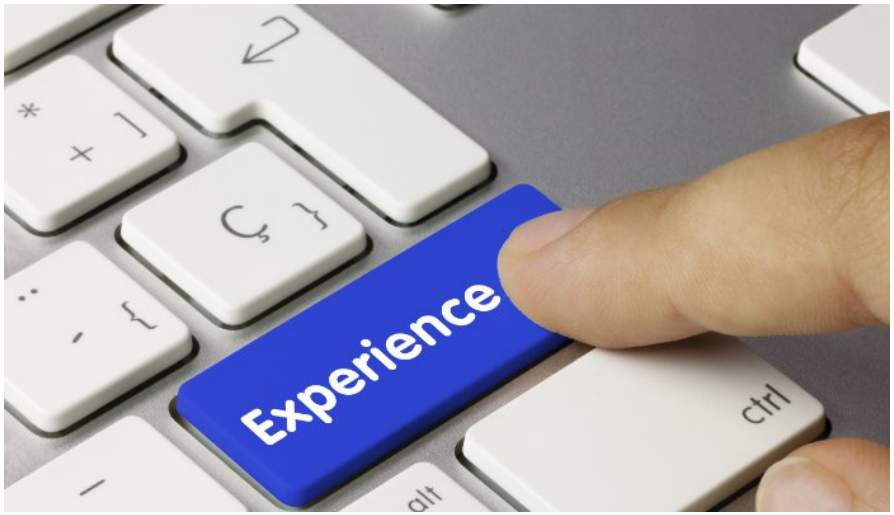
Penentuan Nilai Akhir

Nilai akhir (NA) adalah nilai kumulatif dari nilai ujian tengah semester (UTS), ujian akhir semester (UAS), tugas perorangan (TP), dan tugas kelompok atau proyek akhir (PA).

- UTS dan UAS dilakukan melalui ujian tertulis dengan bobot masing-masing 35%.
- Nilai TP adalah rata-rata dari semua tugas yang diberikan, dan diberi bobot 10%. Nilai PA terdiri dari nilai produk proyek (program komputer, laporan) dan presentasi. Bobot nilai PA adalah 20%.

Catatan: **Tidak ada ujian perbaikan**

What is this course about?



- Processing
- Indexing
- Retrieving
- . . . textual data

Fits in four lines, but much more complex and interesting than that.

Need for IR

- With the advance of WWW - more than 8 Billion documents indexed on Yahoo, Google
- Various needs for information:
  - Search for documents that fall in a given topic
  - Search for a specific information
  - Search an answer to a question
  - Search for information in a different language
  - . . .
  - Search for images
  - Search for music
  - Search for a (candidate) friend
  - . . .

Beberapa Definisi IR

- Salton (1989): ”Information-retrieval systems process files of records and requests for information, and identify and retrieve from the files certain records in response to the information requests. The retrieval of particular records depends on the similarity between the records and the queries, which in turn is measured by comparing the values of certain attributes to records and information requests.”
- Information retrieval mempelajari algoritme dan model untuk memperoleh informasi dari koleksi dokumen
- Information retrieval system : sistem untuk merepresentasikan, menyimpan, mengorganisasikan, dan memproses informasi (Beeza-Yates & Ribeiro-Neto)

IR versus Data Retrieval

IR

- berkaitan dengan natural language text → unstructured and semantically ambigous
- spesifikasi set of words untuk menentukan semantics dari information needed

Data Retrieval

- berkaitan dengan data → well defined structure and semantic
- spesifikasi query expression untuk menentukan constrain yang harus dipenuhi untuk obyek yang akan menjadi himpunan jawaban

IR Principal

- The indexing and retrieval of textual documents.
- Searching for pages on the World Wide Web is the most recent and perhaps most widely used IR application
- Concerned firstly with retrieving relevant documents to a query.
- Concerned secondly with retrieving from large sets of documents efficiently.

