Text Pre-processing

About Module

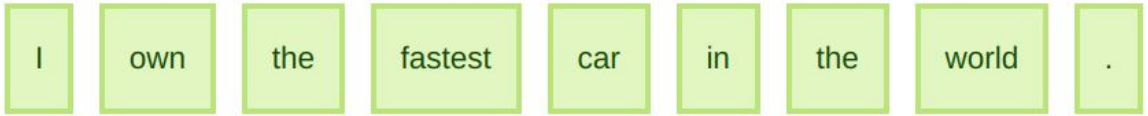
1. Terminologies
2. Tokenization
3. Stopword Removal
4. Normalization
5. Data Exploration

Analytics Vidhya

Terminologies

- Token: Smallest unit of a text.
- Sentence: A sequence of tokens arranged grammatically. Eg,

I own the fastest car in the world.



Terminologies

- Paragraph: A collection of sentences.

> I own the fastest car in the world. It is the Bugatti Chiron Super Sport 300+, and I bought it yesterday for $3.9 million. I keep it in my garage with my other cars. I want to create a collection of supercars. Currently, I own a Rolls Royce, a Koenigsegg Agera, and few Lamborghinis. It's fun to has these cars. I mostly drive my cars on Friday night because at night time they look fantastic.

Terminologies

- Document: A collection of paragraphs.

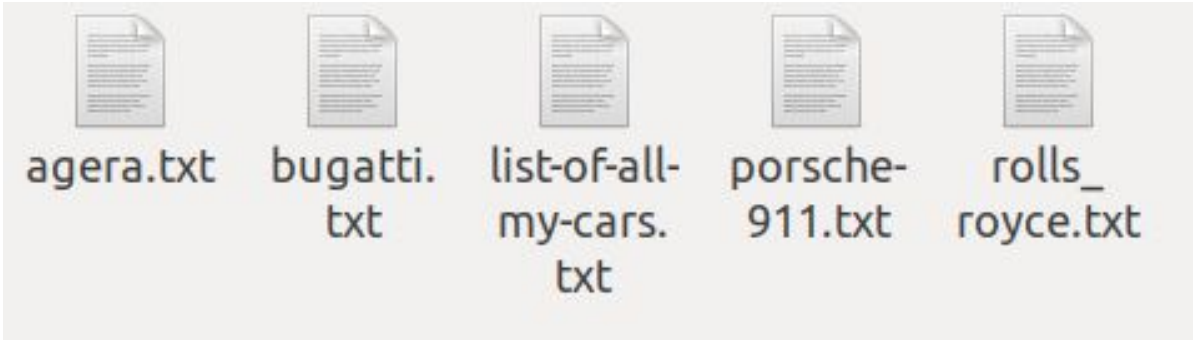
I own the fastest car in the world. It is the Bugatti Chiron Super Sport 300+, and I bought it yesterday for $3.9 million. I keep it in my garage with my other cars. I want to create a collection of supercars. Currently, I own a Rolls Royce, a Koenigsegg Agera, and few Lamborghinis. It's fun to has these cars. I mostly drive my cars on Friday night because at night time they look fantastic.

The Bugatti Chiron Super Sport 300+ can accelerate from 0-100 kilometres per hour (0-60 mph) in 2.4 seconds. As far as its top speed goes, in 2019, a Bugatti test driver was able to achieve a speed of 490.48 kilometres per hour (304.77 mph) under controlled conditions. That's about covering 450 feet in a single second.

It has an incredible 7,993 cc (8.0 L) quad-turbocharged W16 engine which produces about 1600 bhp. There are only 30 beasts like this present in the world, and I have just bought of them.

Terminologies

- Corpus: A collection of text documents.



Analytics Vidhya

Terminologies

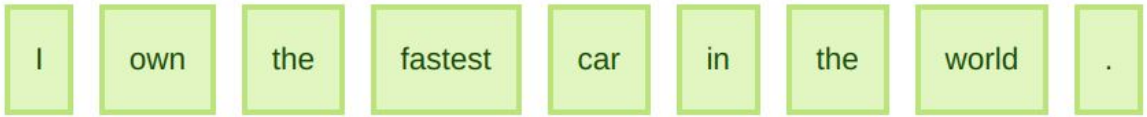
Corpus > Document > Paragraph > Sentence > Token

Analytics Vidhya

Terminologies

- Vocabulary: Set of unique words.

I own the fastest car in the world.



No. of tokens(N) = 9

Vocabulary (V) = { I, own, the, fastest, car, in, world}

Size of Vocabulary = □V□ = 7

Analytics Vidhya

Thank You

Analytics
Vidhya