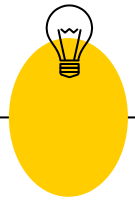
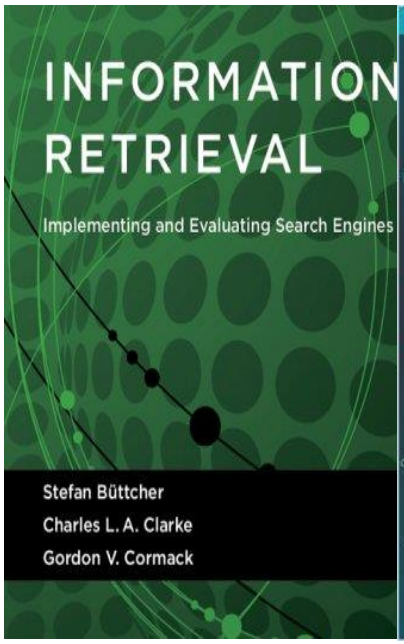
Sistem Temu Kembali Informasi

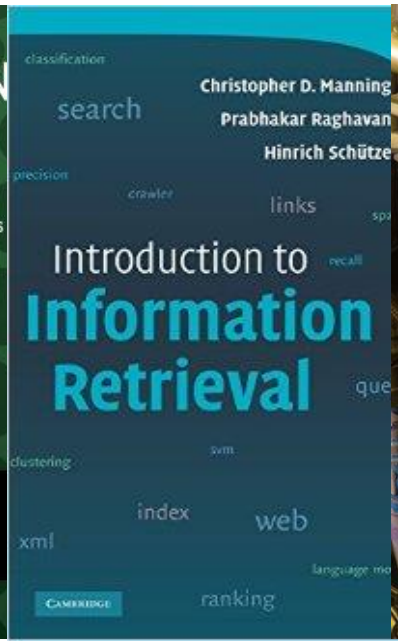
"Klasifikasi Dokumen dengan Naïve Bayes"

