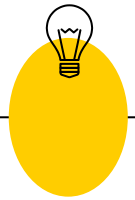
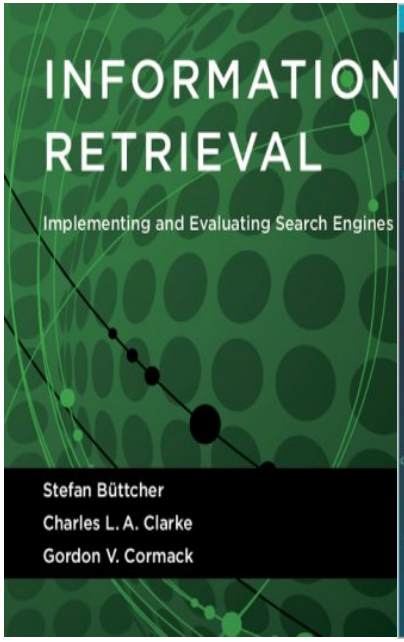
Sistem Temu Kembali Informasi

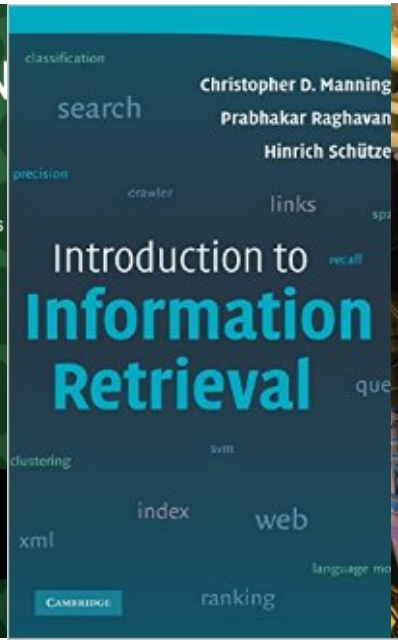
"Peringkasan Dokumen Teks"

