Natural Language Processing

Final Presentation

LSTM model

StackOverflow

Contents

Implementation

1. Loading the data
2. Preprocessing the data
    - Decoding HTML entities (&lt, &gt…)
    - Code patterns are identified like markdown, inline code between single backticks are replaced with CODETOKEN
    - Manage URLs - URLTOKEN
    - Stripping remaining HTML tags

Implementation

2. Preprocessing the data (cont…)
    - Whitespace cleaning and lowercasing
    - Token Pattern identification where Regex is used to distinguish questions and set sentence boundaries

String: "C++ code, URLTOKEN, CODETOKEN”

It becomes → "c++" "code" "," "urltoken" "," "codetoken"]

Implementation

3. Building the vocabulary which involves converting each token to its index
    - Counting token frequencies
    - Initializing special tokens (PAD & UNK)

This helps to map discrete token IDs into token vectors

Implementation

4. Building the vocabulary which involves converting each token to its index
    - Counting token frequencies
    - Initializing special tokens (PAD & UNK)

This helps to map discrete token IDs into token vectors.

5. Tokens are paired into batches with padding and lengths

Implementation

6. Attention layer
    - Handles variable-length sequences whilst ignoring padding and unknowns
    - Computes attention logits by assigning an importance score to each token

DEMO

BiLSTM Model

NEXT STEPS

Build a simple base model

Compare the results (F1 and Accuracy)

Publish Results in the paper

THANK YOU!