

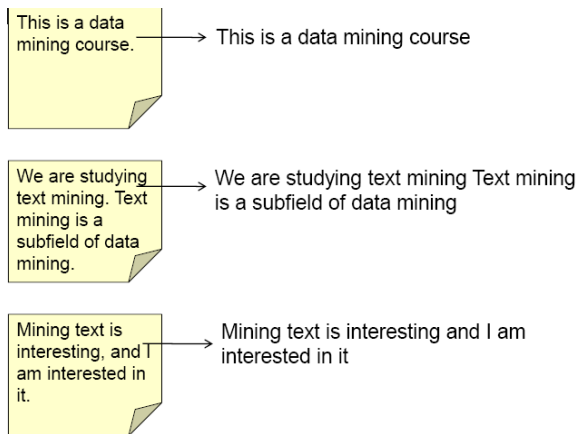
Ilustrasi *Preprocessing* & *Searching* Dalam *Text Mining*

Berikut ini adalah proses keseluruhan dari Text Mining, mencakup *Pre-Processing* dan Perhitungan (*Searching*) Kemiripan antara Query dengan Daftar Dokumen.

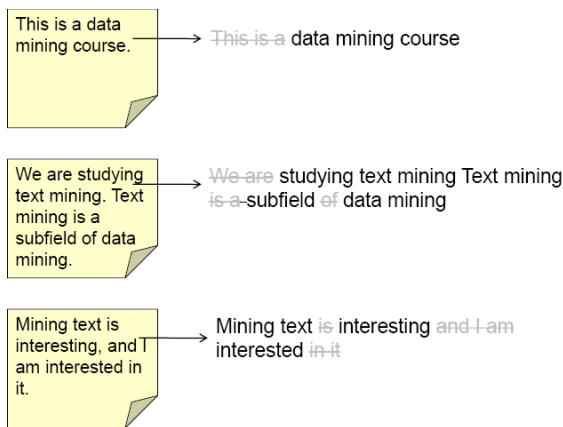
A. Preprocessing Terhadap Daftar Dokumen

1. Langkah 1: Mengekstrak Teks

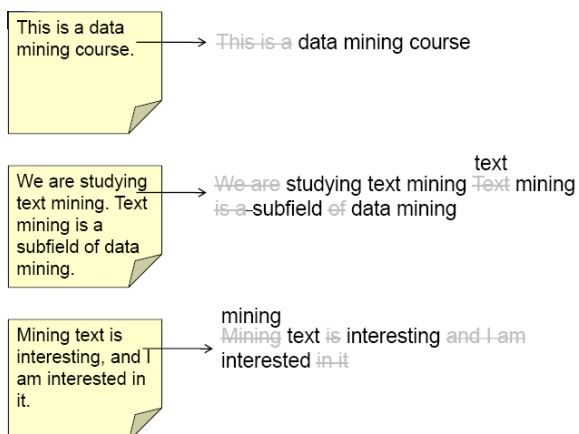
Misalnya terdapat 3 Dokumen seperti di bawah ini:



