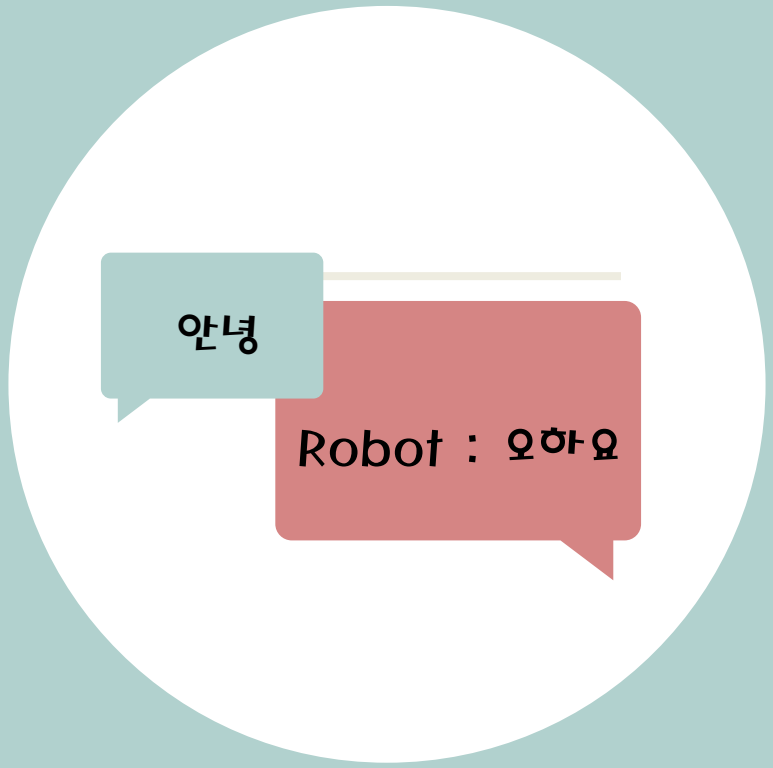


Chabot & Crawling

RNN을 이용한 데이터 분석

2017. 09. 15

CONTENTS

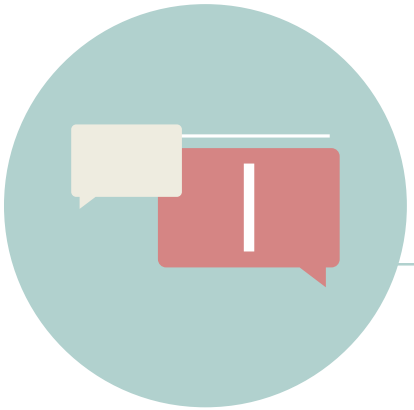
PTLINE
POWERPOINT
TEMPLATE

개요

AI 스피커

CHATBOT 기술 적용

RNN기법을 통한 구현



BIG DATA
데이터 기법 설명

1. Crawling

2. LSTM & GRU

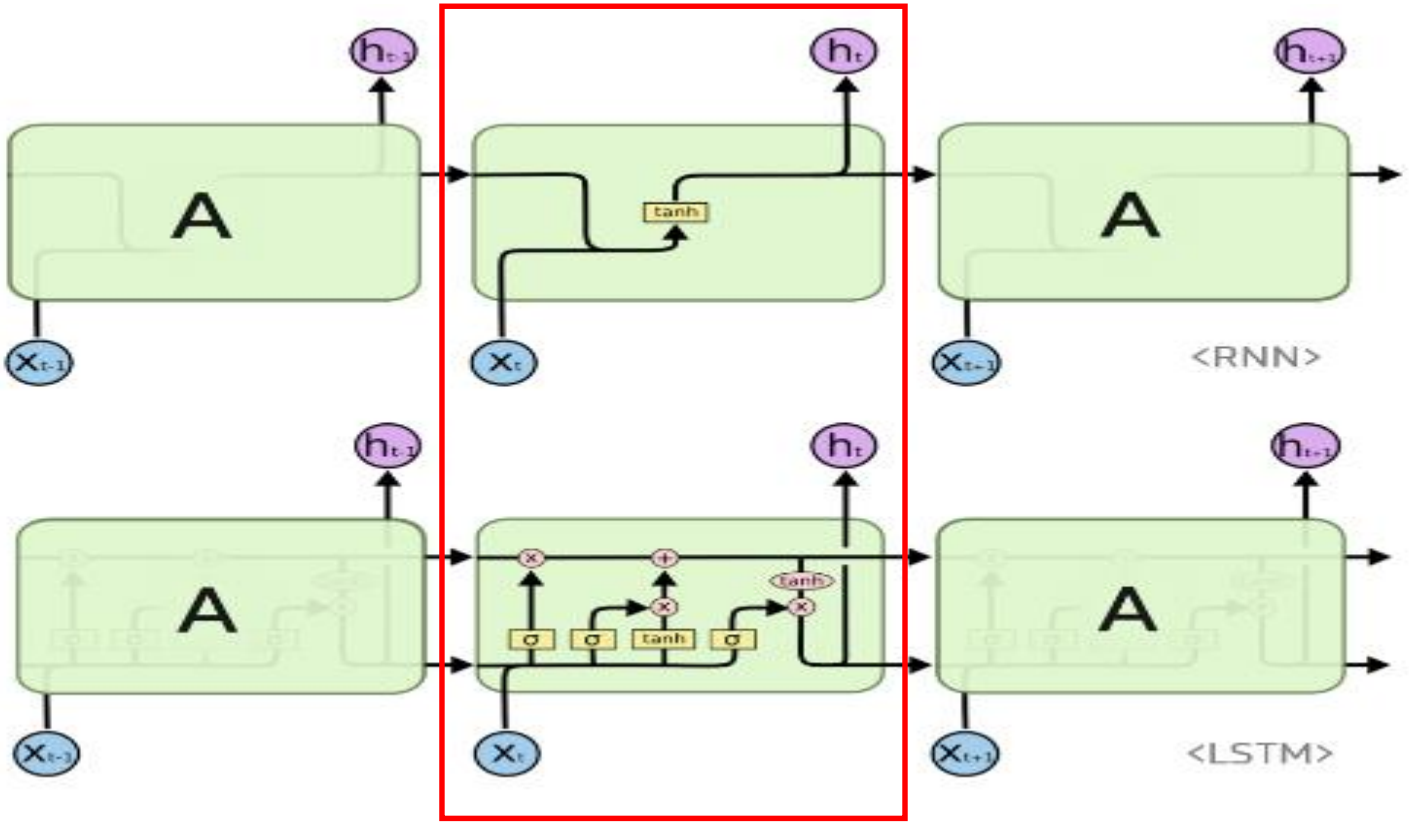
데이터 기법 1. Crawling

- Requests module
- Beautiful soup

데이터 기법 2. LSTM & GRU

RNN 과 LSTM의 차이

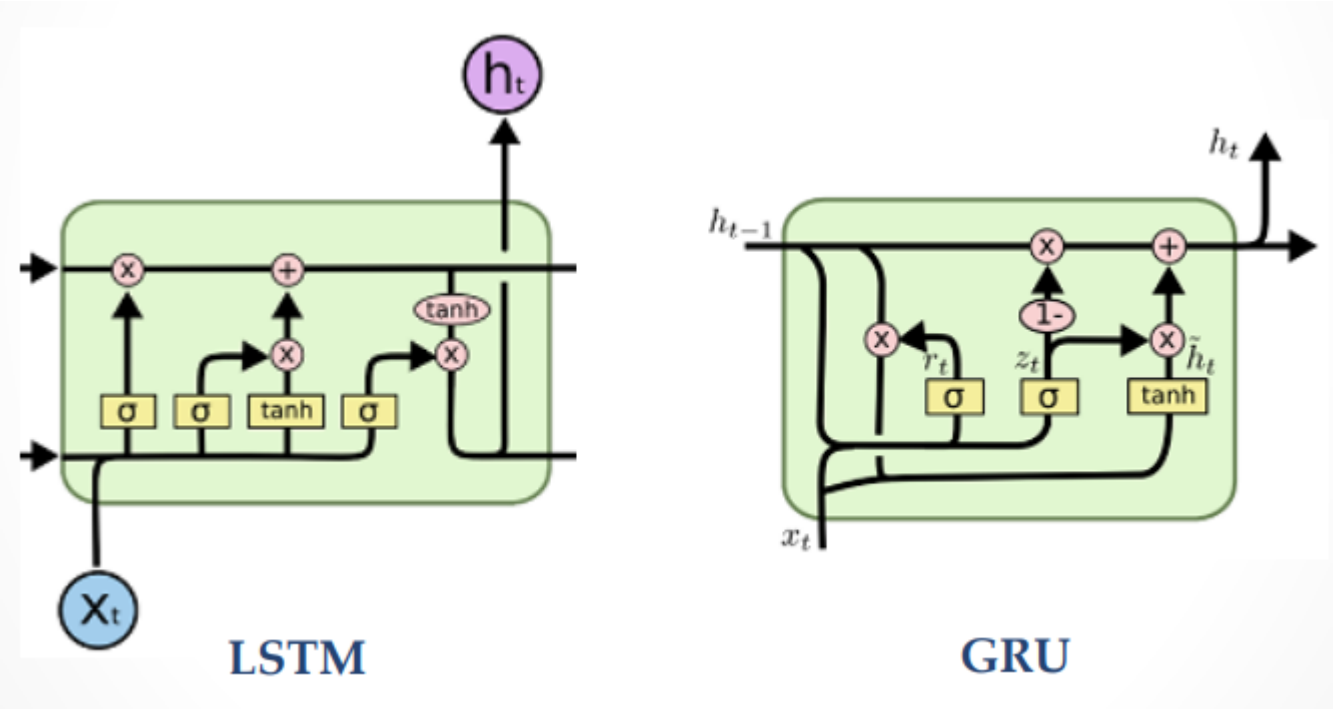
vanishing gradient problem 해결



데이터 기법 2. LSTM & GRU

Seq2Seq Input length ↑ → 망각 ↑

→ Solution 1. LSTM 2. GRU



데이터 기법 2. LSTM & GRU

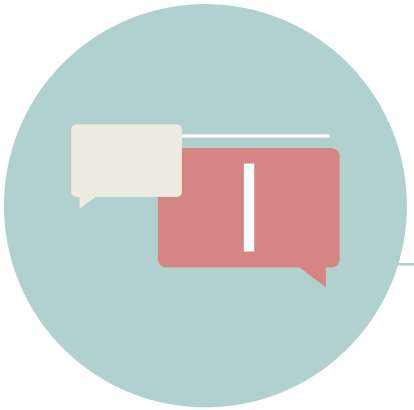
공통점과 차이점

LSTM GRU

공통점 Vanishing gradient 문제 해결을 하여 효과적으로 긴 시퀀스 처리

차이점

1. LSTM GATE 3개 GRU GATE 2개
2. LSTM forget_gate, input_gate를 GRU update_gate로 합침