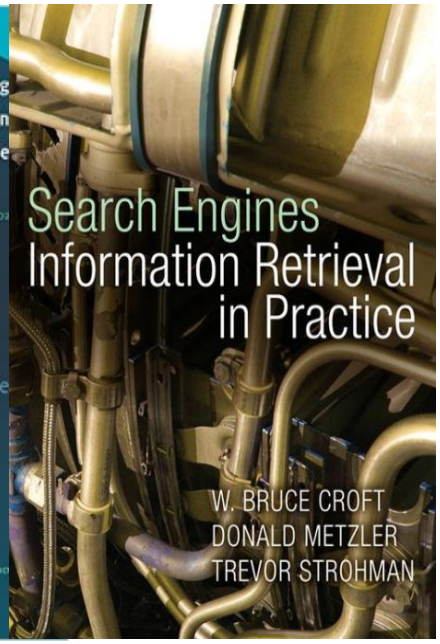


Tim Dosen STKI

Buku Penunjang & Literatur







Course Outline

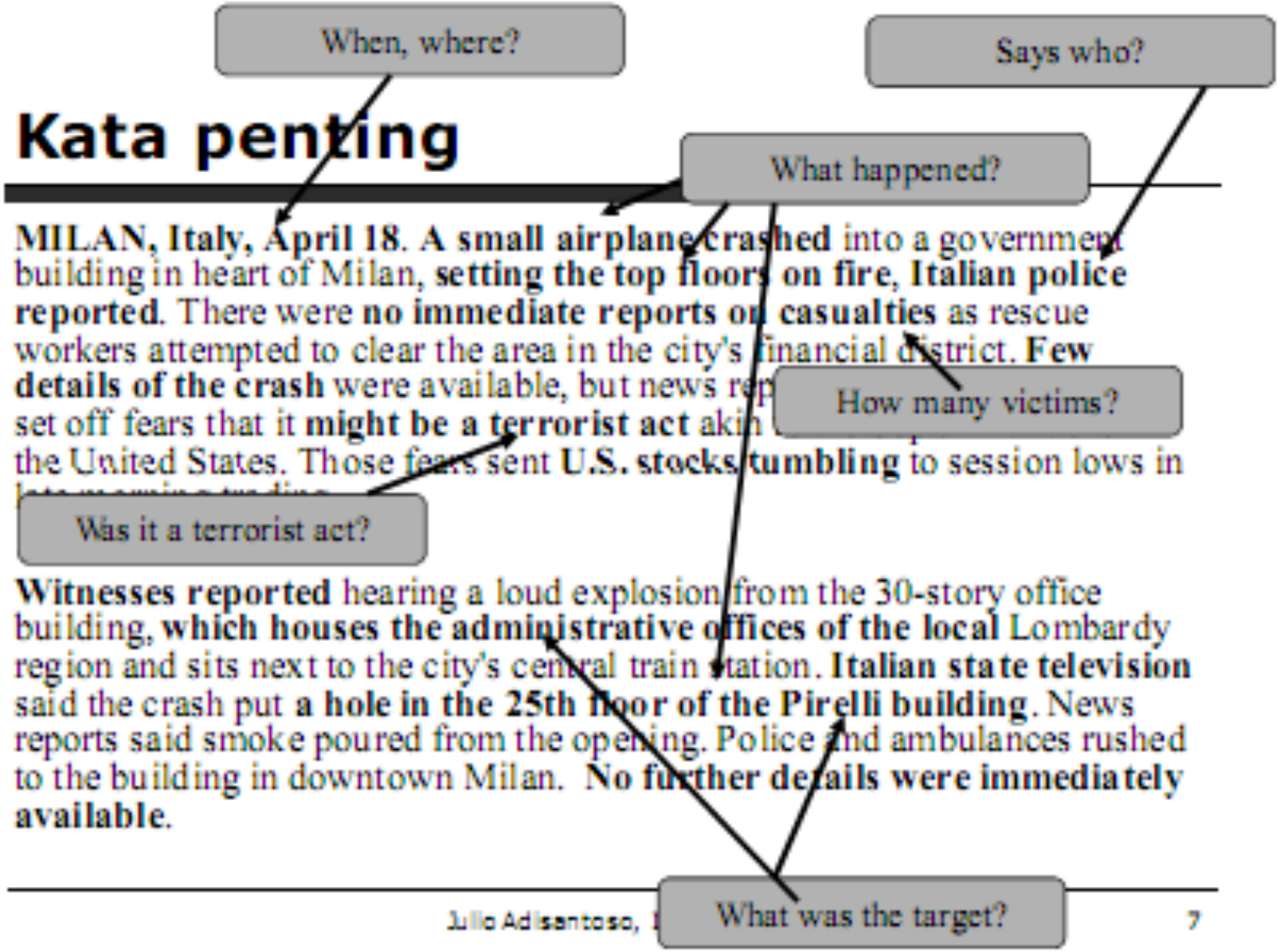
- Peringkasan Dokumen Text
- Feature Based
- Latent Semantic Analysis

Contoh Artikel

- MILAN, Italy, April 18. A small airplane crashed into a government building in heart of Milan, setting the top floors on fire, Italian police reported. There were no immediate reports on casualties as rescue workers attempted to clear the area in the city's financial district. Few details of the crash were available, but news reports about it immediately set off fears that it might be a terrorist act akin to the Sept 11 attacks in the United States Those fears sent akin to the Sept. 11 attacks in the United States. Those fears sent U.S. stocks tumbling to session lows in late morning trading.
- Witnesses reported hearing a loud explosion from the 30-story office building, which houses the administrative offices of the local Lombardy region and sits next to the city's central train station. Italian state television said the crash put a hole in the 25th floor of the Pirelli building. News reports said smoke poured from the opening. Police and ambulances rushed to the building in downtown Milan. No further details were immediately available.

Konten Artikel (Kata Penting)

- MILAN, Italy, April 18. **A small airplane crashed** into a government building in heart of Milan, **setting the top floors on fire**, **Italian police reported**. There were **no immediate reports on casualties** as rescue workers attempted to clear the area in the city's financial district. **Few details of the crash** were available, but news reports about it immediately set off fears that it **might be a terrorist act** akin to the Sept. 11 attacks in the United States. Those fears sent **U.S. stocks tumbling** to session lows in the United States. Those fears sent U.S. stocks tumbling to session lows in late morning trading.
- **Witnesses reported** hearing a loud explosion from the 30-story office building, **which houses the administrative offices of the local** Lombardy region and sits next to the city's central train station. **Italian state television** said the crash put **a hole in the 25th floor of the Pirelli building**. News reports said smoke poured from the opening. Police and ambulances rushed to the building in downtown Milan. **No further details were immediately available.**


