Natural Language Processing

The Jubilee Institute, January 2023

Natural Language

a language that is developed naturally by humans

although there are rules to how words are used, they are not followed 100% of the time

very wide vocabulary, with some words having many different meanings

Programming Language

a language that is used to give instructions to a computer

very strict rules about how phrases need to be written

words do not have multiple definitions, they are used to define very specific functions or methods

Natural Language Processing

a subfield of AI and linguistics (the study of language), which looks at how we can train computers to understand human language

My son and his friend like to play with the **leaves** under the tree.

He hates it when his friend **leaves** to go back home.

I took my kids to the **park** to play.

I had to **park** my car on the street.

You were **right**. I should have turned **right** five minutes ago.

Please **watch** your little sister. I need to get my **watch** fixed at the jewelry store.

Buffalo buffalo Buffalo buffalo buffalo buffalo buffalo Buffalo buffalo.

Training Data

How can we convert this training data into numbers?

Bag of Words

a way to represent a sentence using numbers by identifying which of the all available words are used

all_words = [“hi”, “how”, “are”, “you”, “bye”, “see”, “later”]

all_words = [“hi”, “how”, “are”, “you”, “bye”, “see”, “later”]

“how are you” → **greeting**

all_words = [“hi”, “how”, “are”, “you”, “bye”, “see”, “later”]

“how are you” → [0, 0, 0, 0, 0, 0, 0]

all_words = [“hi”, “how”, “are”, “you”, “bye”, “see”, “later”]

“how are you” → [0, 1, 1, 1, 0, 0, 0]

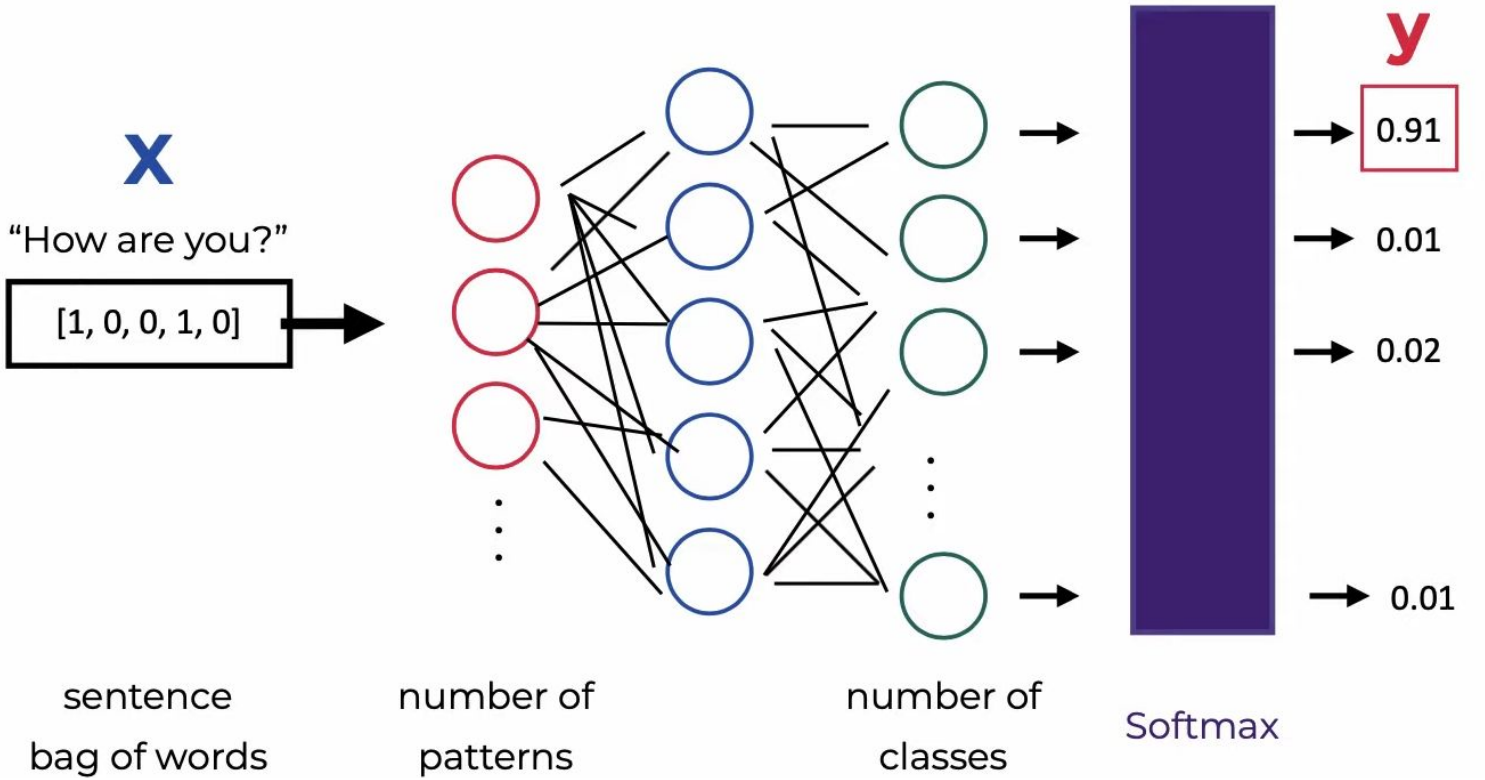
all_words = [“hi”, “how”, “are”, “you”, “bye”, “see”, “later”]

