

## KOM341

### Temu Kembali Informasi

---

KULIAH #7

- Text Classification

### Ad Hoc Retrieval

---

- User mencari informasi dengan memberikan satu atau lebih query terhadap koleksi terkini.
- Contoh: mencari *multicore computer chips* terbaru.
  - Query : multicore AND computer AND chip
  - Akan dieksekusi setiap ada penambahan dokumen baru → standing query
  - Mungkin tidak menemukan artikel baru lain yang relevan, misalnya multicore processors.
  - Gunakan Boolean: (multicore OR multi-core) AND (chip OR processor OR microprocessor)

---

JAS - DEPT. ILMU KOMPUTER IPB 2

### Classification

---

- Lebih mudah kalau dokumen dikelompokkan menjadi misalnya dua kelas, yaitu dokumen tentang multicore computer chips dan dokumen BUKAN tentang multicore computer chips.
- Kelas biasanya merujuk ke topik dokumen.
- Prosesnya sering disebut sebagai text classification, text categorization, topic classification, topic spotting.

---

JAS - DEPT. ILMU KOMPUTER IPB 3

### Categorization/Classification

---

- Given:
  - Deskripsi dokumen  $d \in X$ , dimana  $X$  adalah kumpulan dokumen.
  - Himpunan kelas atau kategori:  $C = \{c_1, c_2, \dots, c_n\}$
- Tujuan:
  - Menentukan kategori dari  $d$ :  $c(d) \in C$ , dimana  $c(d)$  adalah fungsi kategorisasi (*classifier*).

---

JAS - DEPT. ILMU KOMPUTER IPB 4

### Document Classification

---

**Test Data:**

"planning language proof intelligence"

**Classes:**

	(AI)	(Programming)	(HCI)			
	ML	Planning	Semantics	Garb.Coll.	Multimedia	GUI

**Training Data:**

learning	planning	programming	garbage	...	...
intelligence	temporal	semantics	collection		
algorithm	reasoning	language	memory		
reinforcement	plan	proof...	optimization		
network...	language...		region...		

---

JAS - DEPT. ILMU KOMPUTER IPB 5

### Learning Method

---

- Kita mempelajari fungsi klasifikasi yang memetakan dokumen ke kategori tertentu:  $\gamma : X \rightarrow C$
- Disebut juga *supervised learning*, karena supervisor (orang yang menentukan kategori dokumen) berperan langsung di dalam proses pembelajaran.

---

JAS - DEPT. ILMU KOMPUTER IPB 6

### Metode

- Manual
  - Digunakan oleh Yahoo!, Looksmart, about.com, ODP, Medline
  - Sangat akurat karena dilakukan oleh ahli.
  - Konsisten pada saat ukurannya kecil/sedikit.
  - Sulit dan mahal

---

JAS - DEPT. ILMU KOMPUTER IPB 7

### Metode

- Automatic document classification
  - Hand-coded rule-based systems
    - Digunakan oleh CS dept's spam filter, Reuters, CIA, Verity, ...
    - Masukkan ke kategori jika dokumen mengandung kombinasi kata tertentu.
    - Akurasi tinggi jika rule dibuat dengan sangat baik oleh ahli dan kompleks.

---

JAS - DEPT. ILMU KOMPUTER IPB 8

### Metode

- Automatic document classification
  - Supervised learning
    - Beberapa menggunakan machine learning (Autonomy, MSN, Verity, Enkata, Yahoo!, ...)
    - k-Nearest Neighbors (simple, powerful)
    - Naive Bayes (simple, common method)
    - Support-vector machines (new, more powerful)
    - dsb
    - Membutuhkan hand-classified training data
    - Data dapat dibangun oleh amatir
- Banyak sistem komersial menggunakan metode campuran

---

JAS - DEPT. ILMU KOMPUTER IPB 9

### Metode Bayes

- Berbasis teori peluang
- Utamanya teorema Bayes
- Untuk kejadian **a** dan **b**, Bayes Rules:

$$p(a, b) = p(a \cap b) = p(a | b) p(b) = p(b | a) p(a)$$

$$p(\bar{a} | b) p(b) = p(b | \bar{a}) p(\bar{a})$$

$$p(a | b) = \frac{p(b | a) p(a)}{p(b)} = \frac{p(b | a) p(a)}{\sum_{x=a, \bar{a}} p(b | x) p(x)}$$

← Prior

Posterior

---

JAS - DEPT. ILMU KOMPUTER IPB 10

### Naïve Bayes Model

- Supervised learning method
- Multinomial Naïve Bayes Model
- Peluang dokumen **d** dalam kelas **c** :

$$P(c | d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k | c)$$

dimana  $P(t_k | c)$  adalah peluang term  $t_k$  muncul pada dokumen kelas  $c$ ,  $P(c)$  peluang dokumen ada pada kelas  $c$ .

