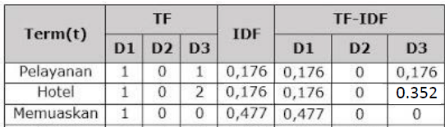
1. Tokenisasi, potong kata (- angka, tnd baca, …).
2. Case folding, huruf menjadi huruf kecil.
3. Stopword, kata ndak guna
4. Stamming, kata dasar
5. Term Weight (TF-IDF)



**Reduksi data** adalah sebuah proses analisis untuk memilih, memusatkan perhatian (focus), menyederhanakan, mengabstraksi serta mentransformasikan data.

Tujuan reduksi data:

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

\*relatif

**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.

**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.

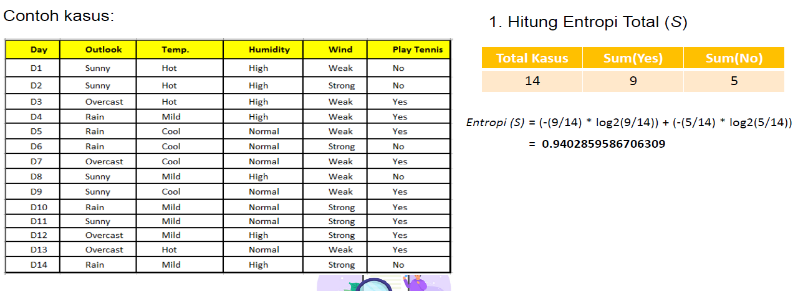
Dimana:

**S** adalah himpunan (dataset) kasus

**k** adalah banyaknya partisi **S**

**Pj** adalah probabilitas yang di dapat

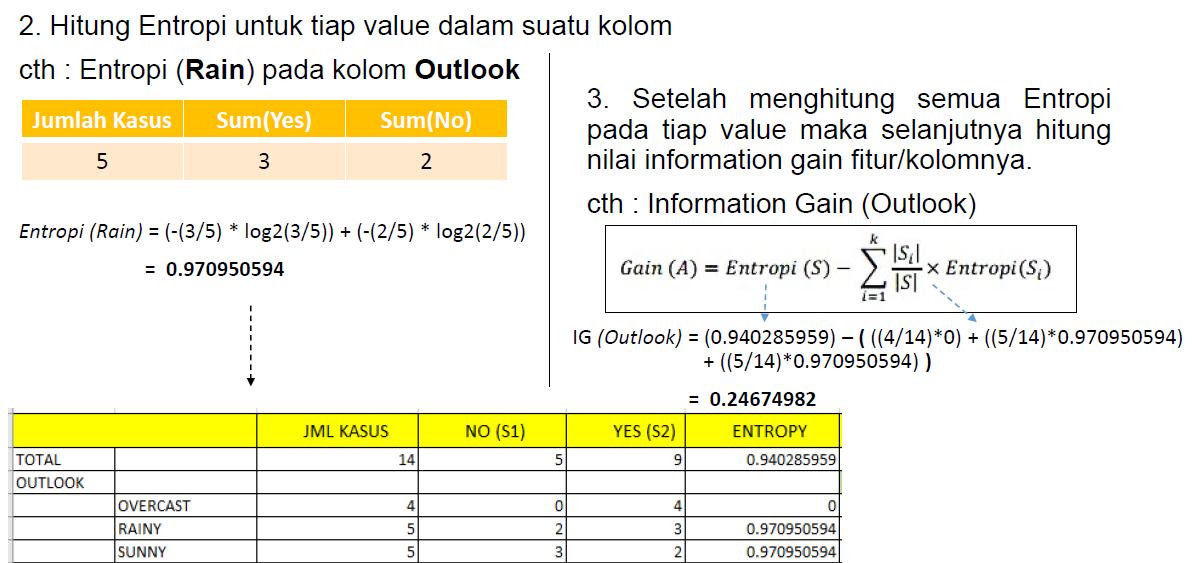
dari Sum(Ya) dibagi Total Kasus.

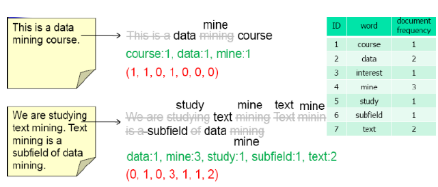


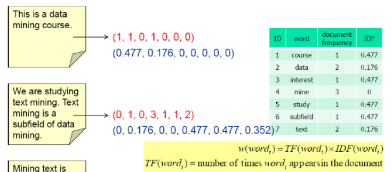
Setelah menemukan nilai entropi maka selanjutnya pemilihan fitur dilakukan dengan nilai information gain terbesar.