Retrieve semua dokumen yang relevan terhadap kueri pengguna dan seminimum mungkin retrieve dokumen yang tidak relevan

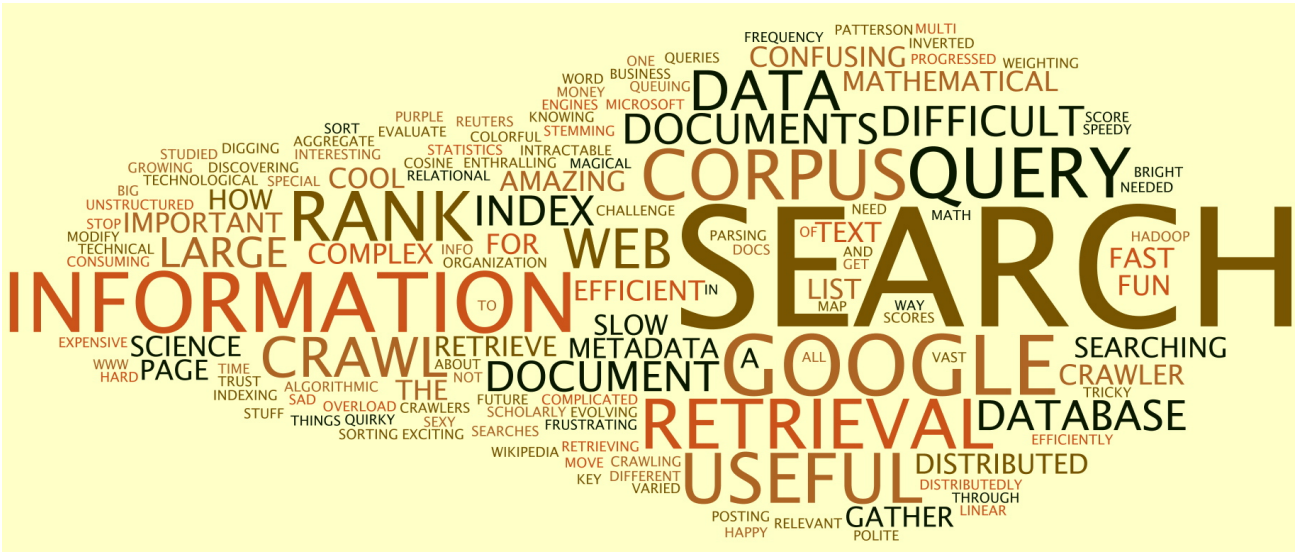
Typical IR Task

Given:

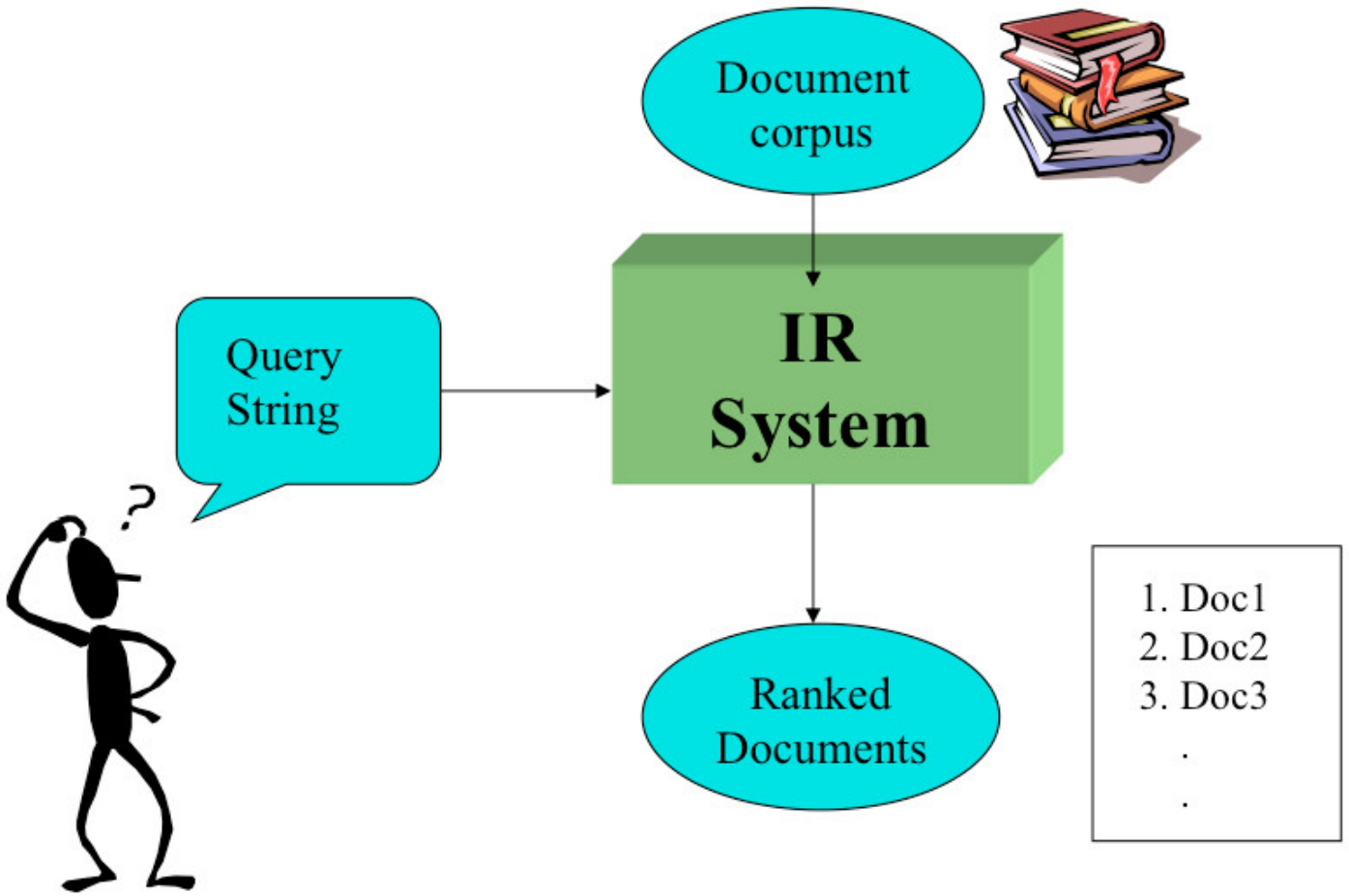
- A corpus of textual natural-language documents.
- A user query in the form of a textual string.

