Text Mining

Ali Ridho Barakbah

Definisi

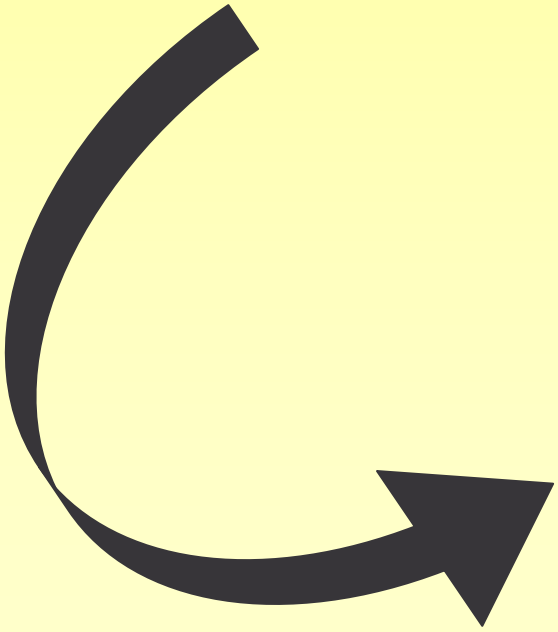
- Menambang data yang berupa teks
- Sumber data biasanya didapatkan dari dokumen
- Tujuannya adalah mencari kata-kata yang dapat mewakili apa yang ada di dalam dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen

Tahapan

- Tokenizing
- Filtering
- Stemming
- Tagging
- Analyzing

Tokenizing

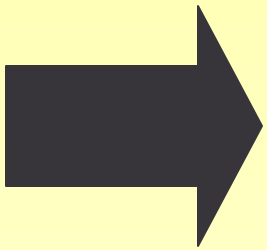
This lecture is talking about
how to mine data



