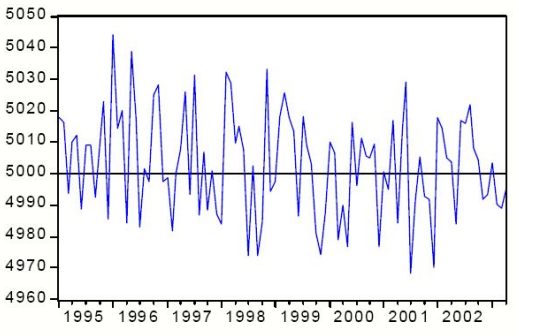
|  |
| --- |
| **ARIMA & LSTM 을 이용한 웹트래픽 시계열 예측** |
| 홍순범 장승민 최규재  서강대학교 컴퓨터공학과  [hannibal0802@naver.com](mailto:hannibal0802@naver.com) [gom626@gmail.com](mailto:gom626@gmail.com) jsgy0000@gmail.com |
| Web -Traffic-Time-Series Forecasting Usingby ARIMA and LSTM |
| Soonbum Hong Seungmin Jang Gyujae Choi  Department of Computer Science and Engineering, Sogang University |
| **요 약**  한국정보과학회는 정보과학에 관한 기술을 발전, 보급시키고 회원상호간의 친목을 도모하기 위하여 1973년 3월 3일에 설립되었으며, 정보통신부에 '사단법인 한국정보과학회'로 등록되었다. 학회의 주요 활동은 1) 컴퓨터 기술 및 이론에 관한 새로운 연구결과를 발표하는 기회를 제공하고, 2) 국내의 컴퓨터 관련 기술 개발에 참여하며, 3) 국제적 학술 교류 및 협력 증진을 도모하고, 4) 회원 상호간의 친목을 증진시키는 것이다. | |

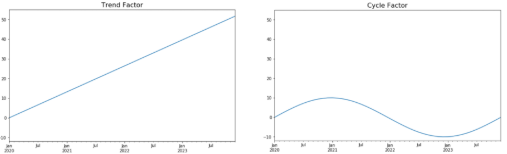
**1. 서론**

본 주제는 복수 시계열의 미래 가치를 예측하는 문제에 초점을 맞추고 있다. 좀 더 구체적으로, 우리는 약 145,000개의 위키백과 기사에 대한 미래의 웹 트래픽을 예측하는 문제에 대해 효과적인 방법을 시험하기 위한 경쟁을 목표로 하고 있다. 순차적 또는 일시적 관찰은 생물학적 데이터, 금융 시장, 날씨 예측, 오디오 및 비디오 처리와 같은 많은 주요 실제 문제에서 나타난다. 시계열 분야는 분석과 추론에서 분류와 예측에 이르는 많은 다양한 문제들을 캡슐화한다.

시계열 분석은 크게 규칙적 시계열 분석과 불규칙적 시계열 분석으로 나뉜다. 여기서 규칙적 시계열이란 트렌드와 분산이 불변하는 시계열 데이터를 말하고, 불규칙적 시계열이란 트렌드 혹은 분산이 변화하는 시계열 데이터를 말하는 것이다. 시계열 데이터를 잘 분석한다는 것은, 불규칙성을 가지는 시계열 데이터에 특정한 기법이나 모델을 적용하여 규칙적 패턴을 적용하거나, 혹은 예측할 수 있도록 하는 것을 의미한다. 아래의 이미지를 보면 직관적으로 이해할 수 있다.



<불규칙 변동>

<규칙적 변동>

불규칙성을 띠는 시계열 데이터에 규칙성을 부여하는 방법으로는 AR, MA, ARMA, ARIMA 모델 등의 분석 방법을 적용하는 것이 가장 널리 알려져 있다. 이러한 모델을 기반으로 도메인과 데이터의 특성에 맞게 발전시킨 모델들이 주로 사용되고 있으며, 최근에는 딥러닝을 이용하여 시계열 데이터의 연속성을 기계 스스로 찾아내도록 하는 방법이 더 좋은 성능을 내고 있다. 가장 대표적인 것이 LSTM이라는 RNN 종류의 네트워크인데, 이전 상태를 꽤나 오랫동안 기억하면서 순환적인 네트워크의 파라미터를 자동적으로 학습하는 방법이다.

