

WEBSCRAPPING AND WORDCLOUD

Berikut adalah dokumentasi untuk file 'WebScrapping_Wordcloud.ipynb' dengan tujuan:

- Melakukan scraping data berita dari situs kompas.com pada berita dari kategori terpopuler
- Analisis berita isi dari berita tersebut
- Tampilkan dalam bentuk wordcloud.

Alamat berita: <https://lifestyle.kompas.com/read/2020/10/16/203104920/hai-perempuan-jangan-lepaskan-pria-dengan-9-karakter-ini?page=all#page2>

Langkah 1: Import semua library yang dibutuhkan

Code:

```
from urllib.request import urlopen, Request
from bs4 import BeautifulSoup as bs
import requests
import re
from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize, word_tokenize
from wordcloud import WordCloud
import matplotlib.pyplot as plt
```

Langkah 2: Ambil Data (Web Scrapping)

Code:

```
#pengambilan data
alamat = "https://lifestyle.kompas.com/read/2020/10/16/203104920/hai-perempuan-jangan-lepaskan-pria-dengan-9-karakter-ini?page=all#page2"
req = Request(alamat, headers=headers)
html = urlopen(req)
data = bs(html, 'html.parser')
artikel = data.find("div", {"class": "read_content"})
#menghapus iklan yang ditandai dengan tag strong
hidelabel = artikel.findAll('strong')
for delete in hidelabel:
    delete.decompose()
items = artikel.findAll('p')
hasil = [item.get_text() for item in items]
```

Dilakukan pengambilan data artikel dan penghapusan bagian iklan. Sehingga hasil data yang diperoleh hanya merupakan artikel tanpa noise.



ANALISA:

Berdasarkan hasil visualisasi dari wordcloud dapat dilihat bahwa berita tersebut banyak membahas tentang kecocokan pasangan laki-laki dan perempuan, yang menyebutkan beberapa jenis karakter seperti humor, setia, independent, mendengarkan, dan cerdas. Maka dapat disimpulkan bahwa artikel ini membahas mengenai kriteria suatu pasangan untuk menemukan kecocokan dan menikah.