Converting Documents to Vectors

Computers store words as a series of characters with each character represented as a number. This representation from number to character or character to number is called **character encoding**. The most popular character encoding is the **ASCII** character encoding.

A	65
...	...
Z	90

a	97
...	...
z	122

ASCII A-Z and a-z

In ASCII, an uppercase "A" is represented as 65 and a uppercase "Z" is represented as 90. All other uppercase english characters are represented between 65 and 90. Lowercase letters are represented in the range of 97 to 122.

Just like a computer, a neural network can't understand words. It only sees a series of numbers. Feeding a neural network these series of numbers would require it to learn the english vocabulary. This requires more data than we have. Instead, let's turn the words into something easer for a neural network to learn from.

Just like ASCII encodes a character as a number, you can encode a word as a number. For example, the text `"my bat is the best bat"` could be changed to `[1, 2, 3, 4, 5, 2]`. Each word has been converted to a number where the number assigned to a word is arbitrary.

It's only required that each number corresponds to a single word and vice versa. This is known as a one to one function. In our case, this mapping is:

- `1 <-> my`
- `2 <-> bat`
- etc.

Quiz

Let's see if you can apply your understanding of converting sentences to vectors. In the following quiz, select a valid array that could represent the following string.

fox jumped over the lazy dog"?

- [ ] [1, 2, 3, 4, 5, 6, 7, 8, 9]
- [ ] [99, 45, 26, 38, 11, 10, 99, 2, 17]
- [ ] [1, 2, 3, 4, 5, 6, 1, 7, 8]

SUBMIT

NEXT