

Source belajar:

- [Github](#)
- [Medium](#)
- [Youtube-1](#)
- [Youtube-2](#)

YAKE merupakan keyword extractor, yake lebih ringan dibanding rake. Parameter YAKE lebih sedikit dibanding RAKE.

```
main.py > ...
1  import yake
2
3  text = "\n\nAdi Haryanto, Koran SI \u00b7 Senin 04 April 2022 10:32 WIB\nBANDUNG BARAT - Dua bal
4
5
6  language = "id"
7  max_ngram_size = 5
8  deduplication_threshold = 1
9  deduplication_algo = 'leve'
10 windowSize = 5
11 numOfKeywords = 5
12
13 custom_kw_extractor = yake.KeywordExtractor(lan=language, n=max_ngram_size, deduplim=deduplicati
14 keywords = custom_kw_extractor.extract_keywords(text)
15
16 for kw in keywords:
17     print(kw)
```

macos@Users-MacBook /Users/macros/Documents/247/yake

```
➤ > python3 main.py
('kepala desa cibodas dindin sukaya', 0.006287263434858747)
('kepala desa cibodas dindin', 0.014110354193119866)
('desa cibodas dindin sukaya', 0.014110354193119868)
('kabupaten bandung barat', 0.014112023074073693)
('kata kepala desa cibodas dindin', 0.014198953438001764)
```

macos@Users-MacBook /Users/macros/Documents/247/yake

Ln 7, Col 19 Spaces: 4

```
class KeywordExtractor(
    lan: str = "en",
    n: int = 3,
    dedupLim: float = 0.9,
    dedupFunc: str = 'seqm',
    windowsSize: int = 1,
    top: int = 20,
    features: Any | None = None,
    stopwords: Any | None = None
)
```

ake.KeywordExtractor(lan=language, n=ma

Paramater Rake terdiri dari:

- lan (language),

- n (maximal panjang kata),
- deduplim (threshold untuk duplikat),
- dedubFun (algo untuk duplikat, ada 'scm', 'leve', dan 'jargo'),
 - Sequence Matching (sqm): Algoritma ini bekerja dengan membandingkan urutan token (kata-kata) dalam kata kunci. Ini menggunakan pendekatan matching substring untuk menentukan sejauh mana suatu kata kunci mirip dengan kata kunci lainnya.
 - Levenshtein (leve): Mengukur jarak atau perubahan minimum yang diperlukan untuk mengubah satu string menjadi string lainnya. Semakin kecil jarak Levenshtein, semakin mirip kedua string tersebut.
 - Jaro-Winkler (jargo): Mengukur kemiripan antara dua string dengan memperhitungkan kesamaan karakter dan posisi karakter yang cocok. Jaro-Winkler memberikan bobot lebih tinggi pada awalan yang sama antara dua string.
- windowsSize (seberapa besar konteks yang dipertimbangkan oleh algoritma penghapusan kata-kata duplikat.),
- top (top keyword yang dihasilkan),
- stopwords(kata yang tidak bermakna)

Berbeda dengan RAKE, pada YAKE scoring nya semakin kecil maka semakin relevan