When, where?
Says who?
Kata penting
What happened?
MILAN, Italy, April 18. A small airplane crashed into a government building in heart of Milan, setting the top floors on fire, Italian police reported. There were no immediate reports on casualties as rescue workers attempted to clear the area in the city's financial district. Few details of the crash were available, but news rep
How many victims?
set off fears that it might be a terrorist act aki
the United States. Those fears sent U.S. stocks tumbling to session lows in
Was it a terrorist act?
Witnesses reported hearing a loud explosion from the 30-story office building, which houses the administrative offices of the local Lombardy region and sits next to the city's central train station. Italian state television said the crash put a hole in the 25th floor of the Pirelli building. News reports said smoke poured from the opening. Police and ambulances rushed to the building in downtown Milan. No further details were immediately available.
Julio Adisantoso,
What was the target?
7

Peringkasan Dokumen Text

Permasalahan : **Information overload**

Goals : **Meringkas Konten Dokumen**

Jenis Ringkasan

- Tujuan
  - **Indicative**, Informative
- Bentuk
  - **Ekstrak** (paragraf/kalimat/frase)
  - Abstrak : suatu ringkasan yg padat dari topic permasalahan dari suatu dokumen
- Konteks
  - Query specific, query independent
- Dimensi
  - Single Document, Multi document

Jenis ringkasan

- Indikatif vs. informatif
  - digunakan untuk kategorisasi secara cepat vs. pemrosesan isi.
- Ekstrak vs. abstrak
  - daftar fragmen teks vs. menyimpulkan kembali isi secara koheren.
- Query-independen vs. query-spesifik
  - mengikuti pandangan penulis vs merefleksikan minat dari user
- Background vs. just-the-news
  - asumsikan jika pengetahuan pembaca sebelumnya tidak banyak vs. sangat mengikuti perkembangan.
- Single-dokumen vs. multi-dokumen
  - berdasarkan pada satu teks vs. penggabungan beberapa teks.

Hasil Ringkasan

- Headlines
- Outlines
- minutes (notulen)
- Biographies
- sound bites
- movie summaries
- chronologies, etc.

Ekstrak VS Abstrak

- Ringkasan Teks
  - Proses penyaringan informasi yang paling penting dari suatu sumber (atau beberapa sumber) untuk menghasilkan suatu versi yang ringkas untuk user.
- Extract vs. Abstrak
  - Suatu extract adalah ringkasan yang isi seluruhnya disalin dari input.
  - Suatu abstrak adalah ringkasan yang paling sedikit ada isinya yang tidak ada pada input, mis. kategorisasi topik, menyarikan kembali isi, dsb.

Feature Based

1. Title Feature
2. Sentence Length
3. Term Weight
4. Sentence Position
5. Sentence to Sentence Similarity
6. Proper Noun
7. Thematic Word
8. Numerical Data

Sumber: Suanmali et. al. “Fuzzy Logic Based Method for Improving Text Summarization”

Title Feature

- kata yang terdapat pada kalimat **dokumen** yang merupakan bagian dari kata dalam **judul.**

$$S\_F1(S) = \frac{No.Title\ word\ in\ S}{No.Word\ in\ Title}$$

Sentence Length

- **Rasio** dari **jumlah kata dalam kalimat** dengan **jumlah kata yang terdapat pada kalimat terpanjang** pada suatu dokumen. (untuk menghilangkan misal ***datelines*** dan ***nama penulis***).

$$S\_F2(S) = \frac{No.Word\ occurring\ in\ S}{No.Word\ occurring\ in\ longest\ sentence}$$

Term Weight

- Menghitung **frekuensi munculnya sebuah term** pada dokumen yang biasa digunakan untuk menentukan penting tidaknya **posisi kalimat** pada sebuah dokumen

$$w_i = tf_i x isf_i = tf_i x log \frac{N}{n_i}$$

$$S_{F3(s)} = \frac{\sum_{i=1}^{k} W_i(S)}{Max\left(\sum_{i=1}^{k} W_i(S_i^N)\right)}$$

Sentence Position

- Apakah **letak suatu kalimat** ada pada **akhir** atau **awal** suatu paragraf dalam dokumen.

$$S\_F4(S) = 5/5 \text{ for } 1^{st},\ 4/5 \text{ for } 2^{nd},\ 3/5 \text{ for } 3^{rd},$$
$$2/5 \text{ for } 4^{th},\ 1/5 \text{ for } 5^{th},$$
$$0/5 \text{ for other sentences}$$