3. forgate_gate 역할이 r(t) 와 z(t) 둘 다에서 나눠짐
4. GRU는 H(t) 출력할 때 비선형 함수를 적용하지 않는다.



BIG DATA
정확도 향상을 위한 노력

1. Input & Output Length Reduction
2. Object Designation
3. Hidden & Epoch Adjustment
4. RNN Or GRU Comparison

1. Input & Output Reduction

초기

	Input	Output
예측	오늘 날씨를 알려줘	오늘 날씨는 OO도이고 비 올 확률은 OO%이고 통합대기수치는 OOO이고 미세먼지와 오존 상태는 OOO입니다.

➡ **Result : Bad**

문제 발생 후 해결책

오늘 날씨는 몇 도야?	오늘 날씨는 OO도 입니다.
오늘 비 올 확률은?	오늘 비가 올 확률은 OO%입니다.
오늘 통합대기수치는?	오늘 통합대기 수치는 OO입니다.
오늘 미세먼지와 오존 상태는?	오늘 미세먼지와 오존 상태는 OOO 입니다.

➡ **Result : Not Bad**

Object Designation

Input and Output Reduction ➡ Result = Not Bad

초기보다 많이 줄였지만, 다소 부정확한 경우가 많이 발생

Solution Output에 특정 객체로 나타내게 지정을 해서 학습시키기

```
어제 보다 기온이 어때?
cc 입니다.
비 올 확률은 얼마야?
비가 올 확률은 dd 입니다.
미세먼지랑 오존 상태가 어때?
오늘 미세먼지와 오존 상태는 ee 입니다.
```

```
    return time, temper, temp_change , rainfall , dust_ozone ,dust_number, ozone_number,total_number,tomo_morning,t

data=main()
aaa=data[0];bbb=data[1];ccc=data[2];ddd=data[3];eee=data[4];fff=data[5];ggg=data[6];hhh=data[7]
iii=data[8];jjj=data[9];kkk=data[10];lll=data[11];mmm=data[12];nnn=data[13]
ooo=data[14];ppp=data[15];qqq=data[16];rrr=data[17];sss=data[18];ttt=data[19]
```

```
weather=["aa","bb","cc","dd","ee","ff","gg","hh","ii","jj",
         "kk","mm","ll","nn","oo","pp","qq","rr","ss","tt"]
dic={"aa":aaa,"bb":bbb,"cc":ccc,"dd":ddd,"ee":eee,"ff":fff,"gg":ggg,"hh":hhh,"ii":iii,
     "jj":jjj,"kk":kkk,"ll":lll,"mm":mmm,"nn":nnn,"oo":ooo,"pp":ppp,"qq":qqq,"rr":rrr,
     "ss":sss,"tt":ttt}
```

➡ Result = Good

Hidden & Epoch Adjustment

```
learning_rate = 0.001
n_hidden = 900
total_epoch = 800
```