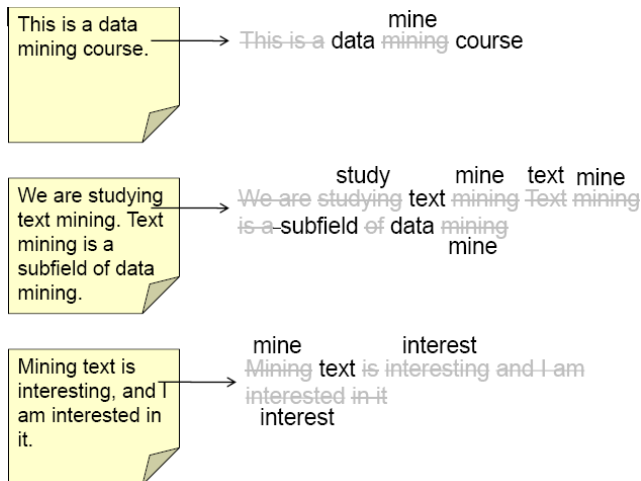
2. Langkah 2: Menghilangkan Stop Words



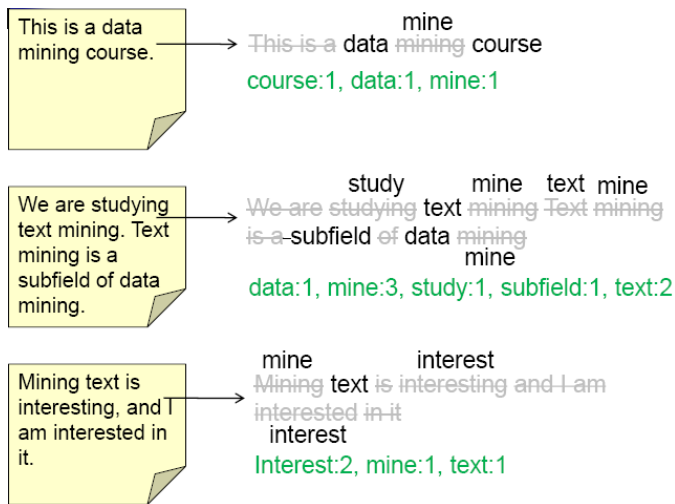
3. Langkah 3: Ubah semua kata ke huruf kecil



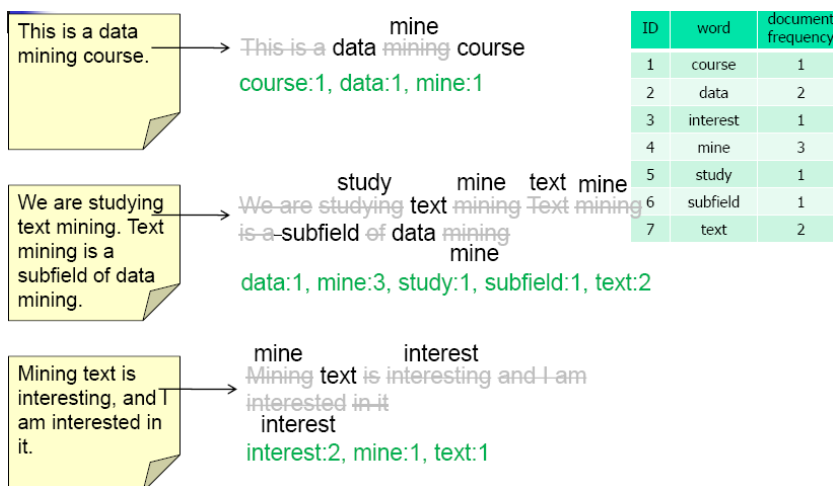
4. Langkah 4: Stemming



5. Langkah 5: Menghitung Frekuensi Kata dari setiap Dokumen (TF)



6. Langkah 6: Membuat File Index



7. Langkah 7: Membuat Model Ruang Vektor

This is a data mining course. → This is a data ^{mine} mining course
 course:1, data:1, mine:1
 (1, 1, 0, 1, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining. → We are studying ^{study} text ^{mine} mining ^{text} Text ^{mine} mining
 is a subfield of data ^{mine} mining
 data:1, mine:3, study:1, subfield:1, text:2
 (0, 1, 0, 3, 1, 1, 2)

Mining text is interesting, and I am interested in it. → Mining ^{mine} text ^{interest} is interesting and I am interested in it
 interest
 interest:2, mine:1, text:1
 (0, 0, 2, 1, 0, 0, 1)

ID	word	document frequency
1	course	1
2	data	2
3	interest	1
4	mine	3
5	study	1
6	subfield	1
7	text	2

8. Langkah 8: Menghitung Inverse Document Frequency (IDF)

This is a data mining course. → (1, 1, 0, 1, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining. → (0, 1, 0, 3, 1, 1, 2)

Mining text is interesting, and I am interested in it. → (0, 0, 2, 1, 0, 0, 1)

$$IDF(word) = \log \frac{\text{total documents}}{\text{document frequency}}$$

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

9. Langkah 9: Menghitung Bobot dari Setiap Kata (TF*IDF)

This is a data mining course. → (1, 1, 0, 1, 0, 0, 0)
 (0.477, 0.176, 0, 0, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining. → (0, 1, 0, 3, 1, 1, 2)
 (0, 0.176, 0, 0, 0.477, 0.477, 0.352)