Sentence to Sentence Similarity

- **Kesamaan antar kalimat**, dimisalkan kalimat *s*, pengukuran kesamaan antara **kalimat *s*** dengan kalimat lainnya dengan menghitung **rasio** dari ringkasan kesamaan kalimat pada kalimat *s* tersebut dengan **maksimum ringkasan** jumlah dari **keseluruhan** kesamaan kalimat pada dokumen.

$$Score\ (S_i) = \frac{Sum\ of\ Sentemce\ Similarity\ in\ S_i}{Max(Sum\ of\ Sentence\ Similarity)}$$

Proper Noun

- **Proper noun** memiliki pengertian sebagai kata benda yang ditulis menggunakan huruf **kapital** di **awal** kata.

$$S\_F6(S) = \frac{No.\ Proper\ nouns\ in\ S}{Sentence\ \ Length\ (S)}$$

Thematic Word

- **Jumlah kata tematik** yang ada dalam kalimat. (dari seluruh koleksi dokumen kata yang sering muncul dan **jumlah kata di tentukan sendiri**).

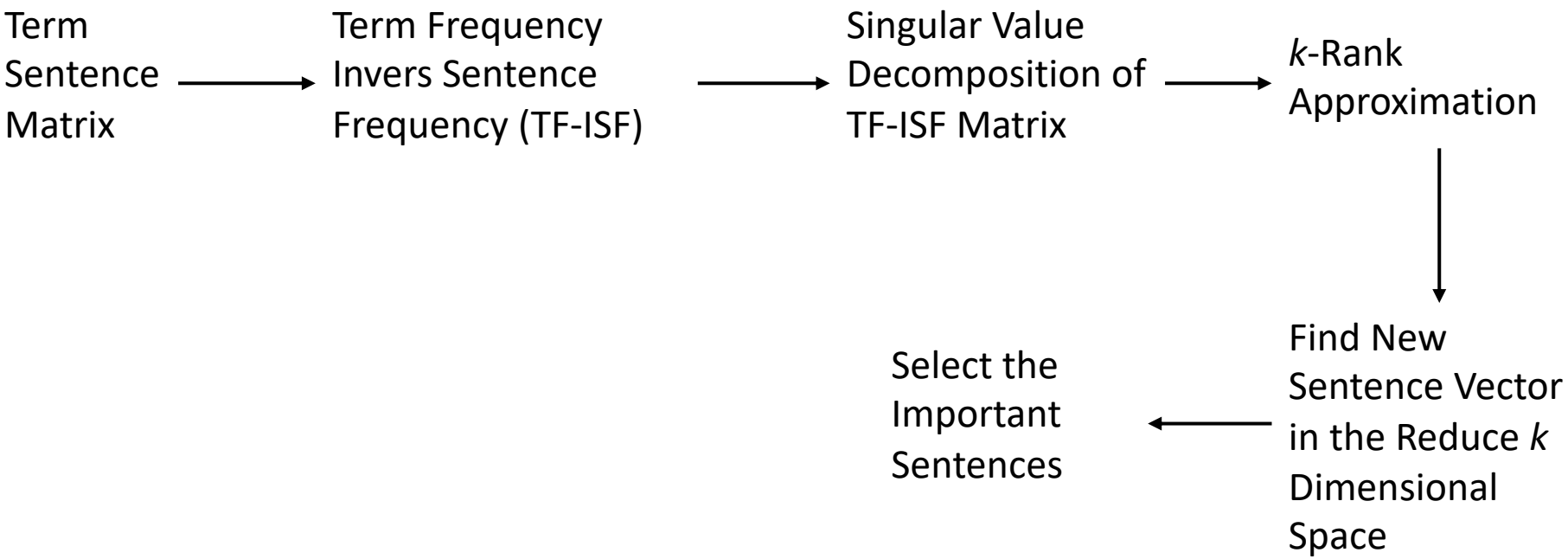
$$S\_F7(S) = \frac{No.\ Thematic\ word\ in\ S}{Max(No.\ Thematic\ word)}$$

