Text Mining

Data Reduction & IR

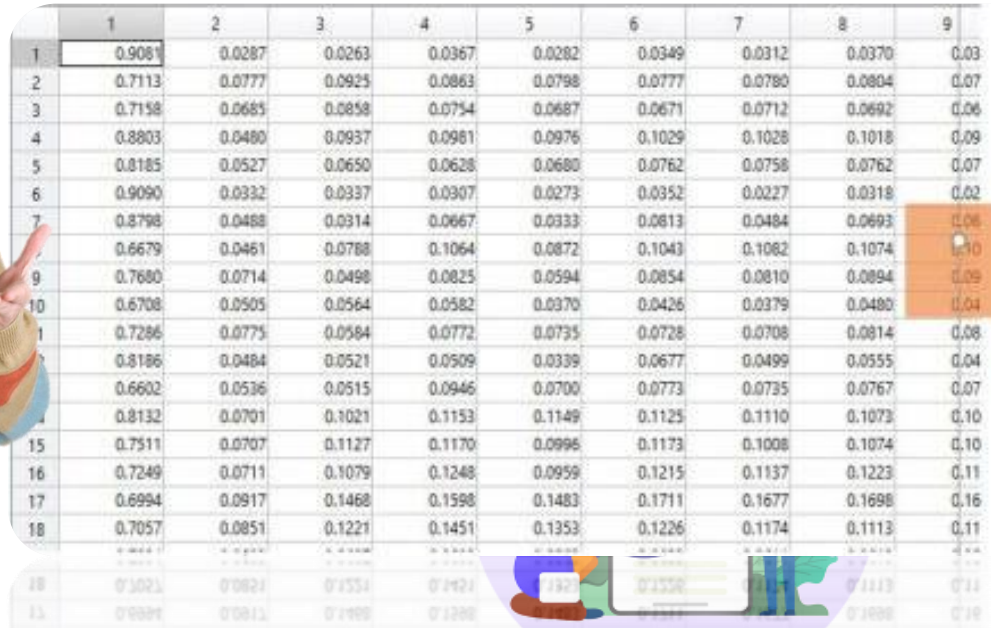
| Outline

1. Data Reduction
2. Feature Selection (Information Gain)
3. Information Retrieval Overview
4. Doc Similarity (Cosine Similarity)

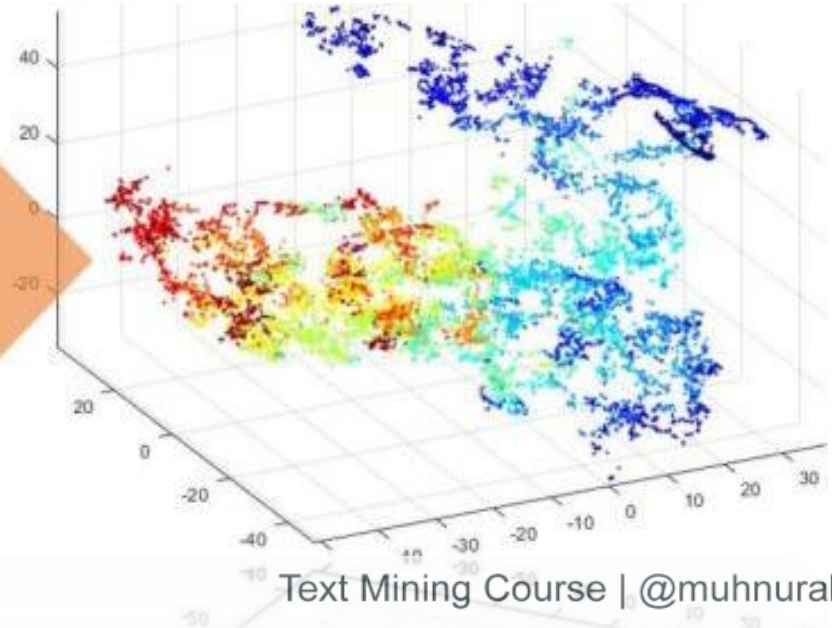
| #1 Data Reduction

- **Reduksi data** adalah sebuah proses analisis untuk memilih, memusatkan perhatian (focus), menyederhanakan, mengabstraksi serta mentransformasikan data.
- Mereduksi data berarti membuat rangkuman, **memilih** hal-hal pokok, memfokuskan pada hal-hal penting, mencari tema dan pola, serta **membuang** yang dianggap tidak perlu.