“how are you” → [0, 1, 1, 1, 0, 0, 0] → **greeting**

all_words = [“hi”, “how”, “are”, “you”, “bye”, “see”, “later”]

“how are you” → [0, 1, 1, 1, 0, 0, 0] → **greeting**

“see you later” → [0, 0, 0, 1, 0, 1, 1] → **goodbye**


x
"How are you?"
[1, 0, 0, 1, 0]
y
0.91
0.01
0.02
0.01
sentence
bag of words
number of
patterns
number of
classes
Softmax

Tokenization

splitting a sentence into meaningful units so we can process it

"what would you do with 1000000$?"

→ [“what”, “would”, “you”, “do”, “with”, “1000000”, “$”, “?”]

“aren’t you happy with so much money?”

→ [“are”, “n’t”, “you”, “happy”, “with”, “so”, “much”, “money”, “?”]

Stemming

converting words into their root forms, to make it easier for the computer to process

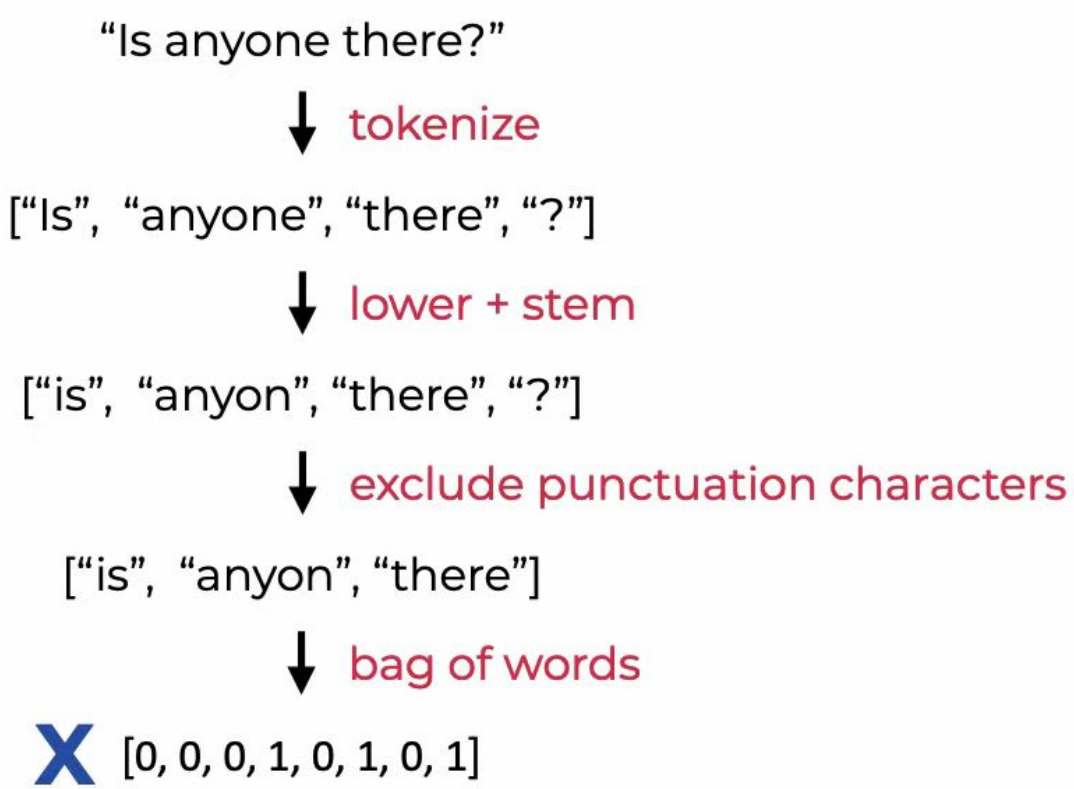
“organize”, “organizes”, “organizing”

“organize”, “organizes”, “organizing”

→ [“organ”, “organ”, “organ”]

[“fit”, “fitting”] → [“fit”, “fit”]

Since he played football he was very **fit**.
The clothes were not **fitting** him very well.


"Is anyone there?"
tokenize
["Is", "anyone", "there", "?"]
lower + stem
["is", "anyon", "there", "?"]
exclude punctuation characters
["is", "anyon", "there"]
bag of words
[0, 0, 0, 1, 0, 1, 0, 1]

tinyurl.com/JubileeGit

lesson4

intents_template.json

14navpreetsingh@gmail.com

Names please in subject line of email.

Come up to test your bot!