**2. 관련 연구**

**2.1 ARIMA**

ARIMA는 Autoregressive Integrated Moving Average의 약자로, Autoregressive는 자기회귀모형을 의미하며, Moving Average는 이동평균모형을 의미한다. 결국 ARIMA 모델은 자기회귀(이전 관측값의 오차항이 이후 관측값에 영향을 주는 모형 - AR(1) : )와 이동평균(관측값이 이전의 연속적인 오차항의 영향을 받는다는 모형 - MA(1) : )을 둘다 고려하는 모형이다.

현실에 존재하는 시계열자료는 불한정(Non-stationary)한 경우가 많다. 따라서 이러한 비정상성을 제거하는 과정을 포함하는 것이 ARIMA모형이며 다음과 같이 표시 한다.

**\* ARIMA (p, d, q)**

**2.2 LSTM(Long Short Term Memory Networks)**

RNN은 은닉층의 과거의 정보가 점점 작아지면서 마지막까지 전달되지 못하는 장기 의존성 문제(The Problems of Long-term Dependencies)의 이슈를 가지고 있다.

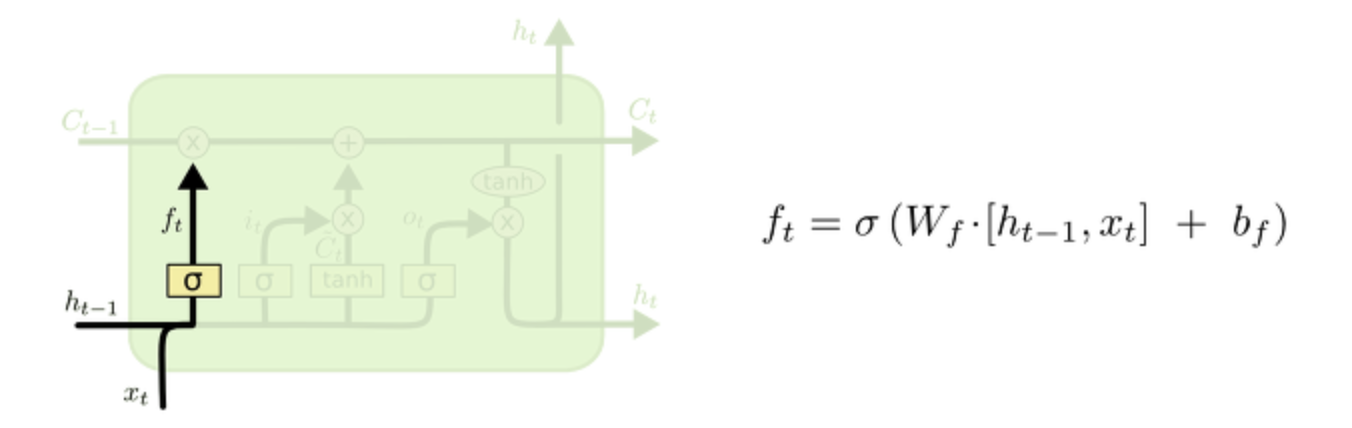
LSTM(Long Short Term Memory Networks, 이하 LSTM)은 이러한 장기 의존성 문제를 해결할 수 있다. LSTM은 RNN과 동일하게 반복되는 체인 구조의 뉴럴 네트워크로 이루어져있다.

RNN과 차이를 만드는 구조는 게이트(gate)를 포함하는 셀 스테이트(Cell State)이다. 이 게이트는 정보를 제거하거나 더하는 기능을 한다. 시그모이드 뉴럴 넷(Sigmoid Neural Net)과 점단위 곱 연산이 이 게이트를 구성한다. 시그모이드 뉴럴 넷은 0이나 1의 결과 값을 출력하는데, 아웃풋에 해당 요소가 얼마나 영향을 줄 것인지를 결정하게 된다. 0은 전혀 영향을 주지 않으며, 1은 확실하게 영향을 주게 되는 방식으로 작동한다.

LSTM은 4가지 단계로 구성되어 있다.