Kenapa data text
harus **direduksi**?



	1	2	3	4	5	6	7	8	9
1	0.9081	0.0287	0.0263	0.0367	0.0282	0.0349	0.0312	0.0370	0.03
2	0.7113	0.0777	0.0925	0.0863	0.0798	0.0777	0.0780	0.0804	0.07
3	0.7158	0.0685	0.0858	0.0754	0.0687	0.0671	0.0712	0.0692	0.06
4	0.8803	0.0480	0.0937	0.0981	0.0976	0.1029	0.1028	0.1018	0.09
5	0.8185	0.0527	0.0650	0.0628	0.0680	0.0762	0.0758	0.0762	0.07
6	0.9090	0.0332	0.0337	0.0307	0.0273	0.0352	0.0227	0.0318	0.02
7	0.8798	0.0488	0.0314	0.0667	0.0333	0.0813	0.0484	0.0693	0.06
8	0.6679	0.0461	0.0788	0.1064	0.0872	0.1043	0.1082	0.1074	0.10
9	0.7680	0.0714	0.0498	0.0825	0.0594	0.0854	0.0810	0.0894	0.09
10	0.6708	0.0505	0.0564	0.0582	0.0370	0.0426	0.0379	0.0480	0.04
11	0.7286	0.0775	0.0584	0.0772	0.0735	0.0728	0.0708	0.0814	0.08
12	0.8186	0.0484	0.0521	0.0509	0.0339	0.0677	0.0499	0.0555	0.04
13	0.6602	0.0536	0.0515	0.0946	0.0700	0.0773	0.0735	0.0767	0.07
14	0.8132	0.0701	0.1021	0.1153	0.1149	0.1125	0.1110	0.1073	0.10
15	0.7511	0.0707	0.1127	0.1170	0.0996	0.1173	0.1008	0.1074	0.10
16	0.7249	0.0711	0.1079	0.1248	0.0959	0.1215	0.1137	0.1223	0.11
17	0.6994	0.0917	0.1468	0.1598	0.1483	0.1711	0.1677	0.1698	0.16
18	0.7057	0.0851	0.1221	0.1451	0.1353	0.1226	0.1174	0.1113	0.11



| #1 Data Reduction (cont)

Tujuan reduksi data:

- **Waktu komputasi**, data yang lebih sederhana dapat mereduksi waktu untuk proses data mining
- **Penyajian/presentasi**, kesederhanaan representasi menjadi model yang lebih mudah dimengerti.
- ***Keakuratan prediksi/deskriptif**, mengukur seberapa baik data dapat disimpulkan dan digeneralisasi ke dalam suatu model

**relatif*

Reduksi data berupa:

- Delete kolom (cth : seleksi fitur)
- Delete baris (cth : delete duplikat data)
- Pengurangan nilai kolom (cth : binning)

High dimensional data→		**columns** ($\mathcal{J}$)
		$1 \cdots j \cdots d$
rows ($\mathcal{I}$)	$\mathbf{x}_1$	$x_{11} \cdots x_{1j} \cdots x_{1d}$
	$\vdots$	$\vdots$
	$\mathbf{x}_i$	$x_{i1} \cdots x_{ij} \cdots x_{id}$
	$\vdots$	$\vdots$
	$\mathbf{x}_n$	$x_{n1} \cdots x_{nj} \cdots x_{nd}$

| #2 Feature Selection

- **Feature Selection** merupakan teknik reduksi fitur/dimensi yang digunakan untuk memperkecil matriks data dengan memperhatikan informasi kata penting yang perlu diproses. Fitur yang dimaksud di sini adalah kata hasil preprocessing dari sebuah dokumen.
- **Information Gain** merupakan salah satu teknik seleksi fitur yang digunakan untuk memilih fitur terbaik yaitu dengan merangking kata-kata yang dianggap penting/ berpengaruh terhadap kelas prediksi.