Learning_rate , n_hidden , total_epoch 조정함으로써 좀 더 적은 오차가 나오는 것을 찾으려고 노력

결론적으로, n_hidden = 900 , total_epoch 800

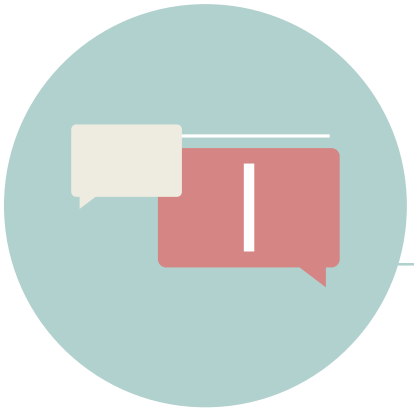
learning_rate = 0.001 일 때 가장 정확

RNN Or GRU Comparison

RNN 과 GRU는 기능 상 거의 차이가 없지만, 좀 더 나은 정확도 향상을 위해

N_HIDDEN , N_EPOCH , LEARNING_RATE 조정을 통해 가장 정확도 좋은 것 찾기 위해 노력

```
#### gru 보다는 lstm 좀 더 정확하다.
with tf.variable_scope("encoder"):
    enc_cell = tf.contrib.rnn.BasicLSTMCell(n_hidden)
    ### 과적합 방지
    #enc_cell =tf.contrib.rnn.GRUCell(n_hidden)
    enc_cell = tf.contrib.rnn.DropoutWrapper(enc_cell, output_keep_prob=0.5)
    outputs, enc_states = tf.nn.dynamic_rnn(enc_cell, enc_input,
                                            dtype=tf.float32)

# 디코더
with tf.variable_scope("decoder"):
    dec_cell = tf.contrib.rnn.BasicLSTMCell(n_hidden)
    #dec_cell =tf.contrib.rnn.GRUCell(n_hidden)
    dec_cell = tf.contrib.rnn.DropoutWrapper(dec_cell, output_keep_prob=0.5)
    outputs, dec_states = tf.nn.dynamic_rnn(dec_cell, dec_input,
                                            initial_state=enc_states,
                                            dtype=tf.float32)

model = tf.layers.dense(outputs, n_class, activation=None)

cost = tf.reduce_mean(
    tf.nn.sparse_softmax_cross_entropy_with_logits(
        logits=model, labels=targets))

optimizer = tf.train.AdamOptimizer(learning_rate).minimize(cost)
```



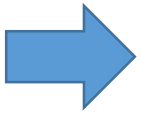
BIG DATA
시연

1. Code
2. Output

Code 1. 대화문

```
나이가 들어 성공한다는 뜻은?
대기만성
먼 훗날까지 걸친 큰 계획은?
백년대계
현재시간을 알려줘
지금 시간은 aa 입니다.
오늘의 기온 좀 알려줘
오늘의 기온은 bb 입니다.
어제 보다 기온이 어때?
cc 입니다.
비 올 확률은 얼마야?
비가 올 확률은 dd 입니다.
미세먼지랑 오존 상태가 어때?
오늘 미세먼지와 오존 상태는 ee 입니다.
미세먼지 정확한 수치는 어때?
오늘 ff 입니다.
오존의 수치는?
gg 입니다.
통합 대기 수치는?
hh 입니다.
내일 오전에는 어때?
내일 오전은 ii 입니다.
내일 오후날씨는?
내일 오후는 jj 입니다.
2일 뒤 오전 날씨는?
2일 후 오전은 kk 입니다.
2일뒤 오후 날씨는?
2일 후 오후는 ll 입니다.
3일뒤 오전기온은?
3일 후 오전은 mm 입니다.
3일뒤 오후기온은?
3일 후 오후는 nn 입니다.
4일뒤 오전은?
4일 후 오전은 oo 입니다.
4일뒤 오후 날씨는?
4일 후 오후는 pp 입니다.
5일뒤 오전은?
5일 후 오전은 qq 입니다.
5일뒤 오후 날씨는?
5일 후 오후는 rr 입니다.
6일뒤 오전은?
6일 후 오전은 ss 입니다.
6일뒤 오후 날씨는?
6일 후 오후는 tt 입니다.
영단어 뜻을 알려줘
zz의 뜻은 vv 입니다.
```

대화문



((*Input*, *Output*),
(*Input* , *Output*))

```
f = open("대화문.txt", 'r')
data2=f.read().splitlines()
script=[]
for i in range(0,len(data2),2):
    ss=[data2[i],data2[i+1]]
    script.append(ss)


train_data = script
train_data
```

```
[['안녕', '오하요'],
 ['가지마', '아이 윌 비 백'],
 ['넌 이름이 뭐니?!', '바리바리양세바리'],
 ['이름이 별로야', '넌 이름이 뭐야'],
 ['난 이성형이야', '뭐야 교회다녀?!'],
 ['너는 뭐가 제일 먹고 싶어?', '치킨이지'],
 ['너의 취미는 뭐야?', 'zZzZ자는거야 '],
 ['나도 마찬가지야!', '나랑 비슷하구나'],
 ['너는 게임을 하니?', '음양사를 해'],
 ['레벨이 몇인데?', '레벨은 52정도?'],
```

