BOOK LAPORAN PPW

Pastikan dalam editor sudah terinstall library yang dibutuhkan yaitu dengan menjalankan perintah pip install name\_lybrary di editor masing – masing :

>> untuk crawling data dari web yang akan kita gunakan

import requests

from bs4 import BeautifulSoup

import csv

>> untuk menjadikan dalam bentuk kalimat

import nltk

from nltk.tokenize import sent\_tokenize

>> untuk proses menghitung tf-idf

import math

from collections import Counter

import pandas as pd

>> digunakan untuk menghitung kesamaan kosinus  
from sklearn.metrics.pairwise import cosine\_similarity

>> untuk menampilkan dalam bentuk graph

import matplotlib.pyplot as plt

Dalam Materi Kali ini kita akan mengulas tentang :

1. Ekstraksi Kalimat dalam bentuk (token sentences)
2. TF – IDF
3. Cosinus Semiliarity
4. Graph
5. Menghitung Closeness Centrality, Menentukan PageRank, Eigenvector Centrality

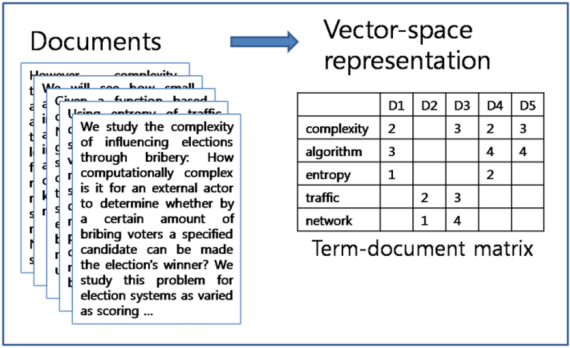
**Ekstraksi Kalimat dalam bentuk (token sentences)**

Langkah pertama yang perlu kita lakukan yaitu menggunakan kalimat sebagai ciri dari sebuah dokumen dimana data kalimat tersebut didapat dari crawling web berita deti.com dengan halaman berjumlah 1112 halaman untuk lebih jelasnya bisa kunjungi link berikut :

[*https://esairina.medium.com/scraping-berita-online-pada-situs-detik-com-menggunakan-google-colab-3a764981384b*](https://esairina.medium.com/scraping-berita-online-pada-situs-detik-com-menggunakan-google-colab-3a764981384b)TF-IDF AND COSINUS SIMILARITY

Tahapa Proses Perangkingan

1. Preproces Dokumen



1. Token Dokumen

1.A. TOKEN DOKUMEN

• D1: Universitas // Trunojoyo // tahun // ini // akan // milad // ke // 10

• D2: Informatika // Trunojoyo // satu-satunya // yang // terakreditasi // A // di // Universitas // Trunojoyo • D3: Mahasiswa // informatika // sedang demo // di // laboratorium

1. Menghapus Stopword

1.B. MENGHAPUS STOPWORD

• D1: Universitas // Trunojoyo // tahun // ini // akan // milad // ke // 21

• D2: Informatika // Trunojoyo // satu-satunya // yang // terakreditasi // A // di // Universitas // Trunojoyo • D3: Mahasiswa // informatika // sedang demo // di // laboratorium

1. Menentukan Term untuk Sentences

1.C. MENENTUKAN TERM



• D1: Universitas // Trunojoyo // tahun // milad • D2: Informatika // Trunojoyo // terakreditasi // Universitas // Trunojoyo

• D3: Mahasiswa // informatika // demo // laboratorium • Kata Kunci (KK): Universitas Trunojoyo

1.C. MENENTUKAN TERM

• D1: Universitas // Trunojoyo // tahun // milad • D2: Informatika // Trunojoyo // terakreditasi // Universitas // Trunojoyo

• D3: Mahasiswa // informatika // demo // laboratorium

• Kata Kunci (KK): Universitas Trunojoyo • Pilih unik term dan urutkan secara ascending

