Text Mining

Text Summarization

| Outline

1. Text Summarization
2. Tipe dan Output PTO
3. Tahapan Peringkasan Teks
4. Pendekatan Peringkasan Teks
5. Peringkasan TF-ISF

| #1 Text Summarization

- Suatu ringkasan adalah suatu teks yang dihasilkan dari satu atau lebih teks yang berisi bagian informasi yang signifikan dalam teks asal, dan yang tidak lebih dari setengah teks aslinya.
  - Hovy, E. H. Automated Text Summa- rization. In R. Mitkov (ed), The Ox- ford Handbook of Computational Linguistics, chapter 32, pages 583–598. Oxford University Press, 2005.
- Ringkasan Teks (Text Summarization) adalah suatu proses penyulingan sebagian besar informasi penting dari sumber (beberapa sumber) untuk menghasilkan suatu ringkasan bagi pemakai atau pekerjaan tertentu.
  - Mani, I., House, D., Klein, G., et al. The TIPSTER SUMMAC Text Summarization Evaluation. In Proceedings of EACL, 1999.

| #2 Peringkasan Teks Otmatis (PTO)

- Ketika proses peringkasan teks dilakukan oleh komputer secara otomatis, maka kita sebut sebagai Automatic Text Summarization (Peringkasan Teks Otomatis – PTO).
- Otomatisasi ringkasan dapat dikenakan terhadap satu dokumen (single document summarization) atau beberapa dokumen (multi-document summarization), satu bahasa (monolingual) atau beberapa bahasa (translingual/multilingual).

| #2 Tipe PTO

- Ringkasan yang umum (Generic Summary)
  - perwakilan dari teks asli yang mencoba untuk mempresentasikan semua feature penting dari sebuah teks asal.
  - mengikuti pendekatan bottom-up (Information Retrieval)
  - pemakai menginginkan segala informasi yang penting
- Ringkasan berpusat pada pemakai (query-driven)
  - peringkasan bersandar pada spesifikasi kebutuhan informasi pemakai, seperti topik atau query.
  - mengikuti pendekatan top-down (Information Extraction).

| #2 Output PTO

- Abstraksi (abstract)
  - abstraksi mengambil nilai-nilai informasi, tidak mesti satu kalimat utuh, mengambil informasi yang penting dan kemudian dirangkai menjadi suatu dokumen baru dengan kalimat penyampaian yang mungkin berbeda dengan kalimat dari dokumen sumber, sebagaimana manusia meringkas sesuatu.
- Ekstraksi (Extract)
  - Ringkasan yang dibangun adalah dokumen yang terdiri dari kalimat-kalimat yang dipilih dan dianggap penting dari dokumen sumber.

Galanis. Dimitrios, Lampouras. Gerasimos and Androutsopoulos. Ion, “Extractive Multi-Document Summarization with Integer Linear Programming and Support Vector Regression”. *COLING 2012,* pp.911-926, 2012.

| #2 Bentuk Output PTO

- Indicative
  - ringkasan yang dapat mengidentifikasikan topik yang terdapat pada teks sumber dan dapat memberikan ide ringkas tentang apa yang tertuang dalam teks sumber
- Informative
  - ringkasan yang dapat mengidentifikasikan informasi tertentu dari dokumen sumber.

| #2 Contoh Bentuk Output PTO

Designing for human-robot symbiosis

Presents the views on the development of intelligent interactive service robots. The authors have observed that a key research issue in service robotics is the integration of humans into the system. Discusses some of the technologies with particular emphasis on human-robot interaction, and system integration; describes human direct local autonomy (HuDL) in greater detail; and also discusses system integration and intelligent machine architecture (IMA). Gives an example implementation; discusses some issues in software development; and also presents the solution for integration, the IMA. Shows the mobile robot.

Saggion, H., Lapalme, G. (2002). Generating Indicative-Informative Summaries with SumUM. Computational Linguistic, Volume 28, Number 4.

| #2 Contoh Bentuk Output PTO (result)

- Indicative

Identified Topics: **HuDL - IMA - aid systems - architecture - holonic manufacturing system - human - human-robot interaction - intelligent interactive service robots - intelligent machine architecture - intelligent machine software - interaction - key issue - widely used interaction - novel software architecture - overall interaction - robot - second issue - service - service robots - software - system - Technologies**