Dimana:

S = ruang (data) sample yang digunakan untuk training.

A = atribut. ISII = jumlah sample untuk nilai V. ISI = jumlah seluruh sample data. Entropi(S) = entropy untuk sample-sample yang memiliki nilai i

****

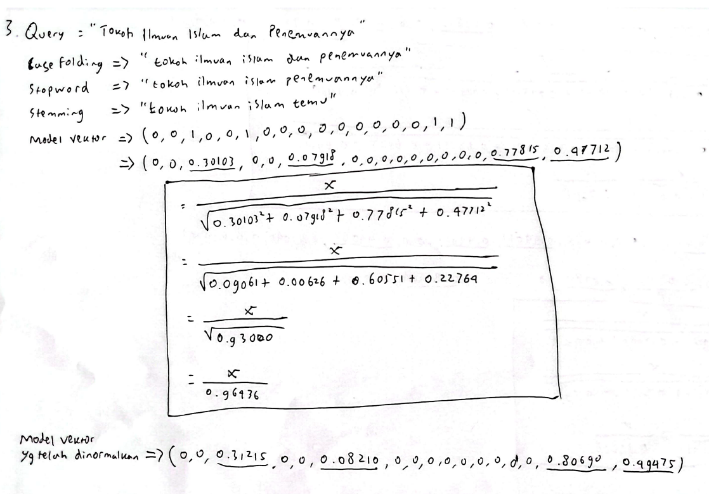


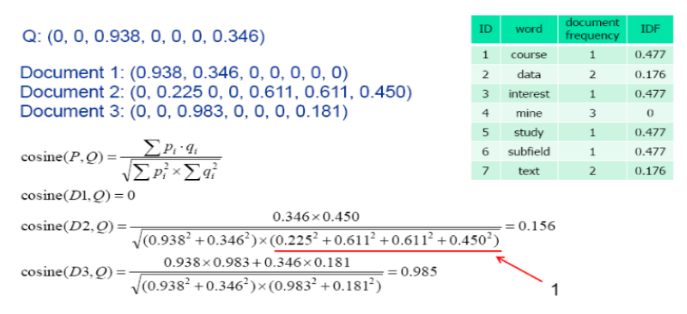
**Query** Buat tabel TF IDFnya. Susun vectornya berdasarkan abjad dan panjang data tabel tf-idf.

Abis itu..

Masukkan nilai IDF\*nWordInDoc ke dalam vector

Abis itu Normalisasi





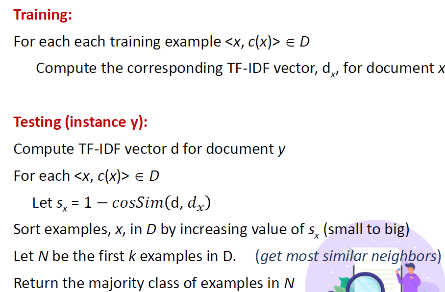
**Information extraction,** pencarian otomatis pada informasi yang terstruktur seperti entitas, hubungan antar entitas, dan atribut yang

menggambarkan entitas dari sumber yang tidak terstruktur. **Tugas?**

Mengumpulkan informasi dari banyak kumpulan teks; identifikasi entitas nama, relasi antar entitas, dan mengekstrak kejadian. **Tujuan?** Membuat informasi menjadi lebih terorganisir dengan baik sehingga berguna untuk manusia & Informasi ditampilkan dalam sebuah format yang

tepat secara semantic sehingga memungkinkan dilakukan inferensi pada tahap selanjutnya Oleh algoritma komputer.

**Named Entity Recognition**, salah satu Subtask yang sangat penting dalam IE : Menemukan dan Mengklasifikasi nama-nama Entitas dalam teks. **Tugas?** Melakukan Indeksi Entitas dsb.;Sentiment bisa disematkan pada perusahaan atau produk; NER juga bermanfaat dalam banyak aplikasi NLP (Natural Language Processing) seperti question-answering, rangkuman dan sistem

**Kelebihan k-NN**: Efektif untuk klasifikasi teks; Menangani kasus multi-class secara langsung; Tidak "membutuhkan model" (hanya menghitung representasi vektor); **Kekurangan**: Membutuhkanfine-tuning untuk nilai k (z 40 untuk klasifikasi teks); Membutuhkan proses adjustment ketika jumlah kelas yang ada pada data tidak berimbang.