Source : Sari, Yunita Arum & Puspaningrum, Eva Yulia, 2013.

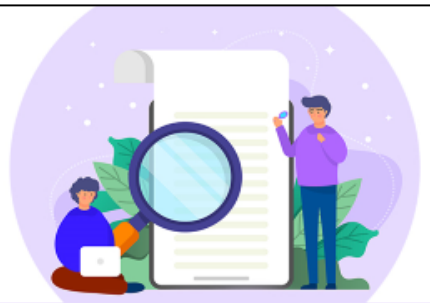
| #2 Feature Selection (Simulasi Information Gain)

- **Entropi**, metode information gain menggunakan konsep entropi, entropi digunakan untuk mengukur “seberapa informatifnya” atau “seberapa pentingnya” sebuah node.
  - Entropi(S) = 0, jika semua contoh pada S berada dalam kelas yang sama.
  - Entropi(S) = 1, jika jumlah contoh positif dan jumlah contoh negative dalam S adalah sama.
  - 0 < Entropi(S) < 1, jika jumlah contoh positif dan negatif dalam S tidak sama.

$$Entropi\ (S) = \sum_{j=1}^{k} -p_j \log_2 p_j$$

Dimana:

- S adalah himpunan (dataset) kasus
- k adalah banyaknya partisi S
- p_j adalah probabilitas yang di dapat dari Sum(Ya) dibagi Total Kasus.

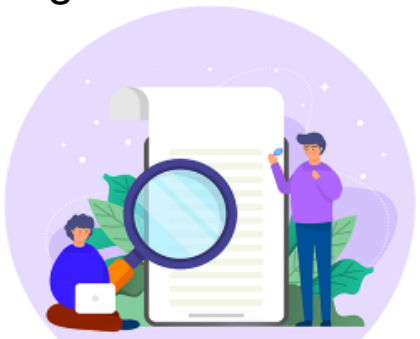


| #2 Feature Selection (Simulasi Information Gain)

- Setelah menemukan nilai entropi maka selanjutnya pemilihan fitur dilakukan dengan nilai information gain terbesar.

$$Gain\ (A) = Entropi\ (S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Entropi(S_i)$$

Dimana:
S = ruang (data) sample yang digunakan untuk training.
A = atribut.
$|S_i|$ = jumlah sample untuk nilai V.
$|S|$ = jumlah seluruh sample data.
Entropi(S_i) = entropy untuk sample-sample yang memiliki nilai *i*



| #2 Feature Selection (Simulasi Information Gain)

Contoh kasus:

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

1. Hitung Entropi Total (*S*)

Total Kasus	Sum(Yes)	Sum(No)
14	9	5

Entropi (S) = (-(9/14) * log2(9/14)) + (-(5/14) * log2(5/14))

= **0.9402859586706309**

| #2 Feature Selection (Simulasi Information Gain)

2. Hitung Entropi untuk tiap value dalam suatu kolom

cth : Entropi (**Rain**) pada kolom **Outlook**

Jumlah Kasus	Sum(Yes)	Sum(No)
5	3	2

Entropi (Rain) = (-(3/5) * log2(3/5)) + (-(2/5) * log2(2/5))

= **0.970950594**

3. Setelah menghitung semua Entropi pada tiap value maka selanjutnya hitung nilai information gain fitur/kolomnya.

cth : Information Gain (Outlook)

$$Gain\ (A) = Entropi\ (S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Entropi(S_i)$$