Find:

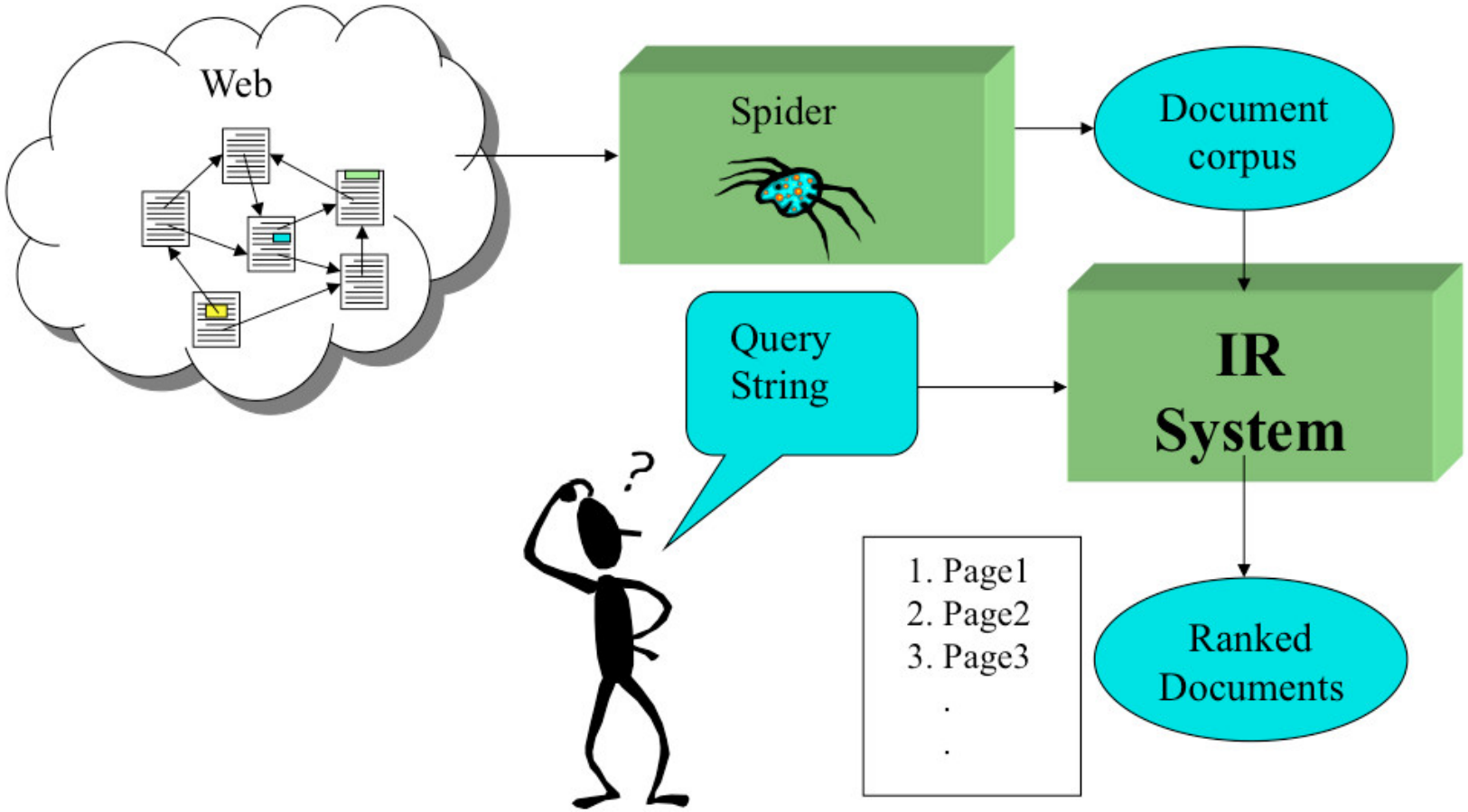
- A ranked set of documents that are relevant to the query.