1. Forget Gate Layer

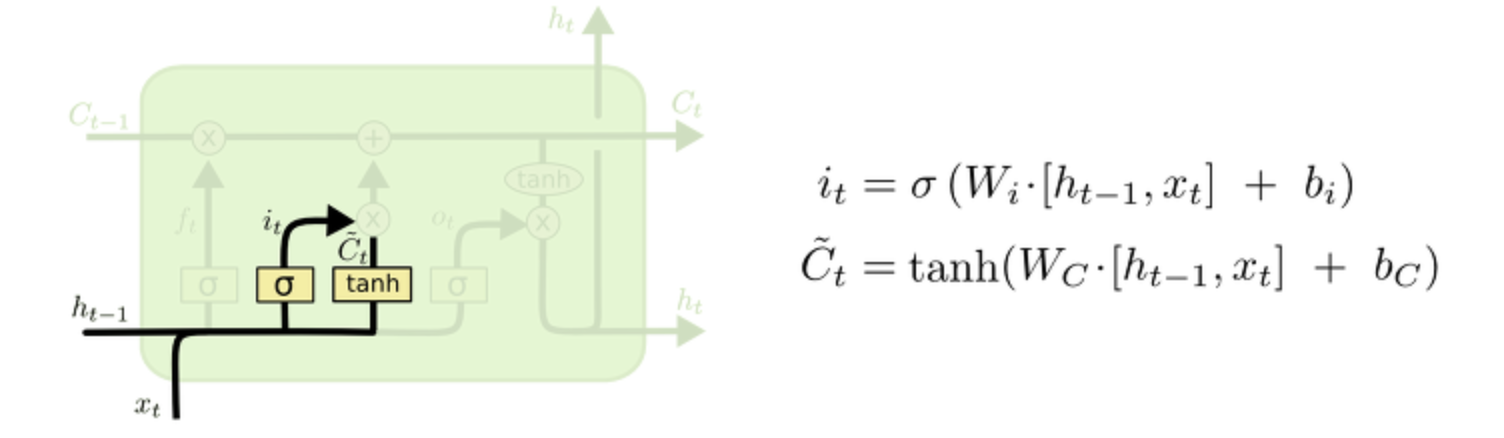
어떠한 정보를 마지막 아웃풋에 반영할 지 정하는 게이트이다. 전 단계 인풋 시퀀스의 hidden stage에서 넘어온 값과 현재 인풋 시퀀스 값을 시그모이드 함수에 넣어 값을 얻는다. 이 값이 1에 가까울 수록 반영이 많이 된다는 의미이고, 0에 가까울 수록 반영이 적게 된다는 의미이다.



<그림 1> Forget Gate Layer 구조 및 식

1. Input Gate Layer

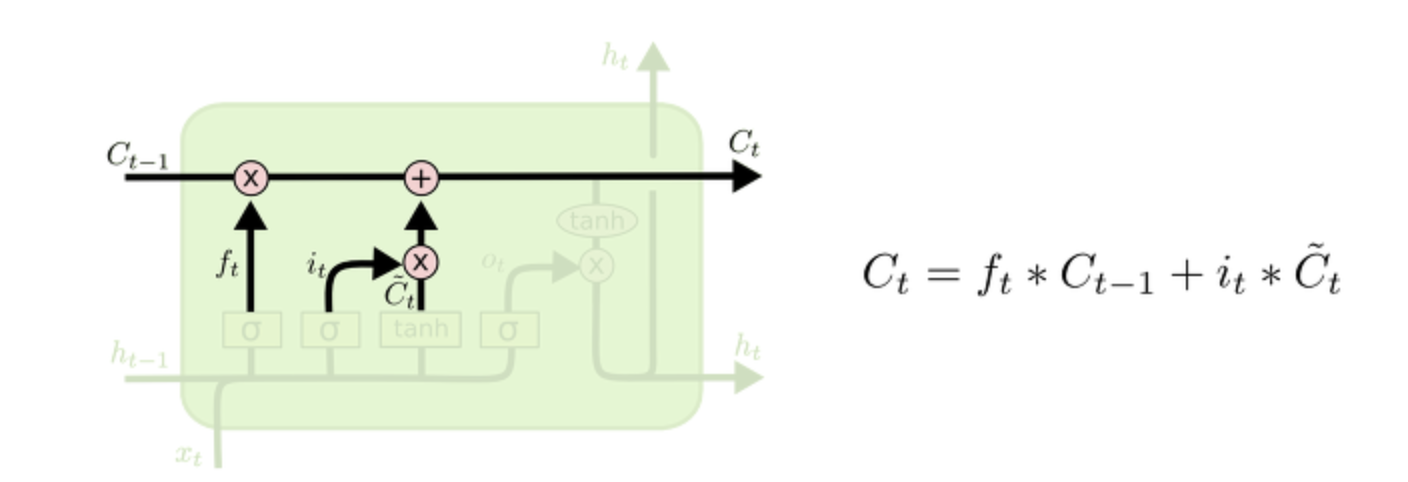
새로운 정보가 Cell State에 저장여부를 정하는 Layer이다. sigmoid와 tanh layer로 구성되어 있다. sigmoid layer는 어떤 값을 업데이트할 지 결정하고, tanh layer는 cell state에 더해질 벡터들의 후보를 업데이트 하는 역할을 담당한다.



