Ahmad Rizky Fauzan

09021281722042

1. Terdapat tiga dokumen berikut ini:

   • D1: "andi pergi ke pasar dan budi pergi ke sekolah"

   • D2: "andi belajar bersama dengan budi"

   • D3: "andi budi dan candra bersama pergi ke sekolah dan pasar"

   Hitung TF-IDF untuk ketiga kalimat di atas secara manual! (stopword juga dihitung)

   Kamus : Andi, Pergi, ke, pasar, dan, budi, sekolah, belajar, Bersama, dengan, candra

Kamus	DF	IDF
Andi	3	Log(3/3) = 0
Budi	3	Log(3/3) = 0
Candra	1	Log(3/1) = 0,477
Pergi	3	Log(3/3) = 0
Ke	3	Log(3/3) = 0
Pasar	2	Log(3/2) = 0,176
Dan	3	Log(3/3) = 0
Sekolah	2	Log(3/2) = 0,176
Belajar	1	Log(3/1) = 0,477
Bersama	2	Log(3/2) = 0,176
dengan	1	Log(3/1) = 0,477

Kamus	TF			IDF	TF-IDF		
	D1	D2	D3		D1	D2	D3
Andi	1	1	1	0	0	0	0
Budi	1	1	1	0	0	0	0
Candra			1	0,477			0,477
Pergi	2		1	0	0		0
Ke	2		1	0	0		0
Pasar	1		1	0,176	0,176		0,176
Dan	1		2	0	0		0
Sekolah	1		1	0,176	0,176		0,176
Belajar		1		0,477		0,477	
Bersama		1	1	0,176		0,176	0,176
dengan		1		0,477		0,477	

HASIL

	A	B	C	P	Ke	Pasar	Dan	S	B	Ber	Deng
D1	0	0	0	0	0	0,176	0	0,176	0	0	0
D2	0	0	0	0	0	0	0	0	0,477	0,176	0,477
D3	0	0	0,477	0	0	0,176	0	0,176	0	0,176	0

2. Jika terdapat kueri berikut ini:

• kueri = "tampilkan laptop dengan prosesor i7, ssd 500GB, dan harga di bawah 10 juta?"

Buatlah program untuk mengambil informasi jenis prosesor, kapasitas harddisk, dan harga maksimal dari kueri tersebut menggunakan fungsi REGEX. Contoh hasilnya programnya: "[i7, 500GB, 10 juta]"

Import re
Mencari pola dari kata yang ingin diambil dari kueri.

```
import re
kueri = "tampilkan laptop dengan prosesor i7, ssd 500GB, dan harga di bawah 10 juta?"

prosesor = re.compile(r'prosesor (?:[a-zA-Z]|[0-9])+')
regex_1= re.findall(prosesor, kueri)
regex_1 = regex_1[0].split('prosesor ')

ssd = re.compile(r'ssd (?:[a-zA-Z]|[0-9])+')
regex_2= re.findall(ssd, kueri)
regex_2 = regex_2[0].split('ssd ')

juta = re.compile(r'\d+\s\w+')
regex_2 = re.findall(juta,kueri)
```

3. Pada artikel Dada et al. (2019) diuraikan beberapa isu (masalah) yang ada pada penelitian spam filtering saat ini. Sebutkan dan jelaskan isu-isu tersebut!

Masalah pada penelitian spam filtering di paper tersebut adalah

- Beberapa makalah berfokus pada metode bebas fitur untuk penyaringan spam email karena telah terbukti memiliki akurasi yang lebih tinggi daripada teknik berbasis fitur. Namun perlu dicatat bahwa teknik bebas fitur memiliki biaya komputasi yang tinggi karena biasanya memakan waktu lebih lama dalam tugas klasifikasi e-mailnya. Itu juga menderita kompleksitas implementasi.
- Beberapa penelitian dianggap menggunakan baris subjek, tajuk, dan badan pesan sebagai fitur paling penting dalam mengklasifikasikan pesan sebagai spam atau ham. Namun, perlu disebutkan bahwa baris subjek, tajuk, dan badan yang mencurigakan saja dapat menyebabkan kesalahan dalam klasifikasi email spam. Pengguna mungkin juga perlu memilih fitur secara manual.
- Peneliti lain menemukan bahwa model tas kata relative fitur efektif untuk memfilter spam dan email phishing, dan email header adalah fitur yang sama pentingnya dengan isi pesan dalam mendeteksi email spam.
- Sebagian besar peneliti tidak memasukkan biaya komputasi Pertimbangan dalam pilihan teknik pembelajaran mesin gunakan untuk memfilter email spam. Fokus utama mereka adalah kinerja akurasi klasifikasi.
- Beberapa peneliti menggunakan pola perilaku spammer sebagai aspek penting dari deteksi spam sementara algoritma pembelajaran mesin digunakan untuk mengekstraksi fitur-fitur penting dari Badan Pesan. Rekayasa fitur yang komprehensif mungkin diperlukan untuk akurasi yang lebih baik.

- Pada Deep learning, teknik pembelajaran mesin lainnya diterapkan pada penyaringan spam email memiliki batasan toleransi kesalahan rata-rata, kurangnya pemrosesan paralel dan kemampuan belajar mandiri yang rendah.