IR System Architecture



Web Search System



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 sekumpulan teks/dokumen 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 Dokumen : 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.

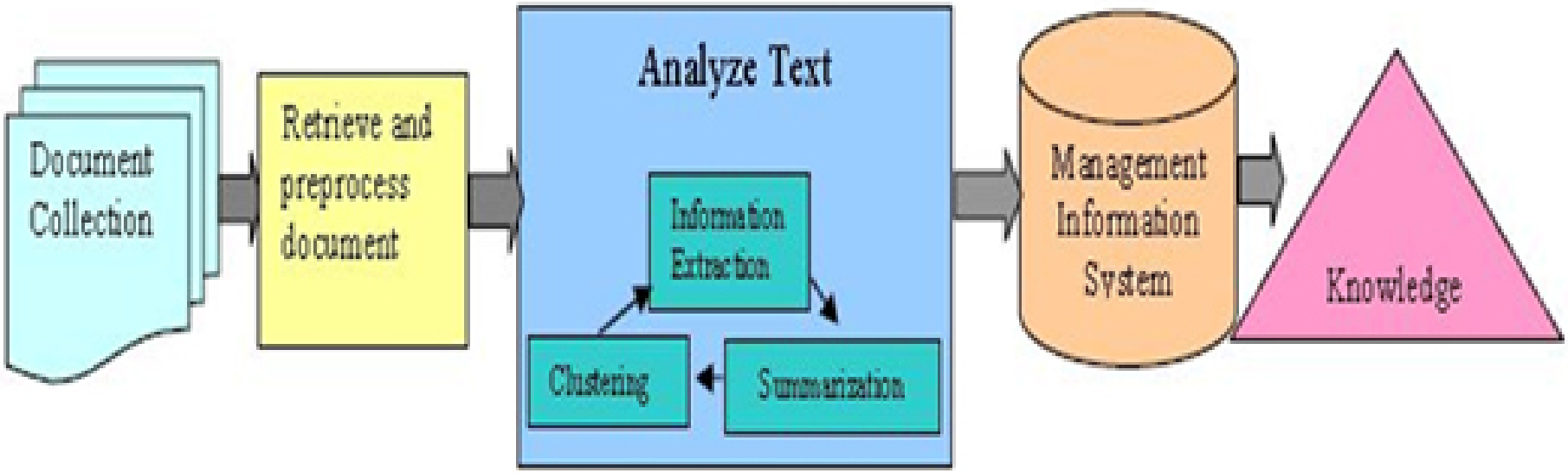
Contoh Dokumen : XML Format

<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>

Pemrosesan Teks

- Dalam IR, langkah awal yang umum dilakukan adalah pengolahan teks menjadi bentuk yang mudah untuk diproses sesuai tujuan tertentu (sering disebut pre-processing).
- Merupakan bagian dari text mining.



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?

Tokenisasi

- Pengertian : suatu tahap pemrosesan di mana teks input dibagi menjadi unit-unit kecil yang disebut token, yang dapat berupa suatu paragraf, kalimat, kata, angka, tanda baca, atau unit lainnya sesuai kebutuhan.
- Konsekuensinya:
  - Perlu mengenali unit secara otomatis
  - Apakah suatu kata itu? Kalimat? Paragraf?
- Pertanyaannya:
  - Bagaimana program komputernya?
  - Bahasa pemrograman apa yang digunakan?