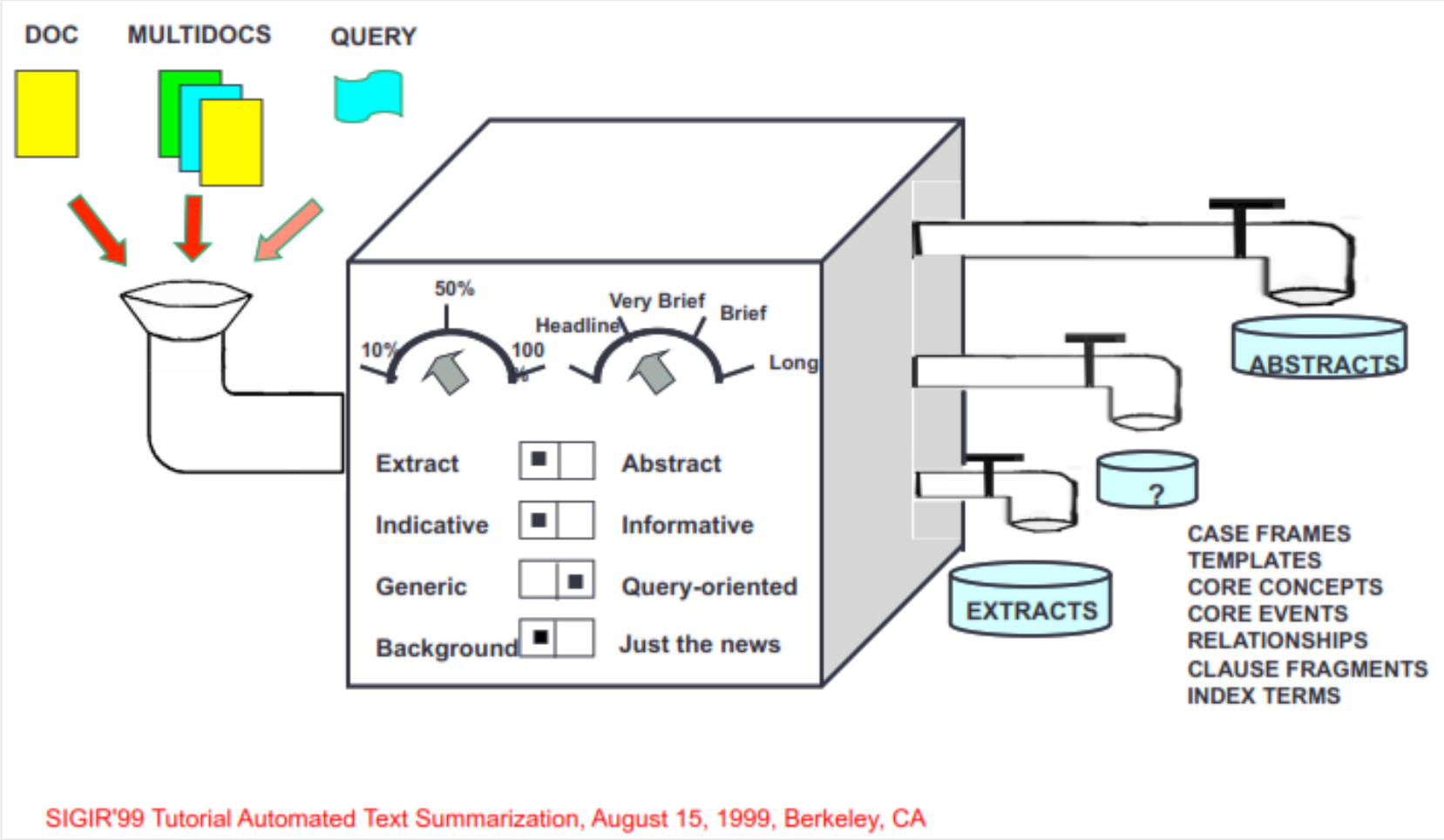
- Informative

Development of a service robot is an extremely challenging task.
In the IRL, we **are using HuDL** to guide the development of a cooperative service robot team.
IMA is a two-level software architecture for rapidly integrating these elements, for an intelligent machine such as a service robot.
A holonic manufacturing system is a manufacturing system having autonomous but cooperative elements called holons (Koestler, 1971).
Communication between the robot and the human is **a key concern for intelligent service robotics.**