---

JAS - DEPT. ILMU KOMPUTER IPB 11

### Pendugaan Parameter

Pendugaan parameter

$$\hat{P}(c) = \frac{N_c}{N}, \quad \hat{P}(t | c) = \frac{T_{ct}}{\sum_{t \in V} T_{ct}}$$

dimana  $N_c$  adalah banyaknya dokumen dalam kelas  $c$ ,  $N$  adalah total dokumen,  $T_{ct}$  adalah banyaknya  $t$  dalam dokumen training dari kelas  $c$ .

---

JAS - DEPT. ILMU KOMPUTER IPB 12

### Laplace smoothing

- Atau Add-One Smoothing.
- Untuk menghilangkan dugaan parameter yang bernilai nol.

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t \in V} (T_{ct} + 1)} = \frac{T_{ct} + 1}{\left(\sum_{t \in V} T_{ct}\right) + B'}$$

dimana  $B = |V|$  = banyaknya term dalam vocabulary.

---

JAS - DEPT. ILMU KOMPUTER IPB 13

### Contoh

	docID	words in document	in c = China?
TRAINING SET	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
TEST SET	5	Chinese Chinese Chinese Tokyo Japan	?

$P(c) = 3/4$  dan  $P(-c) = 1/4$   
 $P(\text{Chinese}|c) = (5+1)/(8+6) = 6/14 = 3/7$   
 $P(\text{Tokyo}|c) = P(\text{Japan}|c) = (0+1)/(8+6) = 1/14$   
 $P(\text{Chinese}|-c) = (1+1)/(3+6) = 2/9$   
 $P(\text{Tokyo}|-c) = P(\text{Japan}|-c) = (1+1)/(3+6) = 2/9$   
 $\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003$   
 $\hat{P}(\bar{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001$

---

JAS - DEPT. ILMU KOMPUTER IPB 14

### Bernoulli Model

- Kejadian Bernoulli
- Multivariate Bernoulli Model
- $\hat{P}(t|c)$  : rasio dokumen dari kelas c yang mengandung term t. Dalam multinomial didefinisikan sebagai rasio token dalam dokumen kelas c yang mengandung term t.

---

JAS - DEPT. ILMU KOMPUTER IPB 15

### Contoh

	docID	words in document	in c = China?
TRAINING SET	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
TEST SET	5	Chinese Chinese Chinese Tokyo Japan	?

$P(c) = 3/4$  dan  $P(-c) = 1/4$   
 $P(\text{Chinese}|c) = (3+1)/(3+2) = 4/5$   
 $P(\text{Tokyo}|c) = P(\text{Japan}|c) = (0+1)/(3+2) = 1/5$   
 $P(\text{Beijing}|c) = P(\text{Shanghai}|c) = P(\text{Macao}|c) = (1+1)/(3+2) = 2/5$   
 $P(\text{Chinese}|-c) = (1+1)/(1+2) = 2/3$   
 $P(\text{Tokyo}|-c) = P(\text{Japan}|-c) = (1+1)/(1+2) = 2/3$   
 $P(\text{Beijing}|c) = P(\text{Shanghai}|c) = P(\text{Macao}|c) = (0+1)/(1+2) = 1/3$

---

JAS - DEPT. ILMU KOMPUTER IPB 16

### Contoh

$$\hat{P}(c|d_5) \propto \hat{P}(c) \cdot \hat{P}(\text{Chinese}|c) \cdot \hat{P}(\text{Japan}|c) \cdot \hat{P}(\text{Tokyo}|c) \cdot (1 - \hat{P}(\text{Beijing}|c)) \cdot (1 - \hat{P}(\text{Shanghai}|c)) \cdot (1 - \hat{P}(\text{Macao}|c))$$

$$= 3/4 \cdot 4/5 \cdot 1/5 \cdot (1 - 2/5) \cdot (1 - 2/5) \cdot (1 - 2/5) \approx 0.005$$

$$\hat{P}(\bar{c}|d_5) \propto 1/4 \cdot 2/3 \cdot 2/3 \cdot 2/3 \cdot (1 - 1/3) \cdot (1 - 1/3) \cdot (1 - 1/3) \approx 0.022$$