Numerical Data

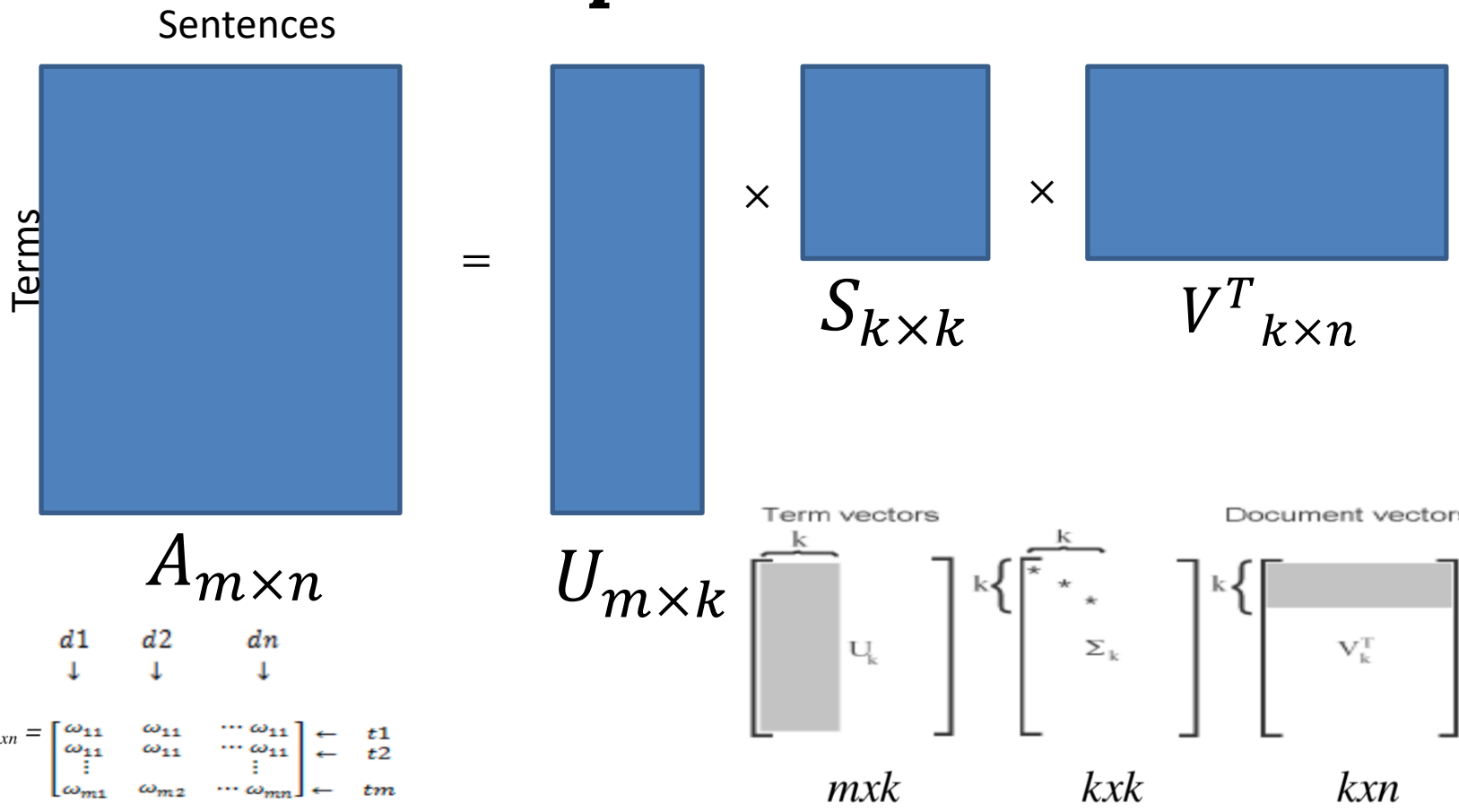
- Jumlah **data numerik** yang ada dalam kalimat, kedudukan kalimat yang mengandung data numeric adalah **penting** karena dimungkinkan akan **masuk** kedalam isi ringkasan dokumen.

$$S\_F8(S) = \frac{No.\ Numerical\ data\ in\ S}{Sentence\ Length\ (S)}$$

Latent Semantic Analysis



Singular Value Decomposition



Kesimpulan

- Peringkasan Dokumen Teks dapat digunakan sebagai “Feature Reduction” pada Tahap Pemrosesan Dokumen
- Meringkas Dokumen dapat digunakan untuk meningkatkan kinerja Proses Text Clustering baik Akurasi maupun waktu komputasi.
- Banyak teknik yang digunakan untuk peringkas dokumen teks

Kuis (Latihan Soal)

- Cari minimal 2 buah (paper / jurnal) yang membahas mengenai peringkasan dokumen teks, lakukan review dan buatkan kesimpulan dari paper / jurnal tersebut mengenai fungsi dari implementasi dokumen teks yang dilakukan.

Thanks!

Any questions ?