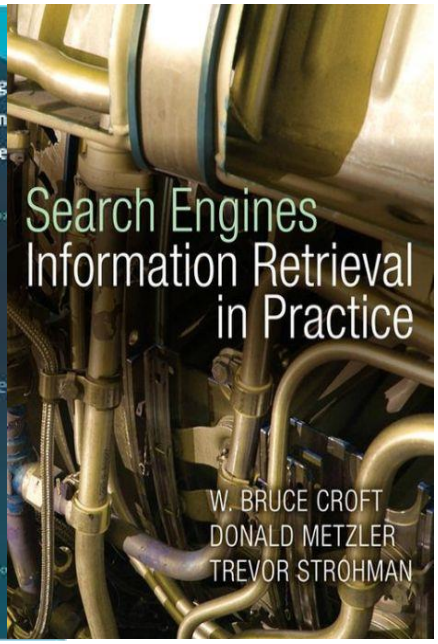


Tim pengampu Dosen STKI

Buku Penunjang & Literatur







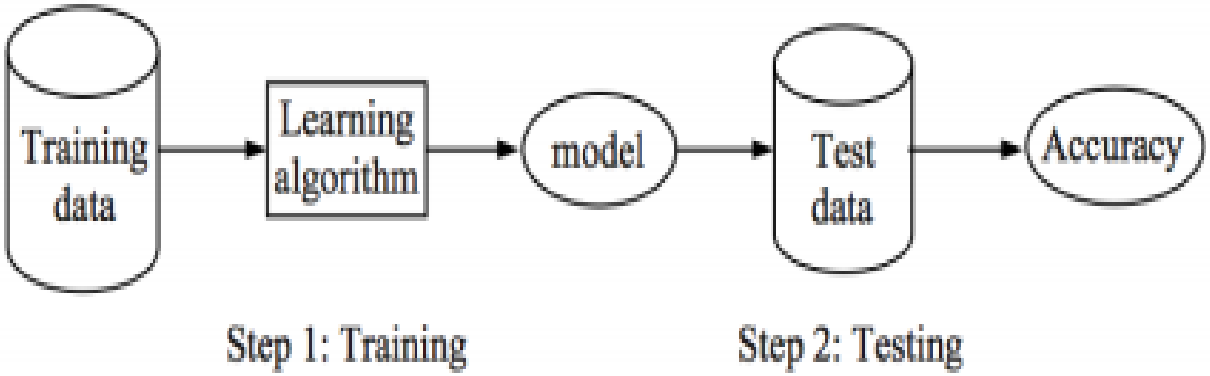
Course Outline

- Klasifikasi Text
- Naive Bayes Classifier (NBC)
- Contoh Penyelesaian Kasus

Klasifikasi Teks

- Masalah klasifikasi adalah bagaimana menentukan suatu objek masuk ke suatu class yang sebenarnya.
  - Dalam pemrosesan text, suatu class lebih bersifat area subjek umum (disebut juga topik)
  - Contoh Implementasi
    - Sentiment detection
    - Mendeteksi encoding dokumen
    - Topic-specific (vertical-search)
    - Mendeteksi otomatis halaman / email spam
    - …….

Proses klasifikasi



Naive Bayes Classifier

- Bayesian Classification adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class.
- Bayesian Classification didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa *decision tree* dan *neural network*.
- Bayesian Classification terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan pada data yang besar.

Naive Bayes Classifier

- *Metode NBC* menempuh dua tahap dalam proses klasifikasi teks, yaitu ***tahap pelatihan*** dan ***tahap klasifikasi***.

Tahap pelatihan:

- Proses analisis terhadap sampel dokumen berupa pemilihan vocabulary.
  (kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen).
- Penentuan probabilitas prior bagi tiap kategori berdasarkan sampel dokumen.

Tahap klasifikasi:

- Ditentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang diklasifikasi

Rumus Teorema Bayes

Teorema Bayes memiliki bentuk umum sbb :

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

Keterangan :

X : data dengan *class* yang belum diketahui

H : hipotesis data X merupakan suatu *class* spesifik

P(H|X) : probabilitas hipotesis H berdasar kondisi X (*posteriori probability*)

P(H) : probabilitas hipotesis H (*prior probability*)

P(X|H) : probabilitas X berdasar kondisi hipotesis H

P(X) : probabilitas dari X

Naive Bayes Classifier

- Diasumsikan suatu koleksi dokumen $D=\{di \mid i=1,2,\ldots|D|\}=\{d1,d2,\ldots,d_{|D|}\}$ dan koleksi kategori / kelas $C=\{cj|j=1,2,\ldots|C|\}=\{c1,c2,\ldots,c_{|c|}\}$.
- Klasifikasi NBC dilakukan dengan cara mencari probabilitas P(C=cj | D=di), yaitu probabilitas category *cj* jika diketahui dokumen *di* dipandang sebagai tuple dari kata-kata dalam dokumen, yaitu $<w_1, w_2,\ldots,w_n>$, yang frekuensi kemunculannya diasumsikan sebagai variable random.
- $P(C|d) = P(C)\prod_{i=1}^{n} P(w_i|C)$ (1)
- $P(w_i|C) = \frac{count(w_i,C)+1}{count(C)+|V|}$ (2)
- Dimana C adalah Class, D adalah dokumen, w_i adalah kata w ke i, $count(w_i, C)$ adalah jumlah kata w_i dalam C, $count(C\ )$ adalah jumlah kata di class C, $|V|$ adalah jumlah *vocabulary*.

Contoh kasus

- Dengan menggunakan algoritma naïve bayes, tentukan class dari D6 berikut:

		Teks	Class
Training	D1	Sepakbola indah menyerang	Olahraga
	D2	Presiden menaikan harga BBM	Politik
	D3	Partai politik Indonesia berburu suara	Poitik
	D4	Manchester United Juara Liga Inggris	Olahraga
	D5	Timnas Indonesia gagal juara AFC	Olahraga
Testing	D6	PSIS berburu juara Liga Indonesia	?