Code 2. Weather Crawling

```
day =input("오늘 날짜를 써줘 형식은 yyyymmdd 이야")
print("고마워 곧 %s에 대한 메인뉴스를 가져올게" %(day))
### 메인뉴스
main_news= "http://weather.naver.com/news/wetrNewsList.nhn?ymd="+day
page = requests.get(main_news)
page.status_code
soup = BeautifulSoup(page.text,'html.parser')
content=soup.find("div",{"id":"content_sub"})
main_adrress=content.find("dt").find("a").get('href')
page = requests.get(main_adrress)
page.status_code
soup = BeautifulSoup(page.text,'html.parser')
content=soup.find("div",{"id":"articleBodyContents"})
num=str(content).find("</script>")
numend=str(content).find("<!-- // 본문 내용 -->")
news = re.compile('[^ ㄱ-ㅣ가-힣]+')
main_news2=news.sub('',str(content)[num:numend]).replace("  "," ").strip(" ")
main_news=['메인뉴스',main_news2]
```

Requests module , BeautifulSoup 을

사용하여 원하는 부분 크롤링 한 후 객체로 저장

```
    return time, temper, temp_change , rainfall , dust_ozone ,dust_number, ozone_number,total_number,tomo_morning,t

data=main()
aaa=data[0];bbb=data[1];ccc=data[2];ddd=data[3];eee=data[4];fff=data[5];ggg=data[6];hhh=data[7]
iii=data[8];jjj=data[9];kkk=data[10];lll=data[11];mmm=data[12];nnn=data[13]
ooo=data[14];ppp=data[15];qqq=data[16];rrr=data[17];sss=data[18];ttt=data[19]
```

Code 3. RNN & GRU

```
for seq in train_data:
    # 인코더 셀의 입력값. 입력단어의 글자들을 한글자씩 떼어 배열로 만든다.
    input = [num_dic[n] for n in seq[0]+'P' * (max_input_text - len(seq[0]))]# P는 Padding 값
    # 디코더 셀의 입력값. 시작을 나타내는 [ 심볼을 맨 앞에 붙여준다. (Seq의 구분)
    output = [num_dic[n] for n in ('[' + seq[1] + 'P' * (max_output_text - len(seq[1])))]
    # 학습을 위해 비교할 디코더 셀의 출력값. 끝나는 것을 알려주기 위해 마지막에 ] 를 붙인다.
    target = [num_dic[n] for n in (seq[1] + 'P' * (max_output_text - len(seq[1])) + ']' )]
    input_batch.append(np.eye(dic_len)[input])
    output_batch.append(np.eye(dic_len)[output])
    target_batch.append(target)
return input_batch, output_batch, target_batch
```

```
#### gru 보다는 lstm 좀 더 정확하다.
with tf.variable_scope("encoder"):
    enc_cell = tf.contrib.rnn.BasicLSTMCell(n_hidden)
    ### 과적합 방지
    #enc_cell =tf.contrib.rnn.GRUCell(n_hidden)
    enc_cell = tf.contrib.rnn.DropoutWrapper(enc_cell, output_keep_prob=0.5)
    outputs, enc_states = tf.nn.dynamic_rnn(enc_cell, enc_input,
                                            dtype=tf.float32)

# 디코더
with tf.variable_scope("decoder"):
    dec_cell = tf.contrib.rnn.BasicLSTMCell(n_hidden)
    #dec_cell =tf.contrib.rnn.GRUCell(n_hidden)
    dec_cell = tf.contrib.rnn.DropoutWrapper(dec_cell, output_keep_prob=0.5)
    outputs, dec_states = tf.nn.dynamic_rnn(dec_cell, dec_input,
                                            initial_state=enc_states,
                                            dtype=tf.float32)

model = tf.layers.dense(outputs, n_class, activation=None)

cost = tf.reduce_mean(
    tf.nn.sparse_softmax_cross_entropy_with_logits(
        logits=model, labels=targets))

optimizer = tf.train.AdamOptimizer(learning_rate).minimize(cost)
```

