RNN을 활용한 단어완성 모델

소프트웨어신기술특론

산업경영공학과
201533258
장예훈

왜 RNN을 활용한 단어완성 모델?

1. 모델 주제

Wide & Deep Neural Network은 활용범위가 워낙 다양해서
모델의 주제를 탐색하는 데 시간이 오래 걸렸음

2. Data

데이터 전처리보다 모델을 만드는데 집중하고 싶었음

간단한 챗봇을 만들자!

왜 RNN을 활용한 단어완성 모델?

챗봇 모델 시작부터 에러 → 챗봇 포기

영화 시나리오(문장) 생성 모델로 도전 → 상당한 데이터 전처리 시간

문장생성 모델 말고 단어완성 모델로 전환

짧은 시 여러 개를 학습시켜 몇 글자가

주어지면 나머지 뒷 글자를 완성시키는 모델

Progress

Data Preprocessing

```
raw_poem = raw_poem.replace("₩n", " ").split(" ")
raw_poem
sentence = [i for i in raw_poem if i is not '']
len(sentence)
```

238

```
poem = " ".join(sentence)
poem
```

"바닷가에 왔드니 바다와 같이 당신이 생각만 나는구려 바다와 같이 당신을 사랑하고만 싶구려 구붓하고 모래톱을 오르면 당신이 앞선 것만 같구려 당신이 뒤선 것만 같구려 그리고 지중지중 물가를 거닐면 당신이 이야기를 하는 것만 같구려 당신이 이야기를 끊은 것만 같구려 바닷가는 개지꽃에 개지 아니 나오고 고기비눌에 하이얀 햇볕만 쇠리쇠리하야 어쩐지 쓸쓸만하구려 섧기만 하구려 밖은 봄철날 따디기의 누굿하니 푹석한 밤이다 거리에는 사람두 많이 나서 흥성흥성 할 것이다 어쩐지 이 사람들과 친하니 싸다니고 싶은 밤이다 그렇건만 나는 하이얀 자리 위에서 마른 팔뚝의 샛파란 핏대를 바라보며 나는 가난한 아버지를 가진 것과 내가 오래 그려오든 처녀가 시집을 간 것과 그렇게도 살틀하든 동무가 나를 버린 일을 생각한다 또 내가 아는 그 몸이 성하고 돈도 있는 사람들이 즐거이 술을 먹으려 다닐 것과 내 손에는 신간서 하나도 없는 것과 그리고 '아서라 세상사'라도 들을 유성기도 없는 것을 생각한다 그리고 이러한 생각이 내 눈가를 내 가슴가를 뜨겁게 하 는 것도 생각한다 산모퉁이를 돌아 논가 외딴 우물을 홀로 찾아가선 가만히 들여다봅니다. 우물 속에는 달이 밝고 구름이 흐르고 하늘이 펼치고 파아란 바람이 불고 가을이 있읍니다. 그리고 한 사나이가 있읍니다. 어쩐지 그 사나이가 미워져 돌아갑니다. 돌아가다 생각하니 그 사나이가 가엾어집니다. 도로 가 들여다보니 사나이는 그대로 있읍니다. 다시 그 사나이가 미워져 돌아갑니다. 돌아가다 생각하니 그 사나이가 그리워집니다. 우물 속에는 달이 밝고 구름이 흐르고 하늘이 펼치고 파아란 바람이 불고 가을이 있고 추억처럼 사나이가 있읍니다. 죽는 날까지 하늘을 우러러 한 점 부끄러움이 없기를 잎새에 이는 바람에도 나는 괴로워했다. 별을 노래하는 마음으로 모든 죽어 가는 것을 사랑해야지 그리고 나한테 주워진 길을 걸어가야겠다 오늘밤도 별이 바람에 스치운다"

```
char_list = []
for i in poem:
    if i != "₩n" and i not in char_list:
        char_list.append(i)
print(char_list[:100])
```

```
['바', '닷', '가', '에', ' ', '왔', '드', '니', '다', '와', '같', '이', '당', '신',
'생', '각', '만', '나', '는', '구', '려', '을', '사', '랑', '하', '고', '싶', '붓',
'모', '래', '톱', '오', '르', '면', '앞', '선', '것', '뒤', '그', '리', '지', '중',
'물', '를', '거', '닐', '야', '기', '끊', '은', '개', '꽃', '아', '비', '눌', '얀',
'햇', '볕', '쇠', '어', '쩐', '쓸', '섧', '밖', '봄', '철', '날', '따', '디', '의',
'누', '굿', '푹', '석', '한', '밤', '람', '두', '많', '서', '흥', '성', '할', '들',
'과', '친', '싸', '렇', '건', '자', '위', '마', '른', '팔', '뚝', '샛', '파', '란',
'핏', '대']
```

```
num_char = {c: i for i, c in enumerate(char_list)}
char_len = len(num_char)
seq_data = poem.split(" ")
```

Progress

단어들의 길이를 맞추기 위한 방법을 찾다가 padding을 먹이는 것 같은 (꼼수)

띄어쓰기를 기준으로 tokenize

```
seq_data = poem.split(" ")
seq_data[:10]
```

```
['바닷가에', '왔드니', '바다와', '같이', '당신이', '생각만', '나는구려', '바다와', '같이', '당신을']
```

token 중의 length가 max인 것을 찾고

그 길이에 맞게 공백 추가

```
max_len = 0
for i in seq_data:
    if len(i) > max_len:
        max_len = len(i)
print(max_len)
```

```
7
```

일종의 **padding**을 먹이는 것과 같은 꼼수 ㅎㅎ

```
for i in range(len(seq_data)):
    seq_data[i] = seq_data[i].ljust(max_len)
seq_data[:10]
```

```
['바닷가에   ',
 '왔드니    ',
 '바다와    ',
 '같이     ',
 '당신이    ',
 '생각만    ',
 '나는구려   ',
 '바다와    ',
 '같이     ',
 '당신을    ']
```

Prediction

```
pred = []
for i in zip([w[:2] + ' ' for w in seq_data], predict_words):
    pred.append(i)
pred[:10]
```

```
[('바닷 ', '바닷가에   '),
 ('왔드 ', '왔드니    '),
 ('바다 ', '바다와    '),
 ('같이 ', '같이     '),
 ('당신 ', '당신이    '),
 ('생각 ', '생각만    '),
 ('나는 ', '나는구려   '),
 ('바다 ', '바다와    '),
 ('같이 ', '같이     '),
 ('당신 ', '당신을    ')]
```

소감

엄청나게 많은 시간과 노력을 투자했지만

막상 결과물은 이런 단순한 미완성의 프로그램인 것을 보면서

~~학문에는 끝이 없으며 멀고도 험하고 역시 이 길은 내 길이 아니... 이게 아니고..~~

이론공부와 단순 예제 학습에서 그치지 않고 이렇게 직접 **데이터 처리**도 해보고 모델을 어떻게 더 **실생활에 접목**시킬 수 있는 지 고민해보는 것이 **실력 향상**에 더 도움이 된다는 것을 경험함!

THANK
Y O U

Reference

- 모두를 위한 딥러닝
- 골빈 해커의 3분 딥러닝 텐서플로맛 (https://github.com/golbin/TensorFlow-Tutorials)
- 백석 <바다>
- 백석 <내가 생각하는 것은>
- 윤동주 <자화상>
- 윤동주 <서시>