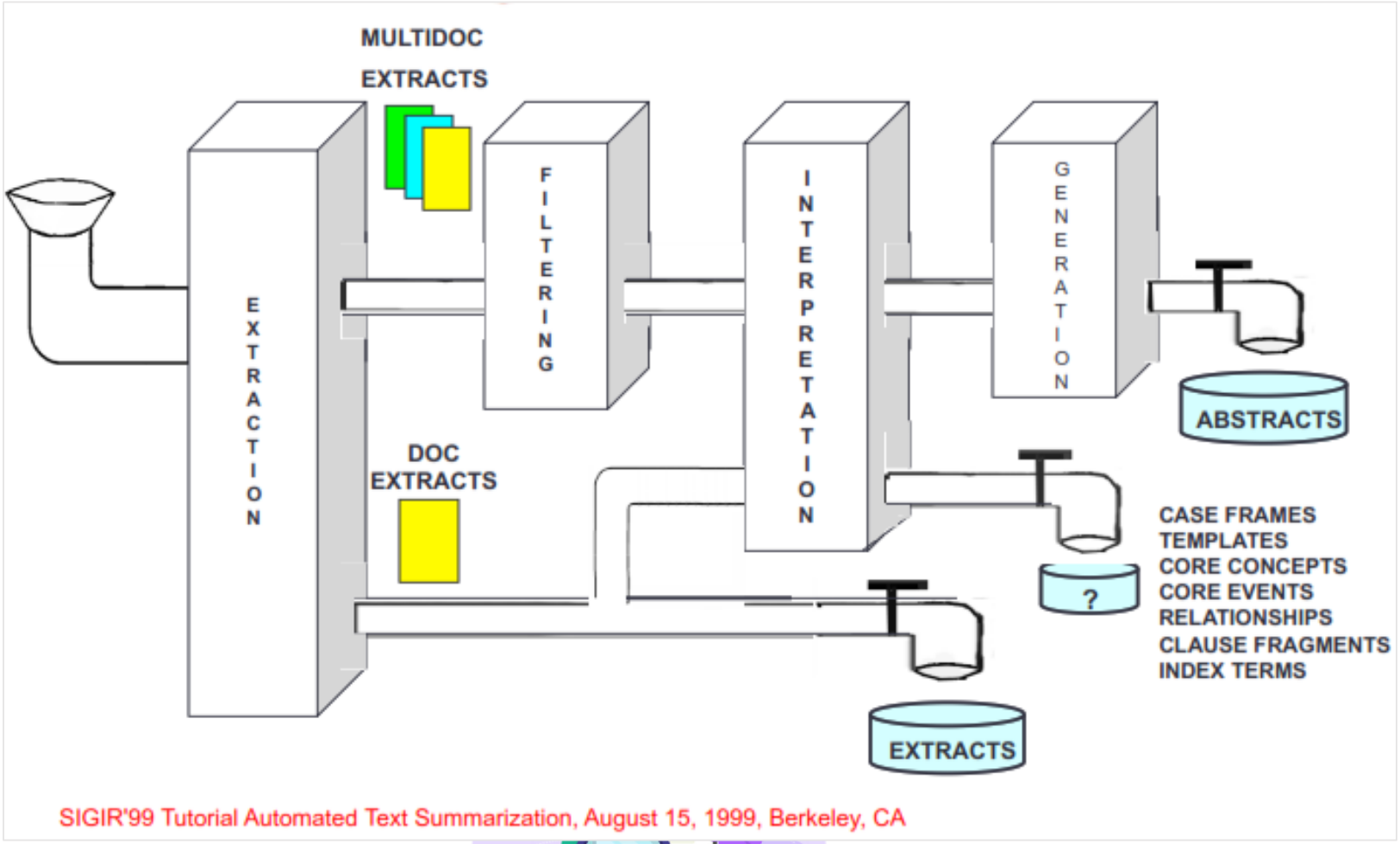
| #3 Tahapan PTO

- Interpretation
  - menafsir teks sumber untuk mendapatkan representasi dari teks
- Transformation
  - mengubah representasi teks menjadi sebuah representasi ringkasan teks
- Generation
  - ringkasan teks dari representasi ringkasan teks

| #3 Mesin Peringkasan Teks



| #3 Mesin Peringkasan Teks



| #4 Pendekatan Peringkasan Teks

- Menurut Many dan Maybury (1999), pendekatan peringkasan teks secara otomatis dapat dikelompokkan berdasar level pemrosesannya.
  - Surface Level
  - Entity Level
  - Discourse Level

| #4 Surface Level

- Pendekatan ini cenderung untuk mewakili informasi dengan mengambil fitur dangkal dan kemudian secara selektif menggabungkan mereka bersama-sama dalam suatu urutan untuk mendapatkan fungsi ciri khas yang dapat digunakan untuk mengekstrak informasi.
  - Thematic features
  - Location
  - Background
  - Cue words

| #4 Surface Level : Thematic

- Pendekatan Thematic bergantung pada statistik kemunculan kata
- sehingga sebuah kalimat dengan kemunculan kata tinggi dalam teks akan memiliki bobot tinggi dari pada lainnya.
- kalimat yang memiliki bobot tinggi diasumsikan adalah penting.
- perlu dilakukan tahapan penyaringan terhadap stop-word.
- perhitungan bobot TF*IDF sangat bermanfaat untuk menentukan keyword dalam teks.