Code 4. 사전

```
def dictionary(word) :
    main_news= "http://endic.naver.com/search.nhn?sLn=kr&isOnlyViewEE=N&query="+word
    page = requests.get(main_news)
    page.status_code
    soup = BeautifulSoup(page.text,'html.parser')
    content=soup.find("div",{"class":"align_right"})
    aa=content.find_all("span")
    b=str(aa)
    ########### <>안에 있는 것들 다 제거하기
    k=re.sub('<.*?>',"",b)
    index=k.find("더보기")
    word2=k[:index-7].replace("[","").replace("]","")
    return word2
```

Code 5. 시연

```
def question():
    number=int(input("몇 개 물어 볼꺼야?"))
    for i in range(1,number+1):
        question=print("{}번째 질문".format(i) )
        answer=input(question)
        output=''.join(predict([answer,'']))
        for a in weather :
            if output.find(a) > -1 :
                output=output.replace(a,dic[a])
        if answer.find("영단어") > -1:
            word1=input("어떤 단어 물어볼거야?")
            word2=dictionary(word1)
            output=output.replace("zz",word1).replace("yy",word2)
        print("A: " + output)

def ask():
    question()
    while True :
        a=input("다 물어 본거야?")
        if a=="응" :
            print("그래 이만 안녕~")
            break
        else  :
            question()
```

Output 1.Weather Crawling

안녕 만나서 반가워

뭐가 궁금하니?날씨

아~ 날씨가 궁금하구나?응

오늘은 2017-9-18 인데, 오늘이 궁금한거지?응

오늘 날짜를 써줘 형식은 yyyymmdd 이야20170918
고마워 곧 20170918에 대한 메인뉴스를 가져올게
오늘의 메인 뉴스야!! ['메인뉴스', '날씨 청명한 하늘강한 자외선한낮 더위해지면 서늘앵커오늘은 전국에 쾌청한 가을 하늘이 펼쳐져있습니다 볕이 뜨겁게 내리쬐면서 기온을 끌어올리고 있는데요자세한 날씨 기상캐스터 연결해 알아보겠습니다 주정경 캐스터캐스터한 주가 시작되는 월요일인 오늘 청명한 가을 하늘이 펼쳐져있습니다오늘 전국 하늘이 맑은 가운데 햇볕이 강하게 내리쬐고 있습니다옛말에 가을볕은 딸에게 쬐인다는 말이 있을 정도로 잠깐씩의 바깥활동은 몸에 이롭습니다만 오늘같이 강한 자외선이 강한 날에는 자외선차단제 꼼꼼하게 바르는 것을 습관들이시고 외출하시는 것이 좋겠습니다낮 기온 쭉쭉오르고 있습니다 현재시각 서울의 기온은 도 대구가 도 광주도 도를 가리키고 있는데요내일은 전국에 구름이 다소 많겠고 오후부터 밤 사이 중부지방과 전북 경북 내륙으로 비가 올 것으로 보입니다 비의 양은 에서 최고 로 많지는 않겠습니다하늘이 흐려 낮기온보다 낮아지겠습니다서울 도 대전 도 예상되고요광주 도 대구 도 등 남부지방은 오늘과 비슷하겠습니다 모레 수요일에는 전북과 영남지역으로도 가을비 소식이 들어있습니다 이후로는 이번 한주 내내 대체로 맑은 날씨가 이어질 전망입니다지금까지 광화문에서 전해드렸습니다 주정경 기상캐스터연합뉴스 기사문의제보 ÞÞ톡라인 연합뉴스 생방송 시청 뉴스스탠드 구독 대한민국 뉴스의 시작 연합뉴스 앱 다운받기']

메인 뉴스는 보여줬고 원하는 구 있어? 아 근데! 서울특별시 쪽만 물어봐 줄래?ㅎ응
아 '응' 라고 생각하는구나 고마워 지금 알아볼게
서울 특별시 구에 대한 정보야 ['강남구', '강동구', '강북구', '강서구', '관악구', '광진구', '도봉구', '동대문구', '동작구', '마포구', '서대문구', '서초구', '성동구', '성북구', '송파구', '양천구', '영등포구', '용산구', '은평구', '종로구', '중구', '중랑구']

원하는 구가 어디야?강남구
강남구동 데이터야 ['서울특별시_강남구_대치동', '서울특별시_강남구_역삼1동', '서울특별시_강남구_압구정동', '서울특별시_강남구_개포2동', '서울특별시_강남구_삼성1동', '서울특별시_강남구_청담동', '서울특별시_강남구_개포동', '서울특별시_강남구_삼성동', '서울특별시_강남구_도곡동', '서울특별시_강남구_도곡1동', '서울특별시_강남구_논현2동', '서울특별시_강남구_도곡2동', '서울특별시_강남구_일원본동', '서울특별시_강남구_일원1동', '서울특별시_강남구_일원2동', '서울특별시_강남구_개포4동', '서울특별시_강남구_자곡동', '서울특별시_강남구_대치2동', '서울특별시_강남구_율현동', '서울특별시_강남구_세곡동', '서울특별시_강남구_논현동', '서울특별시_강남구_일원동', '서울특별시_강남구_신사동', '서울특별시_강남구_수서동', '서울특별시_강남구_역삼2동', '서울특별시_강남구_대치1동', '서울특별시_강남구_삼성2동', '서울특별시_강남구_논현1동', '서울특별시_강남구_개포1동', '서울특별시_강남구_역삼동', '서울특별시_강남구_대치4동']