Jadi, dokumen  $d_5$  diklasifikasikan ke  $-c$  (bukan China)

---

JAS - DEPT. ILMU KOMPUTER IPB 17

### Maximum a Posteriori

- Tujuan klasifikasi: mendapatkan kelas terbaik untuk suatu dokumen.
- Kelas terbaik : sangat mirip atau maximum a posteriori (MAP) kelas  $c_{map}$  :

$$c_{map} = \arg \max_{c \in C} \hat{P}(c|d) = \arg \max_{c \in C} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

---

JAS - DEPT. ILMU KOMPUTER IPB 18

### Maximum a Posteriori

$$c_{map} = \arg \max_{c \in C} P(c | d) = \arg \max_{c \in C} \frac{P(d | c)P(c)}{P(d)}$$

$$= \arg \max_{c \in C} P(d | c)P(c)$$

- Multinomial  $P(d|c) = P(\langle t_1, \dots, t_k, \dots, t_{n_d} \rangle | c)$
- Bernoulli  $P(d|c) = P(\langle e_1, \dots, e_k, \dots, e_M \rangle | c)$

JAS - DEPT. ILMU KOMPUTER IPB 19

### Asumsi Saling Bebas

- Kejadian A dan B saling bebas  
 $P(A \cap B) = P(A, B) = P(A) \cdot P(B)$
- Maka:
  - Multinomial*  $P(d | c) = P(\langle t_1, \dots, t_{n_d} \rangle | c) = \prod_{1 \leq k \leq n_d} P(X_k = t_k | c)$
  - Bernoulli*  $P(d | c) = P(\langle e_1, \dots, e_M \rangle | c) = \prod_{1 \leq i \leq M} P(U_i = e_i | c)$

JAS - DEPT. ILMU KOMPUTER IPB 20

### Multinomial vs Bernoulli

	multinomial model	Bernoulli model
event model	generation of token	generation of document
random variable(s)	$X = t$ iff $t$ occurs at given pos	$U_i = 1$ iff $t$ occurs in doc
document representation	$d = (t_1, \dots, t_k, \dots, t_{n_d}), t_k \in V$	$d = (e_1, \dots, e_i, \dots, e_M), e_i \in \{0, 1\}$
parameter estimation	$\hat{P}(X = t   c)$	$\hat{P}(U_i = e_i   c)$
decision rule: maximize	$\hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(X = t_k   c)$	$\hat{P}(c) \prod_{1 \leq i \leq M} \hat{P}(U_i = e_i   c)$
multiple occurrences	taken into account	ignored
length of docs	can handle longer docs	works best for short docs
# features	can handle more	works best with fewer
estimate for term the	$\hat{P}(X = t   c) \approx 0.05$	$\hat{P}(U_{doc} = 1   c) \approx 1.0$

JAS - DEPT. ILMU KOMPUTER IPB 21

### Vector Space Classification

### Klasifikasi Menggunakan Ruang Vektor

- Setiap dokumen training direpresentasikan sebagai vektor.
- Setiap titik (vektor) dokumen training diberi label sesuai dengan kelasnya.

JAS - DEPT. ILMU KOMPUTER IPB 23

### Test Document = Government ?

Similarity hypothesis true in general?

- Government
- Science
- Arts

JAS - DEPT. ILMU KOMPUTER IPB 24

### Rocchio Classification

Centroid dari kelas  $c$ :

$$\bar{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d)$$

JAS - DEPT. ILMU KOMPUTER IPB 25

### Rocchio Classification

Batas antara dua kelas adalah titik yang memiliki jarak sama ke kedua centroid-nya  $\rightarrow$

$$|a_1| = |a_2|,$$

$$|b_1| = |b_2|,$$

$$|c_1| = |c_2|$$

JAS - DEPT. ILMU KOMPUTER IPB 26

### Rocchio Classification

Dokumen  $d$  dikelompokkan ke dalam kelas  $c$

- Menggunakan jarak
 
$$\arg \min_c |\bar{\mu}_c - \vec{v}(d)|$$
- Menggunakan ukuran kesamaan Cosine
 
$$\arg \max_c \cos(\bar{\mu}(c), \vec{v}(d))$$

JAS - DEPT. ILMU KOMPUTER IPB 27

### Contoh

Dari contoh sebelumnya, diperoleh:

