ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÀI TẬP
MÔN CÁC KỸ THUẬT HỌC SÂU VÀ ỨNG DỤNG

KHOA: KHOA HỌC MÁY TÍNH

HOMEWORK: TEXT GENERATION WITH RNN FOR VIETNAMESE

Giảng viên: Nguyễn Duy Khánh

Nhóm thực hiện:

1. Trương Văn Khải– 21520274

Đề: Làm Text Generation với RNN cho dữ liệu Tiếng Việt tự thu thập, tối thiểu 1000 ký tự

Bài làm:

- **RNN cho cấp độ âm tiết (characters)**

Souce: Tạo văn bản bằng RNN | Text | TensorFlow

1. **Dataset:**
   - Lời nhạc 3 bài hát của ca sĩ Đạt G: Thêm Bao Nhiêu Lâu + Khó Vẽ Nụ Cười + Bánh Mì Không
   - Tiền xử lý dataset:
     - Độ dài: 6354 characters
     - Vocabulary size: 107 unique characters
2. **Mô tả kiến trúc Model:**
   - Embeding Input: 107 embedding vectros, mỗi vector có chiều = embedding_dim = 256
   - Rnn_units = 1024
   - Output: Dense layer : 97
3. **Huấn luyện:**
   - batch_size = 2 (Với dữ liệu ít, để batch_size cao quá mô hình dự đoán ra rất tệ)
   - EPOCHS = 50
   - Optimizer = 'adam'
   - loss = SparseCategoricalCrossentropy

4. Kết quả sinh text:

Số lượng từ cần sinh là: 500

- Trường hợp 1: Từ bắt đầu là **“á”**
- Kết quả: **ánh mắt anh buồn Hoạ nụ cười thêm trên nét môi Nhưng sao chẳng thương có thể em sống cho em Quá dễ để biết tên nhau nhưng đâu có thể nhìn thấy nhau đau ồi hiểu thấu Vẫn hoạ thêm chiếc môi cười tiếp theo Nhưng đau thấu trời Khóc thật nhiều Ngồi khóc thật nhiều Khóc cho đời phong ba lắm đau mà chông thế nàa. … Anh mơ một mai thức giấc chẳng cơn đau nào sẽ ghé qua Em mong ngày mai có nắng sưởi ấm con tim buồn đau tối qua Anh mong dù cho vấp ngã có lắm phong ba bình yên cho em nhưng cũng có thể em số**

- Trường hợp 2: Từ bắt đầu là **“.”**
- Kết quả: **. D em xem mì côi đưng cho em Quá dễ để biết tên nhau nhưng đâu có thể nhìn thấy nhau đau Chiều hoàng hôn buông xuống phía tây nghẹn ngào Uống phai oh Nỗi đau dày vò nỗi đau trời cao có thấu, có biết cơn đau mùi hương thế nào. … Biết đâu những cơn say đầu đời này Phải khóc cho đến hôm nay Hẹn ước chi thế em ơi giờ phải vậy Gửi gió nỗi**

nhớ bay bay Nhìn hoàng hôn kia còn mang em đi xa anh thêm bao nhiêu lâu nữa sẽ trả về Đừng có như vậy, dặn lòng đừng khóc như vậy Mà đời đâu như anh mơ, đâu như anh

- **RNN cho cấp độ từ**

1. Dataset

- Lời nhạc 6 bài hát của ca sĩ Đạt G: Thêm Bao Nhiêu Lâu + Khó Vẽ Nụ Cười + Bánh Mì Không Ngày Mai Em Đi Mất + Còn Buồn Không Em + Anh Tự Do Nhưng Cô Đơn
- Tiền xử lý dataset:
    - Độ dài: 493 từ (11534 characters)
    - Vocabulary size: 493 unique characters

2. Mô tả kiến trúc Model

- Embeding Input: 493 embedding vectros, mỗi vector có chiều = embedding_dim = 256
- Một lớp 1 Bidirectional với cell LSTM 150 node
- Một lớp LSTM với 100 node
- Output: Dense layer : 493

3. **Huấn luyện**

- EPOCHS = 150
- Optimizer = 'adam'
- loss = categorical_crossentropy

4. **Kết quả sinh text:**

- Số lượng từ cần sinh là: 200
- Từ bắt đầu “em ổn không”
- Kết quả: **em ổn không hiểu thấu em như nắng mai trong đời cho em đau ai ngờ để phải nhớ có có khóc xôi ai anh thêm bao nhiêu lâu anh có thêm rơi sẽ hẹn rong chơi và côi rồi nhạc vang lên xe tranh và côi lắm cứ rơi nhưng có yêu ai sao sao đêm nay trong xa là ai anh sao anh khóc cho qua thấu đừng khóc có có thấu đau thế ai ai cười lần vương em có cơn khóc có có xưa còn nói nhớ có vào ùa về và cứ rơi anh tối chẳng khi tố như mưa rơi đâu em hiểu anh dành hơn nửa thứ thứ đó em ở nhớ em có sẽ lánh ở phần anh ơi lại thể anh sẽ chờ em ơi lại cho anh biết vương và vấn khô rồi kêu vẫn mà vẫn phần cũng không biết bầu trời kia lại anh dành thứ chậm em sao anh cũng nhắm mắt em cũng khóc còn vào kiếm ăn rơi còn có cơn nhưng có nhau ăn xôi lắm này đời phiến thăng trầm ấy về rơi nhưng mưa rơi nếu thấu em mưa rơi em yêu như thế mà**


5. Predict the next N words.
[42] test_seq = 'em ổn không'
next_words = 200
for _ in range(next_words):
token_list = tokenizer.texts_to_sequences([test_seq])[0]
token_list = pad_sequences([token_list], maxlen=max_sequence_len-1, padding='pre')
predicted = model.predict(token_list, verbose=0)
output_word = ""
predicted_id = np.argmax(predicted)
if predicted_id in tokenizer.index_word:
output_word = tokenizer.index_word[predicted_id]
if output_word == '<end>':
break
test_seq += " " + output_word
else:
break
print(test_seq)
em ổn không hiểu thấu em như nắng mai trong đôi cho em đau ai ngờ để phải nhớ có có khóc xối ai anh thêm bao nhiêu lâu anh có thêm rơi sẽ hẹn rong chơi và cõi rồi nhạc vang lên xe tru