| #4 Surface Level : Location

- Lokasi merujuk pada posisi dalam teks, paragraf, atau sembarang bagian dalam teks dimana diasumsikan posisi-posisi tersebut mengandung kalimat yang dimasukkan dalam ringkasan.
- Terdapat dua metode:
  - *lead-method*
    - kalimat penting muncul di awal dari teks (paragraf)
    - mengambil sebanyak n kalimat pertama
  - *title-based method*
    - kata-kata dalam judul dan header dianggap berhubungan secara positif dengan ringkasan.

| #4 Surface Level : Background & Cue Word

- Background mengasumsikan bahwa unit arti penting suatu teks ditentukan kemunculan kata dari judul/heading, bagian awal dari teks atau query pemakai.
- *Cue Word / Phrase*
  - suatu kalimat dinilai penting jika berisi frase-frase “bonus” tertentu.
  - contoh: “In this paper we show”, “In conclusion”
  - suatu kalimat dinilai kurang penting jika mengandung “stigma phrase”, seperti “hardly”, “impossible”.
  - metode yang diterapkan adalah menambahkan bobot pada kalimat jika berisi frase “bonus”, dan mengurangi bobot pada kalimat jika berisi “stigma phrase”.

| #4 Entity Level

- Pendekatan ini mencoba membangun suatu representasi teks, memodelkan entitas teks dan relasinya.
- Tujuannya adalah untuk membantu menentukan apa yang menonjol.
- Relasi antar entitas antara lain:
  - Similarity
  - Proximity
  - Co-ocurrence
  - Thesaural relatioships among words
  - Coreference
  - Logical relations
  - Syntatic relations
  - Meaning representation-based relations

| #4 Discourse Level

- Tujuan pada level ini adalah untuk memodelkan struktur global dari teks dan relasinya dalam rangka untuk mencapai tujuan ringkasan yang komunikatif.
- Informasi yang dapat digali:
  - Format Document
  - Threads of Topics
  - Rethorical structure of text
- Aliran topik dari suatu teks dicerminkan oleh pemakaian konstruksi vocabulary dan syntatical.

| #5 Peringkasan Berdasarkan Bobot Kalimat

- Setiap kata dalam kalimat dihitung bobotnya dengan TF- ISF (term frequency - inverse sentence frequency).

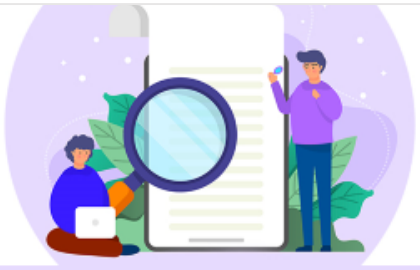
$$TF - ISF(w,s) = TF(w,s) * ISF(w)$$

$$ISF(w) = \log\left(\frac{|S|}{SF(w)}\right)$$

- Selanjutnya, untuk setiap kalimat, s, dihitung rata-rata bobot TS-ISF:

$$Avg - TF - ISF(s) = \frac{\sum_{i=1}^{W(s)} TF - ISF(w_i, s)}{W(s)}$$

W(s) = jumlah kata dalam kalimat s



| #5 Peringkasan Berdasarkan Bobot Kalimat

- Setelah Avg-TF-ISF(s) dihitung, maka berikutnya semua kalimat diurutkan berdasar bobot rata-rata secara descending.
- Pilih kalimat yang memiliki bobot paling tinggi, Max-Avg-TF-ISF(s), sebagai wakil teks.
- Pemakai juga dapat menentukan prosentase kalimat tertinggi yang akan diambil dengan menghitung nilai masukan prosentase (Θ) :

$$\theta_{TF-ISF} = \omega * Max - Avg - TF - ISF$$

- Sistem akan mengembalikan semua kalimat, s, yang memiliki

$$Avg - TF - ISF \geq \theta_{TF-ISF}$$

| #5 Metode Lain

- Latent Semantic Analysis (LSA)
  - http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.87.5199
- Word Cluster
  - http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.6072
- Shortest Path
  - http://dr-hato.se/research/shortpath.pdf
- Ranking dan Relasi Kalimat
  - http://aclweb.org/anthology-new/W/W09/W09-1608.pdf

Thank you

</ti>
FST | UINAM
teknik informatika

UNIVERSITAS ISLAM NEGERI
ALAUDDIN
MAKASSAR