vector	term weights					
	Chinese	Japan	Tokyo	Macao	Beijing	Shanghai
$\vec{d}_1$	0	0	0	0	1.0	0
$\vec{d}_2$	0	0	0	0	0	1.0
$\vec{d}_3$	0	0	0	1.0	0	0
$\vec{d}_4$	0	0.71	0.71	0	0	0
$\vec{d}_5$	0	0.71	0.71	0	0	0
$\bar{\mu}_c$	0	0	0	0.33	0.33	0.33
$\bar{\mu}_c$	0	0.71	0.71	0	0	0

Jarak  $d_5$  terhadap centroid:

- $|\mu_c - d_5| \approx 1.15$  dan  $|\mu_c - d_5| \approx 0.00$
- maka Rocchio mengklasifikasikan  $d_5$  ke kelas  $-c$  (bukan China).

JAS - DEPT. ILMU KOMPUTER IPB 28

### k Nearest Neighbor Classification

- Mengklasifikasikan dokumen  $d$  ke dalam kelas  $c$
- Tentukan k-neighborhood  $N$  atau kNN sebagai  $k$  terdekat dari  $d$
- Hitung banyaknya dokumen  $i$  dalam  $N$  pada kelas  $c$
- Duga nilai  $P(c|d) = i/k$
- Pilih

$$c_{map} = \arg \max_{c \in C} P(c|d)$$

JAS - DEPT. ILMU KOMPUTER IPB 29

### Contoh: k=6 (6NN)

$P(\text{science} | \diamond) ?$

- Government
- Science
- Arts

JAS - DEPT. ILMU KOMPUTER IPB 30

### Ukuran Kemiripan

- Metode kNN tergantung pada ukuran kemiripan (bisa juga jarak) yang digunakan.
- Paling sederhana adalah jarak Euclidean.
- Untuk teks, yang paling efektif adalah ukuran kemiripan cosine dengan bobot vektor tf.idf.
- Skor dokumen di suatu kelas:

$$\text{score}(c, d) = \sum_{d' \in S_c} I_c(d') \cos(\vec{v}(d'), \vec{v}(d))$$

dimana  $I_c(d') = 1$  jika  $d'$  ada dalam kelas  $c$ , dan sebaliknya = 0.

---

JAS - DEPT. ILMU KOMPUTER IPB 31

### Contoh : 1NN

vector	term weights					
	Chinese	Japan	Tokyo	Macao	Beijing	Shanghai
$\vec{d}_1$	0	0	0	0	1.0	0
$\vec{d}_2$	0	0	0	0	0	1.0
$\vec{d}_3$	0	0	0	1.0	0	0
$\vec{d}_4$	0	0.71	0.71	0	0	0
$\vec{d}_5$	0	0.71	0.71	0	0	0

Dengan menggunakan jarak Euclidean, maka:  
 $|d_1 - d_5| = |d_2 - d_5| = |d_3 - d_5| = 1.4171$   
 $|d_4 - d_5| = 0.0000$   
 Maka  $d_5$  lebih dekat ke kelas  $d_4$ .

---

JAS - DEPT. ILMU KOMPUTER IPB 32

### Kombinasi Metode Klasifikasi

Beberapa peneliti menunjukkan bahwa kombinasi beberapa classifier yang berbeda dapat meningkatkan akurasi.

Classifier 1:  
 $X \rightarrow \text{class1}$   
 Classifier 2:  
 $X \rightarrow \text{class2}$   
 Jadi, X dimasukkan kemana?

---

JAS - DEPT. ILMU KOMPUTER IPB 33

### Kombinasi Metode Klasifikasi

- Simple voting  
 Untuk tiap dokumen test, kita klasifikasikan ke kelas  $c_i$  jika mayoritas classifier memasukkan dokumen test ke kelas  $c_i$ .
- Dynamic classifier selection (DCS)  
 Pendekatan kNN dengan ukuran kesamaan Cosine, dilakukan iterasi.
- Adaptive classifier combination (ACC)  
 Kombinasi NB dan kNN

---

JAS - DEPT. ILMU KOMPUTER IPB 34