IG *(Outlook)* = (0.940285959) – **(** ((4/14)*0) + ((5/14)*0.970950594) + ((5/14)*0.970950594) **)**

= **0.24674982**

		JML KASUS	NO (S1)	YES (S2)	ENTROPY
TOTAL		14	5	9	0.940285959
OUTLOOK					
	OVERCAST	4	0	4	0
	RAINY	5	2	3	0.970950594
	SUNNY	5	3	2	0.970950594

| #2 Feature Selection (Simulasi Information Gain)

4. Hitung IG semua fitur atau kolom dan merangking dari nilai terendah hingga tertinggi

		JML KASUS	NO (S1)	YES (S2)	ENTROPY	INFORMATION GAIN
TOTAL		14	5	9	0.940285959	
OUTLOOK						0.24674982
	OVERCAST	4	0	4	0	
	RAINY	5	2	3	0.970950594	
	SUNNY	5	3	2	0.970950594	
TEMP						0.029222566
	COOL	4	1	3	0.811278124	
	HOT	4	2	2	1	
	MILD	6	2	4	0.918295834	
HUMIDITY						0.151835501
	HIGH	7	4	3	0.985228136	
	NORMAL	7	1	6	0.591672779	
WIND						0.04812703
	WEAK	8	2	6	0.811278124	
	STRONG	6	3	3	1	

Download file excelnya di sini

| #3 IR Overview

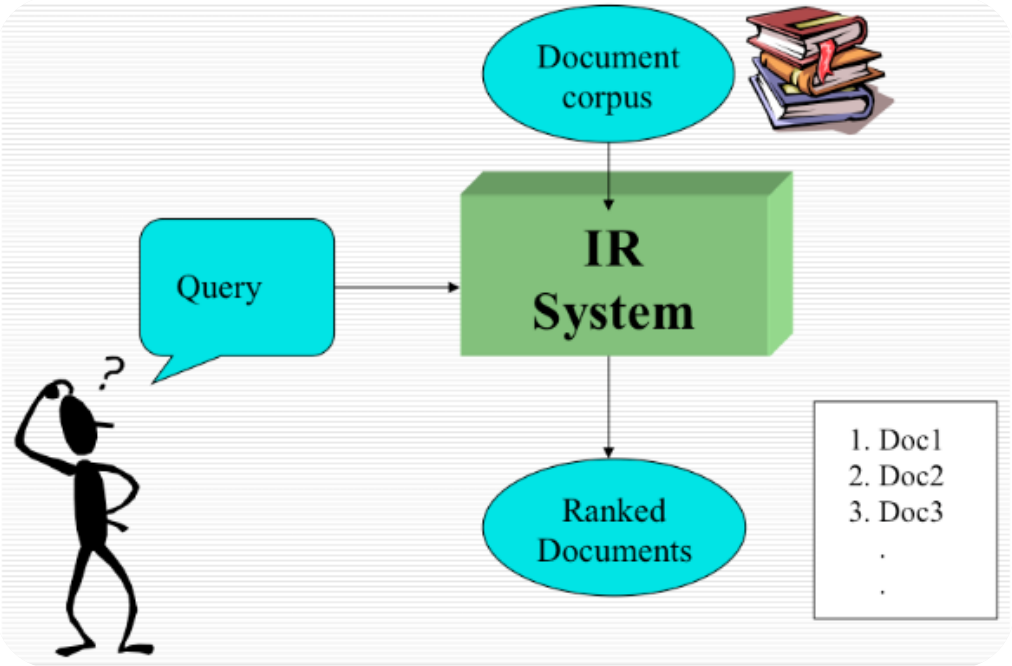
Definition

- Manning *et al* (2007): Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfy an information need from within large collections (usually stored on computers).
- 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.
- Beeza-Yates & Ribeiro-Neto: Information retrieval system adalah sistem untuk merepresentasikan, menyimpan, mengorganisasikan, dan memproses informasi.