Mining text is interesting, and I am interested in it. → (0, 0, 2, 1, 0, 0, 1)
 (0, 0, 0.954, 0, 0, 0, 0.176)

$$w(word_i) = TF(word_i) \times IDF(word_i)$$

$TF(word_i)$ = number of times $word_i$ appears in the document

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

10. Langkah 10: Normalkan Semua Dokumen ke Panjang Unit

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

This is a data mining course. → (1, 1, 0, 1, 0, 0, 0)
 (0.938, 0.346, 0, 0, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining. → (0, 1, 0, 3, 1, 1, 2)
 (0, 0.225 0, 0, 0.611, 0.611, 0.450)

Mining text is interesting, and I am interested in it. → (0, 0, 2, 1, 0, 0, 1)
 (0, 0, 0.983, 0, 0, 0, 0.181)

$$w(\text{word}_i) = \frac{w(\text{word}_i)}{\sqrt{w^2(\text{word}_1) + w^2(\text{word}_2) + \dots + w^2(\text{word}_n)}}$$

Contoh Perhitungan Normalisasi:

This is a data mining course. → (1, 1, 0, 1, 0, 0, 0)
 (0.477, 0.176, 0, 0, 0, 0, 0)

$$w(\text{course}) = \frac{0.477}{\sqrt{0.477^2 + 0.176^2 + 0 + 0 + 0 + 0 + 0}} = 0.938$$

(0.938, 0.346, 0, 0, 0, 0, 0)

B. Penanganan Query

Bagaimana Query ditangani? Hampir sama dengan *preprocessing* dokumen (bukan query), kemudian hitung kemiripan antara query dengan dokumen yang telah dipreprocess juga. Berikut ini adalah apa yang harus dilakukan jika terdapat query “**interested in interesting data and text**”:

Query Awal : (**interested in interesting data and text**)

- Langkah 1: Hilangkan semua stop word: (**interested interesting data text**)
- Langkah 2: Stemming: (**interest interest data text**)
- Langkah 3: Hilangkan duplikasi: (**interest data text**)
- Langkah 4: Bangun suatu model ruang vektor: (**0, 1, 1, 0, 0, 0, 1**)
- Langkah 5: Hitung bobot dari setiap kata: (**0, 0, 0.477, 0, 0, 0, 0.176**)
- Langkah 6: Normalkan model ruang vektor: (**0, 0, 0.938, 0, 0, 0, 0.346**)

Ingat tabel index dan bobot dari 3 dokumen yang telah dipreprocess?

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

7. Hitung kemiripan antara Query dan Daftar Dokumen menggunakan metode Cosine Similarity.

Q: (0, 0, 0.938, 0, 0, 0, 0.346)

Document 1: (0.938, 0.346, 0, 0, 0, 0, 0)

Document 2: (0, 0.225, 0, 0, 0.611, 0.611, 0.450)

Document 3: (0, 0, 0.983, 0, 0, 0, 0.181)

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

$$\text{cosine}(P, Q) = \frac{\sum p_i \cdot q_i}{\sqrt{\sum p_i^2 \times \sum q_i^2}}$$

$$\text{cosine}(D1, Q) = 0$$

$$\text{cosine}(D2, Q) = \frac{0.346 \times 0.450}{\sqrt{(0.938^2 + 0.346^2) \times (0.225^2 + 0.611^2 + 0.611^2 + 0.450^2)}} = 0.156$$

$$\text{cosine}(D3, Q) = \frac{0.938 \times 0.983 + 0.346 \times 0.181}{\sqrt{(0.938^2 + 0.346^2) \times (0.983^2 + 0.181^2)}} = 0.985$$

Kesimpulan: **Mengembalikan Dokumen #3**

C. Soal Latihan

Diberikan suatu query "W4 W5" dan koleksi 3 dokumen berikut:

- Dokumen 1: <W1 W2 W3 W4 W5 >
- Dokumen 2: <W6 W7 W4 W5>
- Dokumen 3: <W8 W3 W9 W4 W10>

Gunakan model ruang vektor (VSM), skema pembobotan TF/IDF, dan ukuran kemiripan vektor Cosine untuk mendapatkan dokumen yang paling relevan terhadap query tersebut!

Sumber awal: http://itee.uq.edu.au/%7Einfo4203/Lecture/Lesson07_Text_Mining_2011.pdf