1. Hitung TF-IDF

MAPPING TERM DAN FREQUENSINYA

Memasukkan nilai frekuensi term (tf) yaitu jumlah kemunculan term dalam masing-masing dokumen

• D1: Universitas //

Trunojoyo // tahun // milad

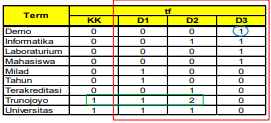
• D2: Informatika //

Trunojoyo // terakreditasi // Universitas // Trunojoyo • D3: Mahasiswa //

informatika // demo // laboratorium

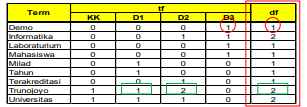
• Kata Kunci (KK):

Universitas Trunojoyo



MENGHITUNG DOKUMEN FREKUENSI (DF)

Df adalah jumlah dokumen yang didalamnya memuat term tertentu



MENGHITUNG NILAI TF DAN IDF

• idf = log (D/df)

– D = Jumlah Dokumen yang di perbandingkan

• Wdt = tf \* Idf



RANGKING berdasarkan TF-IDF

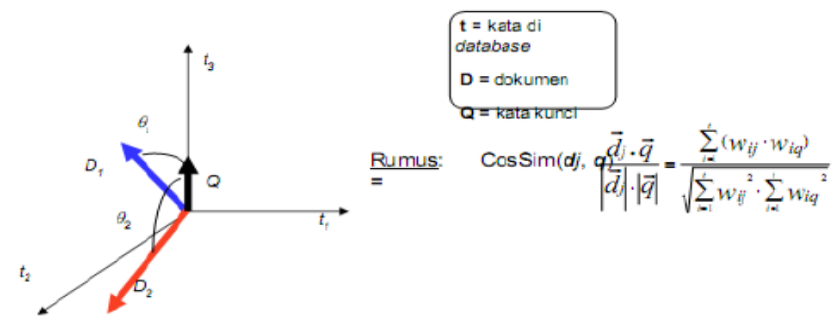
• D2 = 0,352 + 0.176 = 0,528 • D1 = 0,176 + 0,176 = 0,352 • D3 = 0.000 + 0.000 = 0,000

• Dari nilai tf-idf ini dapat disimpulkan bahwa D2 memiliki kesamaan paling tinggi kemudian D1. serta D3 tidak memiliki nilai kesamaan

1. **Hitung Cosinus Similarity**

Vector Space Model

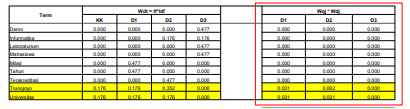
• Menghitung nilai cosinus sudut dari 2 vector yaitu W dari dokumen dengan W dari kata kunci



Menghitung Wij\* Wiq

• Melakukan perkalian skalar nilai

– Wkk (W kata Kunci) dengan – Wdi (W dokumen ke i)



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Term** | **Wdt = tf\*Idf** | | | |  | **Wqj \* Wdj** | | |
| **KK** | **D1** | **D2** | **D3** | **D1** | **D2** | **D3** |
| Demo | 0.000 | 0.000 | 0.000 | 0.477 |  | 0.000 | 0.000 | 0.000 |
| Informatika | 0.000 | 0.000 | 0.176 | 0.176 | 0.000 | 0.000 | 0.000 |
| Laboraturium | 0.000 | 0.000 | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 |
| Mahasiswa | 0.000 | 0.000 | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 |
| Milad | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Tahun | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Terakreditasi | 0.000 | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 | 0.000 |
| Trunojoyo | 0.176 | 0.176 | 0.352 | 0.000 | 0.031 | 0.062 | 0.000 |
| Universitas | 0.176 | 0.176 | 0.176 | 0.000 | 0.031 | 0.031 | 0.000 |