| #3 IR Overview (cont)

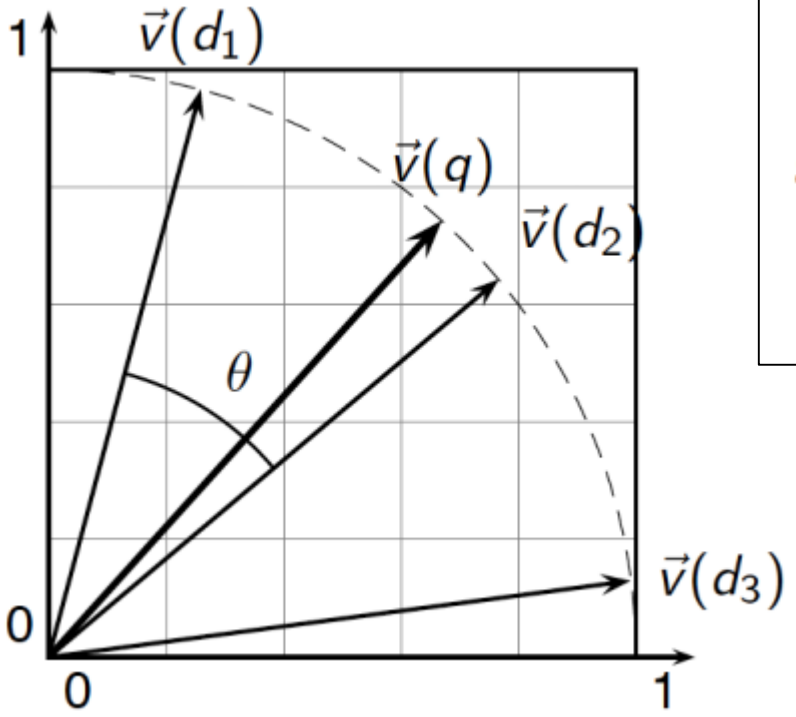
Konsep dasar dari IR adalah **pengukuran kesamaan**

- Sebuah perbandingan antara dua dokumen, mengukur sebearapa mirip keduanya.
- Setiap input query yang diberikan, dapat dianggap sebagai sebuah dokumen yang akan dicocokan dengan dokumen-dokumen lain.



| #4 Doc Similarity

- Salah satu metode yang paling populer dalam mengukur kemiripan sebuah dokumen yaitu ***cosine similarity*.**
- ***Cosine similarity*** termasuk dalam metode vector space, yaitu mencari kesamaan antar dokumen melalui penggambaran vektor sebagai suatu dokumen.



$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Ingat?

	0°	30°	45°	60°	90°
cos α	1	$\frac{1}{2}\sqrt{3}$	$\frac{1}{2}\sqrt{2}$	$\frac{1}{2}$	0

| #4 Doc Similarity (Simulasi Cosine Similarity)

Langkah 1: Mengekstrak Teks

Misalnya terdapat 3 Dokumen seperti di bawah ini:

This is a data mining course.	This is a data mining course
We are studying text mining. Text mining is a subfield of data mining.	We are studying text mining Text mining is a subfield of data mining
Mining text is interesting, and I am interested in it.	Mining text is interesting and I am interested in it

2. Langkah 2: Menghilangkan Stop Words

This is a data mining course.	~~This is a~~ data mining course
We are studying text mining. Text mining is a subfield of data mining.	~~We are~~ studying text mining Text mining ~~is a~~ subfield ~~of~~ data mining
Mining text is interesting, and I am interested in it.	Mining text ~~is~~ interesting ~~and I am~~ interested ~~in it~~

| #4 Doc Similarity (Simulasi Cosine Similarity)

