Metni Biçimlendirme (Tekonize Etme) ve Cümleler İçin Diziler Oluşturma

Bu colab dosyasında, metni TensorFlow modelleriyle kullanım için hazırlamanın ilk aşaması olarak metnin nasıl tekonize edeceğimizi ve cümleler için diziler oluşturacağımıza bakacağız.

İçeri Aktarma İşlemi: `Tekonizer`

In [1]:

```
from tensorflow.keras.preprocessing.text import Tokenizer
```

Bazı Cümleler Yazalım ¶

In [2]:

```
sentences = [
    'My favorite food is ice cream',
    'do you like ice cream too?',
    'My dog likes ice cream!',
    "your favorite flavor of icecream is chocolate",
    "chocolate isn't good for dogs",
    "your dog, your cat, and your parrot prefer broccoli"
]
```

Kelimelerin Simgelenmesi (Tekonize Edilmesi)

Bir makine öğrenimi modelinde kullanılacak metni hazırlamanın ilk adımı, metni tekonize etmek diğer bir deyişle kelimeler için sayılar üretmektir.

In [3]:

```
# İsteğe bağlı olarak belirtilecek maksimum kelime sayısını ayarlayın.
# (OOV) belirteci, sözcük dağarcığı dizinde olmayan sözcükleri temsil eder.
# Her kelime için benzersiz sayılar oluşturmak için belirteç üzerinde fit_on_text() öğesini
tokenizer = Tokenizer(num_words = 100, oov_token="<OOV>")
tokenizer.fit_on_texts(sentences)
```

Kelime Dizinin Görüntülenmesi

Metni simgeleştirdiktem sonra belirteç tüm sözcükler ve sayılar için anahtar/değer çiftlerini içeren sözcük dizinin sahip olur.

Kelimler anahtardır ve sayılar ise değerleri temsil eder.

`OOV` belirtecinin ilk giriş olduğuna dikkat edin.

In [4]:

```
# Kelime dizinini inceleyin
word_index = tokenizer.word_index
print(word_index)
```

```
{'<OOV>': 1, 'your': 2, 'ice': 3, 'cream': 4, 'my': 5, 'favorite': 6, 'is':
7, 'dog': 8, 'chocolate': 9, 'food': 10, 'do': 11, 'you': 12, 'like': 13, 't
oo': 14, 'likes': 15, 'flavor': 16, 'of': 17, 'icecream': 18, "isn't": 19,
'good': 20, 'for': 21, 'dogs': 22, 'cat': 23, 'and': 24, 'parrot': 25, 'pref
er': 26, 'broccoli': 27}
```

In [5]:

```
# Belirli bir kelimenin numarasını alın
print(word_index['favorite'])
```

```
6
```

Cümleler İçin Dizilerin Oluşturulması

Sözcükleri simgeleştirdikten sonra, sözcük dizini her sözcük için benzersiz bir sayı içerir. Ancak, kelime dizinindeki sayılar sıralı değildir. Bir cümledeki kelimelerin bir sırası vardır. Bu nedenle, kelimeleri tokenize ettikten sonraki adım, cümleler için diziler oluşturmaktır.

In [6]:

```
sequences = tokenizer.texts_to_sequences(sentences)
print (sequences)
```

```
[[5, 6, 10, 7, 3, 4], [11, 12, 13, 3, 4, 14], [5, 8, 15, 3, 4], [2, 6, 16, 1
7, 18, 7, 9], [9, 19, 20, 21, 22], [2, 8, 2, 23, 24, 2, 25, 26, 27]]
```

Kelime Dizininde Olmayan Kelimeleri İçeren Sıralı Cümlelerin Görüntülenmesi

Sıralanan cümle, kelime dizininde olmayan kelimeler içeriyorsa ne olduğuna bir bakalım.

The Out of Vocabluary (OOV) belirteci, kelime dizinindeki ilk giriştir. Kelime dizininde olmayan herhangi bir kelimenin yerine dizilerde göründüğünü göreceksiniz.

In [7]:

```
sentences2 = ["I like hot chocolate", "My dogs and my hedgehog like kibble but my squirrel

sequences2 = tokenizer.texts_to_sequences(sentences2)
print(sequences2)
```

```
[[1, 13, 1, 9], [5, 22, 24, 5, 1, 13, 1, 1, 5, 1, 1, 1, 24, 5, 1, 13, 3, 4,
1, 1]]
```