Laporan Praktikum
Menghitung Frekuensi Kata Menggunakan Python

KMTI21132 - Information Retrieval Dosen:
Salman El Farisi, S.Kom., M.Kom.
Asdos: Elyas Randi Renaldi (0110223277)

Nama	:	Al Hijir
NIM	:	0110224222
Program Studi	:	Teknik Informatika
Rombel	:	TI01
URL Github Tugas	:	https://github.com/Alhijir/Infromation-Retrieval

Jawablah pertanyaan dibawah ini berdasarkan hasil praktikum yang sudah dilakukan:

1. Tuliskan dalam bentuk tabel 10 kata dengan frekuensi tertinggi. Urutkan mulai dari kata dengan frekuensi tertinggi.
2. Tuliskan dalam bentuk tabel 10 kata dengan frekuensi terendah. Urutkan mulai dari kata dengan frekuensi terendah.
3. Tampilkan gambar distribusi frekuensi kata pada korpus “kompas-beritaPart1.xml”
4. Apakah gambar distribusi frekuensi pada korpus “kompas-beritaPart1.xml” mengikuti prinsip distribusi zipf?
5. Tuliskan 10 kata yang menurut kamu janggal / aneh. Menurut kamu, apa penyebabnya?

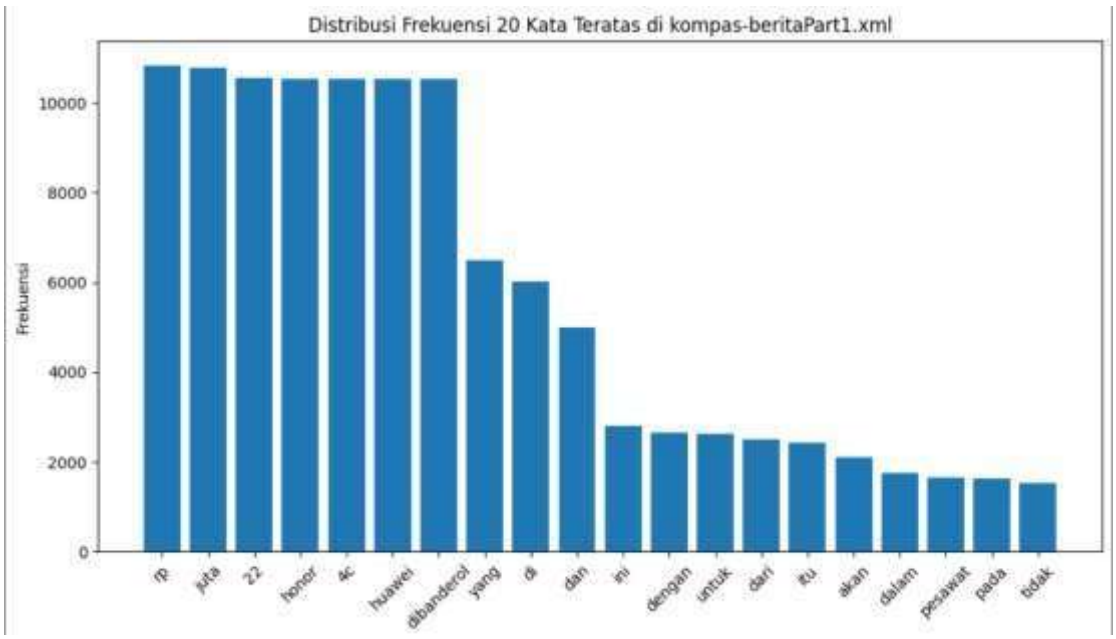
1. 10 Kata Dengan Frekuensi Tertinggi

No	Kata	Freq	Rank
1	Yang	6450	1
2	Di	5824	2
3	Dan	4915	3
4		4264	4
5	Ini	2744	5
6	Dengan	2625	6
7	Untuk	2583	7
8	Dari	2469	8
9	Itu	2437	9
10	Akan	2055	10

