**Crawling adalah proses mesin pencari untuk menemukan halaman website gambar video dokumen dan lain sebagainya yang telah di update di sebuah situs.**

**Untuk pencarian sebuah konten di search engine dengan keyword tertentu maka akan mencari indeks dan konten yang mana yang paling sesuai untuk user tersebut.**

**Ada banyak beragam pilihan tools yang dapat digunakan untuk melakukan web Crawling salah satunya disebut sebagai web robot atau web Spider dan juga proses web rolling ini tidak dapat dilakukan secara manual**

**cara kerja Web Crawler**

1. **Web crawler akan menuju ke laman situs dan link**
2. **Masukkan URL situs di Google Search Console**
3. **Web Crawling akan melakukan pencatatan pada setiap link di indeks.**

**note : halaman/link yang bersifat privat tidak bisa diambil informasinya**

1. **informasi yang sudah terumpul maka akan di simpan di dalam indeks search engine sehingga muncul di konten dengan keyword yang sama**

**Contoh Web Crawler**

1. **Googlebot, web crawler milik Google yang paling banyak digunakan saat ini. Gogglebot akan membuat indeks yang akan bertugas untuk mengumpulkan informasi dari berbagai website.**
2. **HTTrack, web crawler yang bersifat open source. ketika user sudah mendownload aplikasi ini maka user bia membuka konten situs tanpa melalui koneksi internet**
3. **Cyotek WebCopy, sama halnya seperti HTTrack yang membedakan user dapat memilih bagian mana yang ingin di download.**
4. **Webhose, web crawler yang mengubah konten menjadi datafeeds**

# Text Preprocessing

**Data teks perlu dibersihkan dan dikodekan ke nilai numerik sebelum diberikan ke model pembelajaran mesin, proses pembersihan dan pengkodean ini disebut sebagai preprocessing teks.**

**Perlu dipersiapkan untuk library ini yakni natural language toolkit atau disingkat nltk yang merupakan library Python dengan pemodelan teks nltk ini juga menyediakan alat yang sebelum digunakan pada mesin learning atau Algo algoritma Deep learning untuk menginstal nltk menggunakan pip pada command Line atau terminal**

**pip install nltk**

**Sebelum menginstal nltk user harus mengunduh paket nltk**

**import nltk**

**nltk.download()**

1. Python Sastrawi / streaming bahasa Indonesia,**merupakan pengembangan dari proyek PHP sastrawi dan juga merupakan library sederhana yang dapat mengubah kata berimbuhan bahasa Indonesia menjadi bentuk dasarnya**

**instal melalui “pip”.**

**pip install Sastrawi**

**Penggunaan Python Sastrawi sangat sederhana seperti baris kode dibawah ini :**

**# import StemmerFactory class**

**from Sastrawi.Stemmer.StemmerFactory import StemmerFactory**

**# create stemmer**

**factory = StemmerFactory()**

**stemmer = factory.create\_stemmer()**

**# stemming process**

**sentence = 'Perekonomian Indonesia sedang dalam pertumbuhan yang membanggakan'**

**output = stemmer.stem(sentence)**

**print(output)**

**# ekonomi indonesia sedang dalam tumbuh yang bangga**

**print(stemmer.stem('Mereka meniru-nirukannya'))**

**# mereka tiru**

1. **Case Folding, Text Processing yang sederhana dan efektif yang mempunyai tujuan untuk merubah semua huruf didalam dokumen menjadi huruf kecil abjad huruf a sampai z yang diterima karakter selain objek tersebut dihilangkan dan diangkat delimiter**

## Mengubah text menjadi lowercase

**Lower case untuk mesin pencarian sangatlah penting contohnya untuk mencari dokumen yang mengandung Indonesia namun tidak ada yang muncul karena indonesia di indeks sebagai INDONESIA**

**Ini contoh source code Python untuk mengubah teks menjadi lower case**

**kalimat = "Berikut ini adalah 5 negara dengan pendidikan terbaik di dunia adalah Korea Selatan, Jepang, Singapura, Hong Kong, dan Finlandia."**

**lower\_case = kalimat.lower()**

**print(lower\_case)**

**# output**

**# berikut ini adalah 5 negara dengan pendidikan terbaik di dunia adalah korea selatan, jepang, singapura, hong kong, dan finlandia.**

## Menghapus angka

**Tulisan angka yang tidak relevan dengan apa yang akan dianalisa seperti nomor rumah nomor telepon dan lain-lain regular expression dapat juga digunakan untuk menghapus karakter angka**

**Contoh source code paytren untuk menghapus angka dalam sebuah kalimat**

**import re # impor modul regular expression**

**kalimat = "Berikut ini adalah 5 negara dengan pendidikan terbaik di dunia adalah Korea Selatan, Jepang, Singapura, Hong Kong, dan Finlandia."**

**hasil = re.sub(r"\d+", "", kalimat)**

**print(hasil)**

**# ouput**

**# Berikut ini adalah negara dengan pendidikan terbaik di dunia adalah Korea Selatan, Jepang, Singapura, Hong Kong, dan Finlandia.**

## Menghapus tanda baca

**Sama halnya dengan angka, tanda baca dalam kalimat tidak memiliki pengaruh pada *text preprocessing*. Menghapus tanda baca seperti [!”#$%&’()\*+,-./:;<=>?@[\]^\_`{|}~] dapat dilakukan di pyhton seperti dibawah ini :**

**kalimat = "Ini &adalah [contoh] kalimat? {dengan} tanda. baca?!!"**

**hasil = kalimat.translate(str.maketrans("","",string.punctuation))**

**print(hasil)**

**# output**

**# Ini adalah contoh kalimat dengan tanda baca**

## Menghapus whitepace (karakter kosong)

**Untuk menghapus spasi di awal dan akhir, anda dapat menggunakan fungsi strip()pada pyhton. Perhatikan kode dibawah ini :**

**kalimat = " \t ini kalimat contoh\t "**

**hasil = kalimat.strip()**

**print(hasil)**

**# output**

**# ini kalimat contoh**

**2. Tokenizing, Merupakan proses teks menjadi potongan-potongan yang disebut token yang akan dianalisa kata angka simbol dan tanda baca yang memiliki entitas penting dianggap sebagai token**

**Fungsi split()pada pyhton dapat digunakan untuk memisahkan teks. Perhatikan contoh dibawah ini :**

**kalimat = "rumah idaman adalah rumah yang bersih"**

**pisah = kalimat.split()**

**print(pisah)**

**# output**

**# ['rumah', 'idaman', 'adalah', 'rumah', 'yang', 'bersih']**