Pembahasan

- Mengacu pada persamaan :
- $P(C|d) = P(C) \prod_{i=1}^{n} P(w_i|C)$
- $P(w_i|C) = \frac{count(w_i,C)+1}{count(C)+|V|}$
- Dimana C adalah Class, D adalah dokumen, w_i adalah kata w ke i, $count(w_i, C)$ adalah jumlah kata w_i dalam C, $count(C\ )$ adalah jumlah kata di class C, $|V|$ adalah jumlah *vocabulary*.

Tokenisasi

$$P(C|d) = P(C)\prod_{i=1}^{n} P(w_i|C)$$

$$P(w_i|C) = \frac{count(w_i,C)+1}{count(C)+|V|}$$

$|V|$ = 20 (total vocabulary)

$count(C)$ = 12 (untuk kelas Olahraga)

$count(C)$ = 9 (untuk kelas Politik)

P (Politik) = 2/5 = 0,4

P (Olahraga) = 3/5 = 0,6

1	sepakbola	
2	indah	
3	menyerang	
4	presiden	
5	menaikan	
6	harga	
7	bbm	
8	partai	
9	politik	
10	indonesia	
11	berburu	
12	suara	
13	Manchester	
14	United	
15	Juara	
16	liga	
17	inggris	
18	timnas	Indonesia
19	gagal	
20	AFC	

Data testing

- PSIS berburu juara Liga Indonesia

$$P(w_i|C) = \frac{count(w_i,C)+1}{count(C)+|V|}$$

KELAS OLAHRAGA

P(PSIS | Kelas Olahraga) = (0+1)/(12+20) = 0,03125

P(berburu | Kelas Olahraga) = (0+1)/(12+20) = 0,03125

P(juara | Kelas Olahraga) =(2+1)/(12+20) = 0,09375

P(liga | Kelas Olahraga) =(1+1)/(12+20) = 0,0625

P(Indonesia | Kelas Olahraga) =(1+1)/(12+20) = 0,0625

Data testing

KELAS POLITIK

P(PSIS | Kelas Politik) = (0+1)/(9+20) = 0,03448

P(berburu | Kelas Politik) = (1+1)/(9+20) = 0,06897

P(juara | Kelas Politik) =(0+1)/(9+20) = 0,03448

P(liga | Kelas Politik) =(0+1)/(9+20) = 0,03448

P(Indonesia | Kelas Politik) =(1+1)/(9+20) = 0,06897

Data testing

- $P(C|d) = P(C)\prod_{i=1}^{n} P(w_i|C)$
- **Kelas Olahraga**
- $P(C|d)$ =0,6 * 0,03125 * 0,03125 * 0,09375 * 0,0625 * 0,0625 = 2,14577 * 10^{-7}
- **Kelas Politik**
- $P(C|d)$ = 0,4 * 0,03448 * 0,06897 * 0,03448 * 0,03448 * 0,06897 = 7,79978 * 10^{-8}

*Hasil: Dokumen Testing termasuk kedalam **Kelas Olahraga***

Kesimpulan & Review

- Bayesian Classification adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class.
- Bayesian Classification terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan pada data yang besar.
- Metode NBC menempuh dua tahap dalam proses klasifikasi teks yaitu :

1. Tahap pelatihan:
   - Proses analisis terhadap sampel dokumen berupa pemilihan vocabulary.
   - Penentuan probabilitas prior bagi tiap kategori berdasarkan sampel dokumen.
2. Tahap klasifikasi:

   Menentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang di klasifikasi

Kuis (Latihan Soal)

- Dengan menggunakan algoritma naïve bayes, tentukan class dari D7 berikut:
- Preprocessing: tokenization, stemming

		Teks	Class
Training	D1	Timnas Indonesia gagal juara AFC	Olahraga
	D2	PSIS berburu juara Liga Indonesia	Olahraga
	D3	Partai politik Indonesia berburu suara	Poitik
	D4	Hasil putusan Sidang Elit Dewan	Politik
	D5	Harga cabai naik dipasar tradisional	Ekonomi
	D6	Upah minimum regional diprediksi naik	Ekonomi
Testing	D7	Demo buruh menaikan upah minimum	?

Thanks!

Any questions ?