Menghitung ∑ Wij\* Wiq

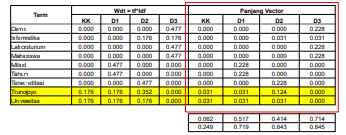
Nilai – ∑ Wd1\*Wkk = 0,062 – ∑ Wd2\*Wkk = 0,093 – ∑ Wd3\*Wkk = 0,000



|  |  |  |  |
| --- | --- | --- | --- |
| **Term** | **Wqj \* Wdj** | | |
| **D1** | **D2** | **D3** |
| Demo | 0.000 | 0.000 | 0.000 |
| Informatika | 0.000 | 0.000 | 0.000 |
| Laboraturium | 0.000 | 0.000 | 0.000 |
| Mahasiswa | 0.000 | 0.000 | 0.000 |
| Milad | 0.000 | 0.000 | 0.000 |
| Tahun | 0.000 | 0.000 | 0.000 |
| Terakreditasi | 0.000 | 0.000 | 0.000 |
| Trunojoyo | 0.031 | 0.062 | 0.000 |
| Universitas | 0.031 | 0.031 | 0.000 |
|  | 0.062 | 0.093 | 0.000 |

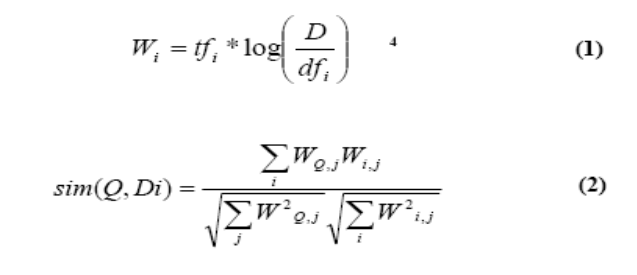
Menghitung √∑Wij 2 \* √∑Wiq 2

• Nilai Wdt Dipangkatkan 2 • Hitung jumlah(sigma) masing2 Di dan Kk • Mengakar pangkat 2 Hasil Sigma Di dan Kk



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Term** | **Wdt = tf\*Idf** | | | | **Panjang Vector** | | | |
| **KK** | **D1** | **D2** | **D3** | **KK** | **D1** | **D2** | **D3** |
| Demo | 0.000 | 0.000 | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 | 0.228 |
| Informatika | 0.000 | 0.000 | 0.176 | 0.176 | 0.000 | 0.000 | 0.031 | 0.031 |
| Laboraturium | 0.000 | 0.000 | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 | 0.228 |
| Mahasiswa | 0.000 | 0.000 | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 | 0.228 |
| Milad | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 | 0.228 | 0.000 | 0.000 |
| Tahun | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 | 0.228 | 0.000 | 0.000 |
| Terakreditasi | 0.000 | 0.000 | 0.477 | 0.000 | 0.000 | 0.000 | 0.228 | 0.000 |
| Trunojoyo | 0.176 | 0.176 | 0.352 | 0.000 | 0.031 | 0.031 | 0.124 | 0.000 |
| Universitas | 0.176 | 0.176 | 0.176 | 0.000 | 0.031 | 0.031 | 0.031 | 0.000 |
|  | | | | | 0.062 | 0.517 | 0.414 | 0.714 |
| 0.249 | 0.719 | 0.643 | 0.845 |

Rumus TF-IDF dan Cossim



Mendapatkan nilai CosSim()

• Cossim(Q,D1) = 0,062 / (0,249\*0,719) = 0,3462 • Cossim(Q,D2) = 0,093 / (0,249\*0,643) = 0,5808 • Cossim(Q,D3) = 0,000 / (0,249\*0,845) = 0

Menentukan Kesamaan

• Cossim(Q,D2) = 0,5808 • Cossim(Q,D1) = 0,3462 • Cossim(Q,D3) = 0

• Dari nilai diatas yang dianggap mirip adalah yg memiliki nilai cossim mendekati nilai 1 (dalam hal contoh soal adalah D2)