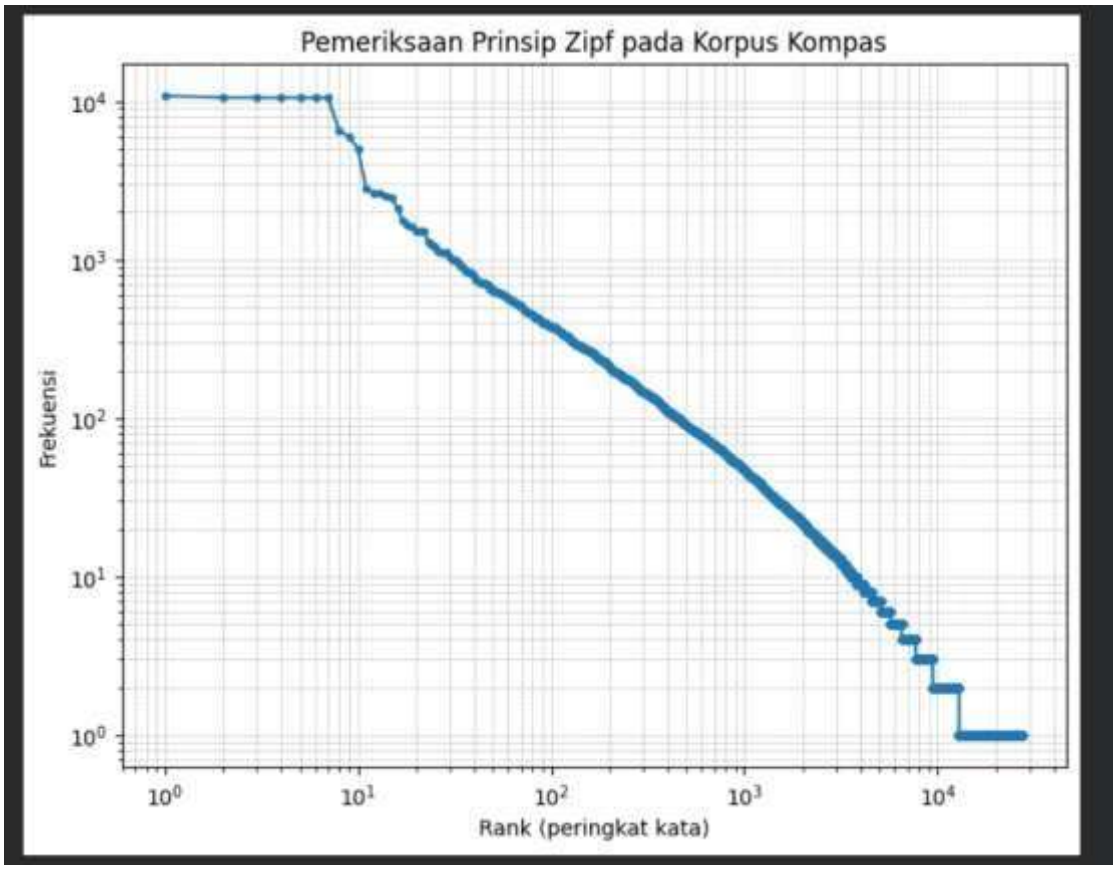
2. 10 Kata Dengan Frekuensi Terendah

No	Kata	Freq	Rank
1	xml	1	1
2	version10	1	2
3	jutakami	1	3
4	department	1	4
5	jakartaponsel	1	5
6	Kirin	1	6
7	Pembuatanya	1	7
8	hisilicon	1	8
9	jutajudulisijakarta	1	9
10	data_beritaberitasumberkompascomsumbertanggal20150701tanggal kategoriteknologikategorijudulponsel	1	10

3. gambar distribusi frekuensi kata pada korpus “kompas-beritaPart1.xml”



4. frekuensi pada korpus "kompas-beritaPart1.xml" mengikuti prinsip distribusi zipf, karena grafik **log–log** di bawah terlihat **garis lurus miring ke bawah**, berarti korpus kamu **mengikuti prinsip Zipf**.



5. **10 kata janggal / aneh**

No	Kata Janggal	Dugaan Penyebab
1	nbsp	sisa kode HTML untuk spasi non-breaking ( )
2	br	tag HTML yang tidak dibersihkan
3	lt / gt	karakter < atau > yang dikonversi jadi <, >
4	xa0	karakter spasi aneh dari encoding UTF-8
5	said	campuran teks berbahasa Inggris dalam artikel Indonesia
6	kompascom	sisa domain dari watermark atau atribusi sumber berita
7	foto	berasal dari caption gambar, bukan isi berita utama
8	detik	muncul jika berita mengutip sumber lain (tidak relevan konteks)
9	jakartacom	hasil parsing URL atau teks web yang tidak difilter
10	— (strip panjang)	simbol tanda baca yang tidak difilter dengan regex biasa