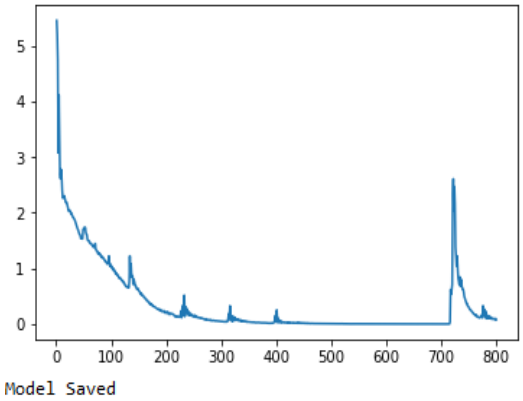
Output 2. 시연

```
몇 개 물어 볼꺼야?10
1번째 질문

None안녕
A: 오하요
2번째 질문

None네를 비꼬는 듯이 말하는 것은?
A: 뉘예~뉘예~
3번째 질문

None너가 좋아하는 한자성어 있어?
A: 계륵ㅎㅎ넌 계
4번째 질문

None나이가 들어 성공한다는 뜻은?
A: 대기만성
5번째 질문

None현재시간을 알려줘
A: 지금 시간은 15시 현재 입니다.
6번째 질문

None비 올 확률은 얼마야?
A: 비가 올 확률은 강수확률 10% 입니다
7번째 질문

None내일 오전에는 어때?
A: 내일 오전은 오전 17.0℃ 구름많음강수확률 20% 입니다.
8번째 질문

None미세먼지랑 오존 상태가 어때?
A: 오늘 미세먼지와 오존 상태는미세먼지 좋음 오존 보통e입니다
9번째 질문

None6일 뒤 오후 날씨는?
A: 6일 후 오후는 일요일 오후 26.0℃ 구름조금 입니다.
10번째 질문

None어제 보다 기온이 어때?
A: 어제보다 -1.0℃ 입니다.

다 물어 본거야?아니
```

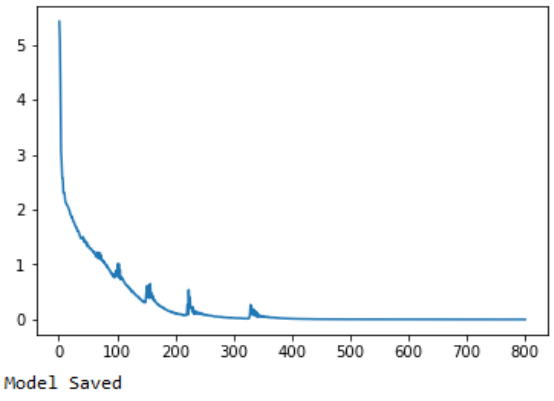
```
몇 개 물어 볼꺼야?3
1번째 질문

None영단어 뜻을 알려줘

어떤 단어 물어볼거야?돌고래
A: 돌고래의 뜻은 dolphin; (몸집이 작은) porpoise, 한 무리의 돌고래, a school of d
입니다.
2번째 질문

None영단어 뜻을 알려줘

어떤 단어 물어볼거야?교회
A: 교회의 뜻은 church, 교회에 다니다, attendgo to 입니다.
3번째 질문

None영단어 뜻을 알려줘

어떤 단어 물어볼거야?church
A: church의 뜻은 명사, 교회 (건물), a church tower, 교회탑 입니다.

다 물어 본거야?응
그래 이만 안녕~


In [7]:
```

Output 3. 오차그래프 (GRU & LSTM)