3. Langkah 3: Ubah semua kata ke huruf kecil

This is a data mining course. → ~~This is a~~ data mining course

We are studying text mining. Text mining is a subfield of data mining. → ~~We are~~ studying text mining ~~Text~~ text mining ~~is a~~ subfield ~~of~~ data mining

Mining text is interesting, and I am interested in it. → ~~Mining~~ mining text ~~is~~ interesting ~~and I am~~ interested ~~in it~~

4. Langkah 4: Stemming

This is a data mining course. → ~~This is a~~ data ~~mining~~ mine course

We are studying text mining. Text mining is a subfield of data mining. → ~~We are~~ ~~studying~~ study text ~~mining~~ mine ~~Text~~ text ~~mining~~ mine ~~is a~~ subfield ~~of~~ data ~~mining~~ mine

Mining text is interesting, and I am interested in it. → ~~Mining~~ mine text ~~is~~ ~~interesting~~ ~~and I am~~ ~~interested~~ ~~in it~~ interest interest

| #4 Doc Similarity (Simulasi Cosine Similarity)

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

This is a data mining course. → This is a data mining course
course:1, data:1, mine:1

We are studying text mining. Text mining is a subfield of data mining. → We are studying text mining Text mining is a subfield of data mining
data:1, mine:3, study:1, subfield:1, text:2

Mining text is interesting, and I am interested in it. → Mining text is interesting and I am interested in it
Interest:2, mine:1, text:1

6. Langkah 6: Membuat File Index

This is a data mining course. → This is a data mining course
course:1, data:1, mine:1

We are studying text mining. Text mining is a subfield of data mining. → We are studying text mining Text mining is a subfield of data mining
data:1, mine:3, study:1, subfield:1, text:2

Mining text is interesting, and I am interested in it. → Mining text is interesting and I am interested in it
interest:2, mine:1, text: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

| #4 Doc Similarity (Simulasi Cosine Similarity)

7. Langkah 7: Membuat Model Ruang Vektor

This is a data mining course.

This is a data 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 text mining Text mining is a subfield of data 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 text is interesting and I am interested in it

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

#4 Doc Similarity (Simulasi Cosine Similarity)

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)

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

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

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

| #4 Doc Similarity (Simulasi Cosine Similarity)

10. Langkah 10: Normalkan Semua Dokumen ke Panjang Unit

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)

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

$$w(word_i) = \frac{w(word_i)}{\sqrt{w^2(word_1) + w^2(word_2) + \cdots + w^2(word_n)}}$$

Normalisasi panjang vektor

$$\vec{d} = [x_1, y_1, x_1]$$

$$\left|\vec{d}\right| = \left[\frac{x_1}{\sqrt{x_1^2 + y_1^2 + z_1^2}}, \frac{y_1}{\sqrt{x_1^2 + y_1^2 + z_1^2}}, \frac{z_1}{\sqrt{x_1^2 + y_1^2 + z_1^2}}\right]$$

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)

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

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

| #4 Doc Similarity (Simulasi Cosine Similarity)

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**)

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

| #4 Doc Similarity (Simulasi Cosine Similarity)

Penanganan Query

7. Hitung kemiripan antara Query dan Daftar Dokumen menggunaan 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

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

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

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

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

1

Kesimpulan: **Mengembalikan Dokumen #3**

Source : https://www.itee.uq.edu.au/

TASK

Hitung Cosine Similarity

Instruksi

1. Ambil hasil TF-IDF pada Quiz 1
2. Buat 1 query yang terdiri dari **minimal 3 kata** yang memiliki similarity dengan minimal 2 dokumen yg ada pada quiz 1 Anda
3. Hitung nilai *cosine similarity*nya lalu urutkan dokumen mana yang paling mirip
4. Contoh cara hitung cosine similarity dan laporan serta formatnya dapat dilihat di classroom

Thank you

</ti>
FST | UINAM
teknik informatika

UNIVERSITAS ISLAM NEGERI
ALAUDDIN
MAKASSAR