**Chapter 14 : Counting Words**

Dijelaskan bahwa sejahu pemahaman yang telah didapatkan dari bab 2 sampai 7. Menegaskan bahasa memiliki struktur yang luar biasa, sebagian terlihat, dan sebagian tersembunyi, dan masih banyak yang harus ditemukan. Dijelaskan oleh penulis bahwa sampai dengan buku ini diterbitkan penelitian yang dilakukan adalah qualitative dimana merujuk kepada pemaknaan yang muncul dari fonem atau grafem(Tag bagian dari ucapan, milik konstituen, dll., adalah bagian dari informasi kualitatif yang dapat kita kumpulkan pada morfem, kata, atau kelompok kata,) dilakukan dengan melakukan tanya jawab.

Terhadap pendekatan kuantitatif, dilakukan dalam tahapan dan langkah yang berbeda. menyimpulkan dengan melampirkan informasi numerik ke unit bahasa. Lalu diberikan beberapa contoh proses perhitungan informasi numerik daei unit bahasa dari saman ahli ahli taurat Yahudi yang disebut Masoret sampai kepada Matematikawan Rusia (1913) A Makrov mengambil 20.000 grafem dari sebuah novel. Selanjutnya penulis berkesimpulan bahwa komputer pun dapat melakukan tugas menghitung dengan baik. Sehingga n inilah sebabnya pendekatan statistik terhadap bahasa mengalami peningkatan luar biasa pada masa-masa awal komputasi dan peningkatan besar kedua pada tahun sembilan puluhan ketika komputer pribadi dan peningkatan daya pemrosesan membuat metode statistik yang rumit dapat diakses oleh setiap orang.

1. **Tokenization dan Segmentasi**

Tokenisasi adalah proses memecah teks menjadi unit-unit yang lebih kecil yang disebut token, yang bisa berupa kata, tanda baca, atau entitas lain tergantung konteks dan bahasa. Konsep “kata” sendiri sulit didefinisikan secara universal, sehingga pendekatan statistik menggunakan tokenisasi untuk menghindari definisi baku dan membuatnya bergantung pada bahasa dan tugas tertentu. Dalam buku disajikan contoh tokenisasi dan pengunaannya dalam Python. Misalkan dengan menggunakan *split() , nltk.word\_tokenize,* juga dengan *spacy.* Untuk bahasa tanpa spasi antar kata misalkan seperti cina dijelaskan dalam buku bahwa menghadirkan tantanagn tersendiri. Sehingg segmentasi menjadi sanggat penting. Segmentasi sendiri merupakan proses mendeteksi batas antar token dalam teks. Segmentasi adalah bagian dari proses tokenisasi.

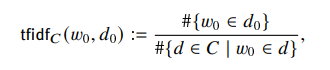
1. **Aturan Zipf’s**

Hukum Zipfs menyatakan bahwa dalam korpus teks yang cukup besar, frekuensi sebuah kata berbanding terbailik dengan peringkatnya jika diurutkan dari yang paling sering digunakan. Artinya : Semakin tinggi frekuensi sebuah kata digunakan dalam bahasa, maka semakin rendah peringkat(rangkinG) KATANYA dalam daftar frekuensi dan sebalikya. Pada buku diberikan contoh percobaan pada korpus whovian . dari gambar yang ditampilkan kurva mendekati antidiagonal menunjukan kecocokan parsial dengan hukum Zipf’s . Zipf’s Law bekerja pada data yang sudah diurutkan dari frekuensi terbesar ke terkecil. Rank adalah urutan dari kata yang paling sering muncul (rank 1) ke paling jarang (rank n). Baru setelahnya diuji apakah distribusinya mendekati 

1. **Stop Words dan tfidf**

Menurut penjelasan, Stop words adalah kata kata umum yang sering muncul dalam teks tapi kurang memberikan makna kontekstual : *the, is, and, to, of*. Dalam NLP, stop words sering dihapus dari analisis karena dianggap tidak membantu dalam membedakan dokumen atau konteks. Tidak semua stop words harus dihapus. Dalam beberapa konteks atau tugas, kata-kata ini justru penting secara semantik. Contoh : kalimat terkenal "To be or not to be" dari Hamlet hanya terdiri dari stop words—menghapusnya bisa menghilangkan makna penting.

tf-idf digunakan untuk menilai seberapa penting suatu kata dalam dokumen relatif terhadap seluruh korpus. memberi bobot pada kata berdasarkan seberapa penting kata itu dalam dokumen tertentu.

yaitu Jumlah kemunculan w di d dan Jumlah dokumen yang mengandung w.

tf (term frequency): seberapa sering kata muncul dalam dokumen tertentu. Sedangkan idf (inverse document frequency): semakin banyak dokumen yang mengandung kata tersebut, semakin rendah nilai idf-nya.

Pada buku diberikan contoh kasus episode ”Heaven Sent” dari Doctor Who. Didapatkan bahwa Kata dengan nilai tf-idf tertinggi adalah : clara, hybrid, shepherd . dan untuk nilai terendah adalah er, wait, sure, good. Kelemahan metode ini adalah tidak dapat membedakan apakah sebuah kata muncul 1 atau 1000 kali dalam dokumen lain. Yakni hanya memperhatikan Jumlah dokumen bukan frekuensi total dari dokumen tersebut.

1. **Collocations**

Adalah pasangan

1. **N-Grams**
2. **Vektor Semantik**
3. **LSA dan ESA**
4. **Skip-gram Embeddings**
5. **Teori**
6. **Analogi Kata**
7. **Visualisasi**
8. **FastText**
9. **Further Reading**
10. **Latihan**