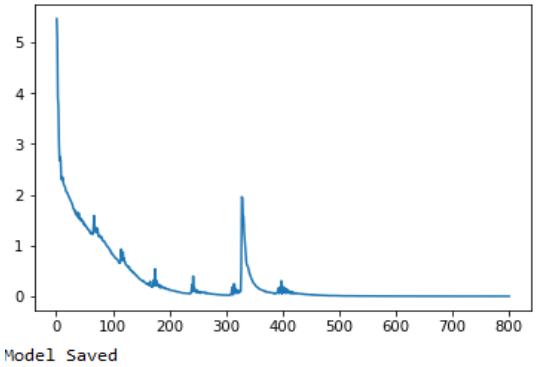
GRU_H800E800



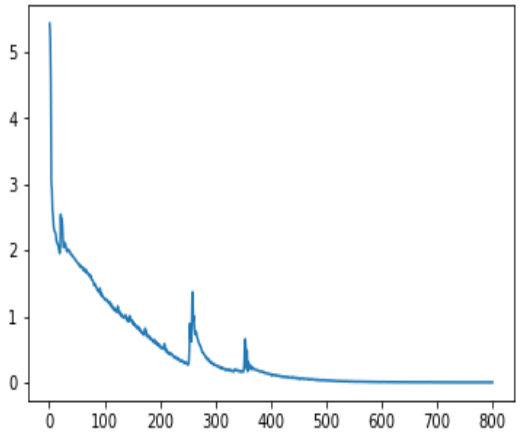
GRU_H900E800



GRU_H1000E800

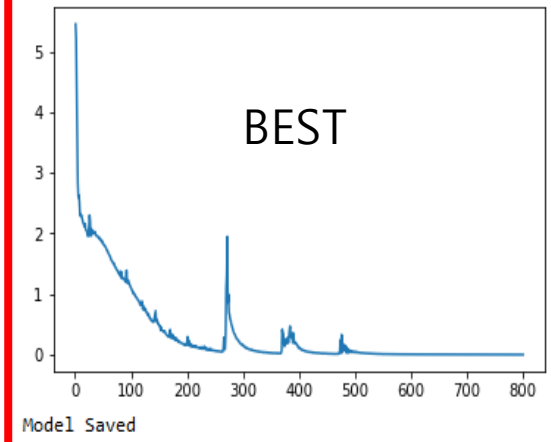


LSTM_H800E800

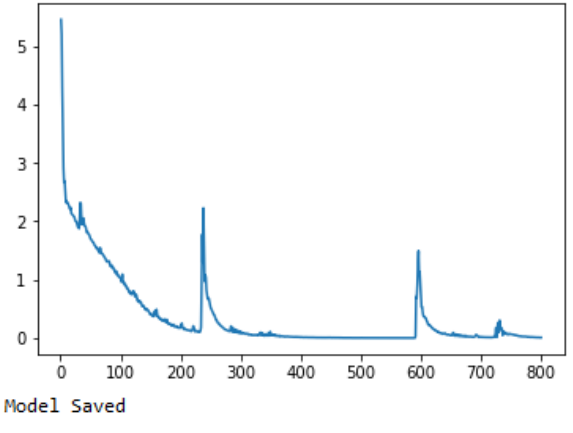


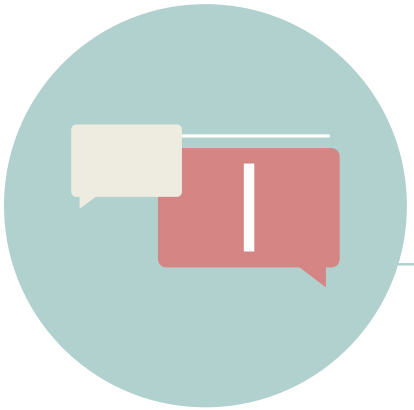
LSTM_H900E800

BEST



LSTM_H1000E800





BIG DATA
개선 해야 할 점

1. Input & Output length
2. 1 : 1 -> N : N
3. Better Code than ever before
4. Platform

Input & Output length

Reduction 과 Object Designation 을 통해 어느정도 Length 에 대한 부분은 해결

그러나 긴 질문을 해야 될 때와 자율적인 긴 대답을 해야 하는 경우가 있기 때문에
이러한 부분에 대해서 좀 더 좋은 code 나 다른 기법이 있다면 찾아봐야 함.

1 : 1 ➔ N : N

지금 코드는 일문일답이므로 한 질문에 대해서 다양한 대답을 얻어 낼 수가 없다.

그리고 답과 다양한 물음을 다시 사용자에게 물어 보는 기능이 없다.

어떻게 해야 하는지는 아직 더 고민이 필요함

Better Code than ever before

코딩실력이 미숙해서 일반적인 상황으로 코딩을 하지 못함

좀 더 공부를 해서 일반적인 상황에서도 적용이 될 수 있게 노력해야 함

Platform

음성 기능을 어떻게 하는지 몰라서 아쉬웠다.

Django와 같은 프로그램을 잘 다루지 못해서 spyder 에서만 구현을 했다.

Django와 같은 프로그램을 공부해서 인터넷화면에서 가능하게 하고 싶다.

Thank you

PTLINE POWERPOINT
TEMPLATE