this
lecture
is
talking
about
how
to
mine
data

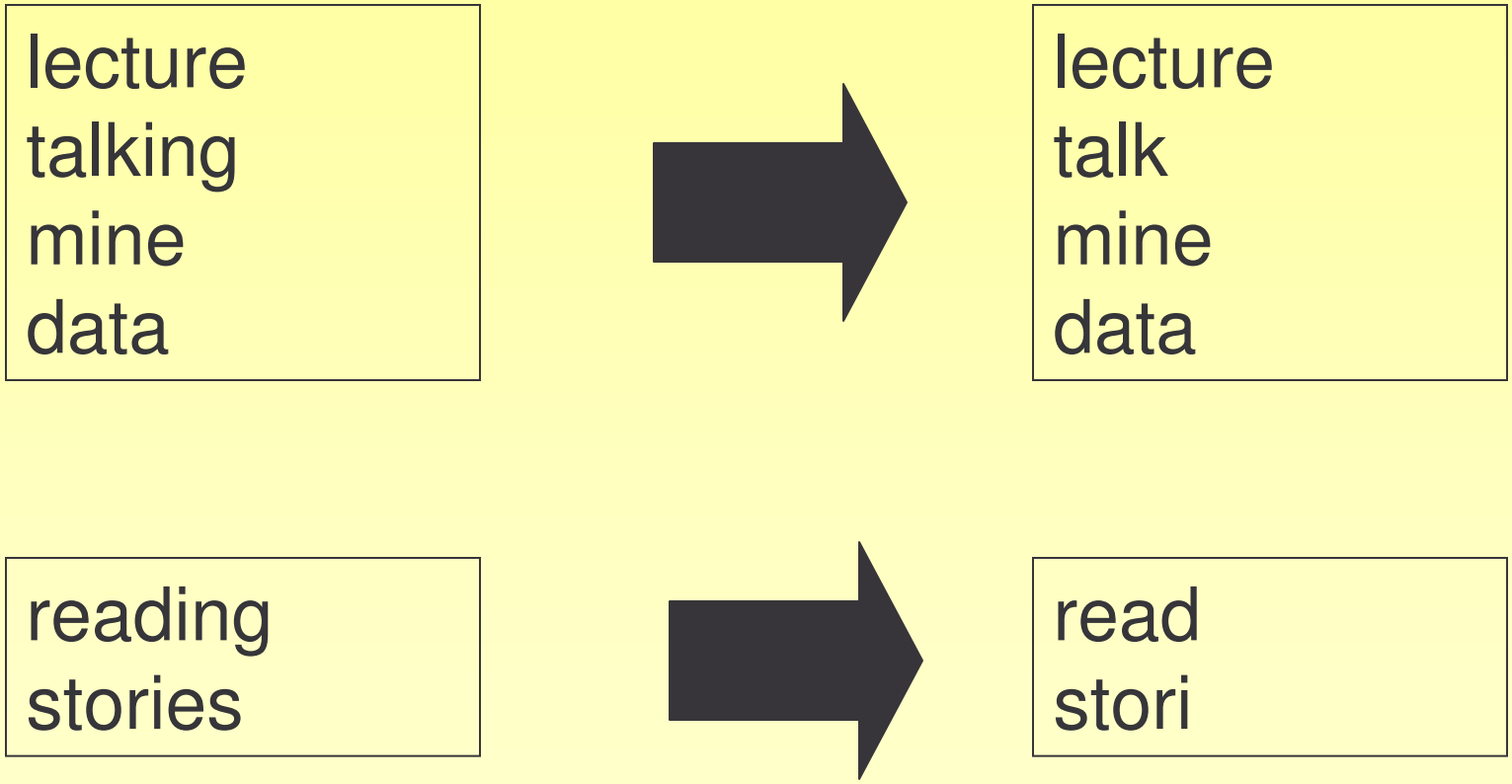
Filtering

this
lecture
is
talking
about
how
to
mine
data

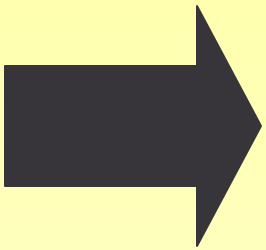


lecture
talking
mine
data

Stemming


lecture
talking
mine
data
lecture
talk
mine
data
reading
stories
read
stori

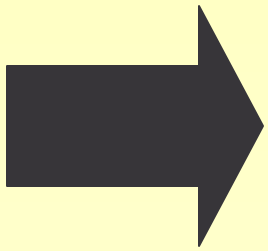
Tagging

thougt was stori		think be story

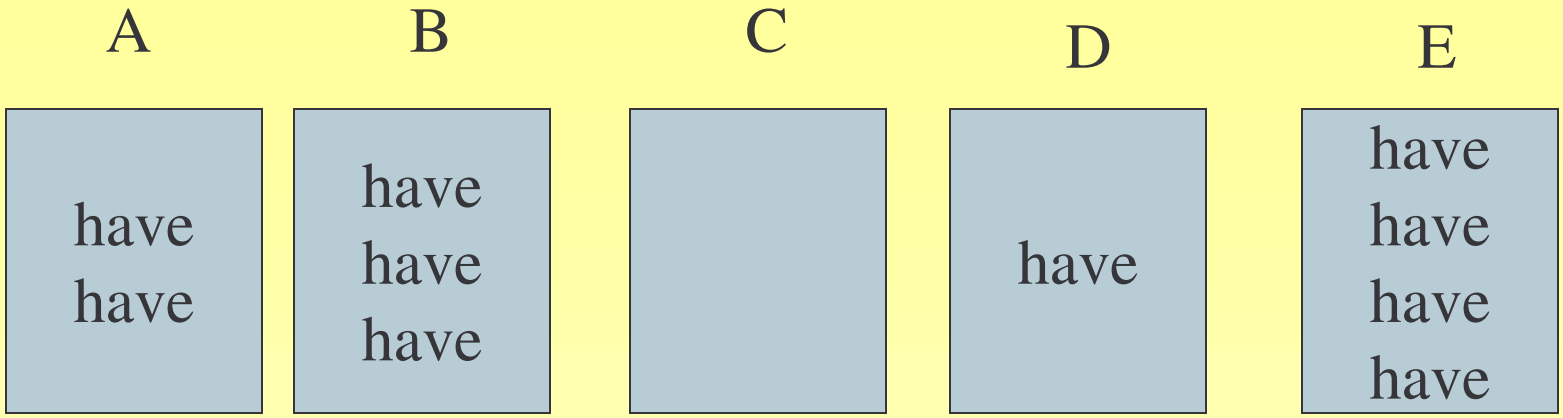
Analyzing

- Mencari seberapa jauh keterhubungan antar kata-kata antar dokumen
- Term Frequency-Inversed Documen Frequency (TF-IDF) → Algoritma yang paling sederhana yang biasanya dipakai untuk scoring

lecture
talk
mine
data



Lecture → 0.8
Talk → 0.34
Mine → 0.7
Data → 0.45



$$TFIDF_{d,t} = FREQ_{d,t}(1 + \log \frac{N}{DFREQ_t})$$

$$\text{TFIDF}_{have,B} = 3 \text{ x } (1 + \log(5 / 4))$$