<그림 2> Input Gate Layer 구조 및 식

1. Update Cell

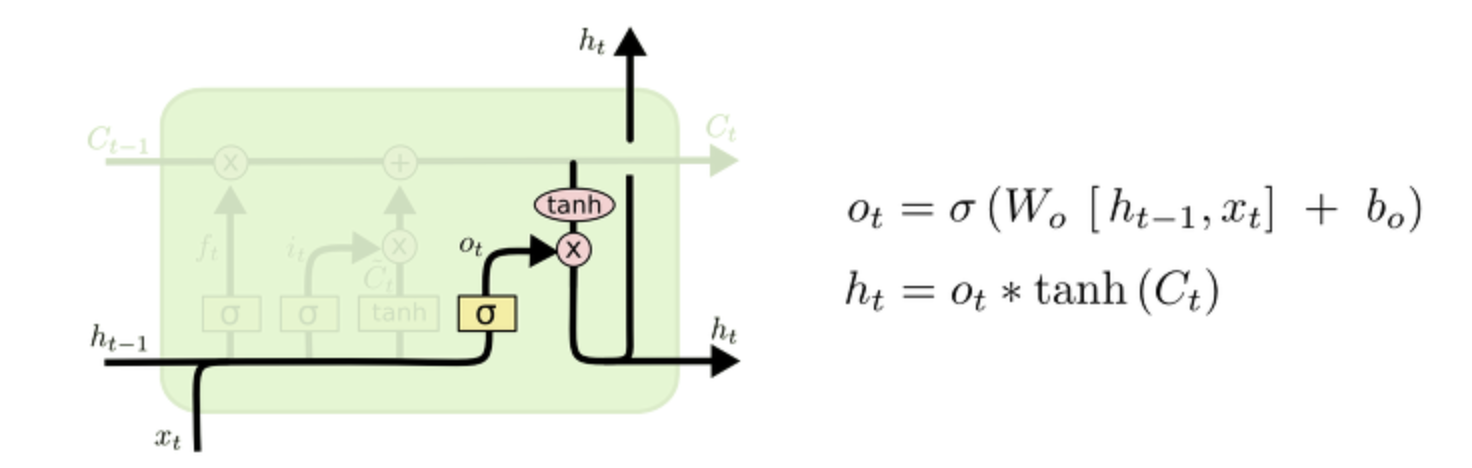
앞 두 단계에서 얻은 값들을 곱 연산을 한 후 현재의 cell state 값을 업데이트 한다.



<그림 3> Cell State 업데이트

1. Output Gate Layer

마지막 출력 값을 정하는 단계이다. 수식은 다음과 같다.

<그림 4> Output Layer 구조 및 식

**3. 구현**

**3.1 ARIMA 모델 학습**

arima = ARIMA(data,params[key])

result = arima.fit(disp=False)

.

**3.2 LSTM 모델 학습**

LSTM 알고리즘은 keras Documentation을 이용하여 구현하였고 코드는 다음과 같은 순으로 이루어 진다.

1. RNN 초기화 (레이어 선형 연결 구성)

LstmModel = Sequential()

1. ADD 메소드를 통해 LSTM 메소드 추가

LstmModel.add(LSTM(units = 8, activation = 'relu', input\_shape = (None, 1)))

3) Output layer 추가

LstmModel.add(Dense(units = 1))

4) Compile 메소드를 통해 학습 방식 설정

LstmModel.compile(optimizer = 'rmsprop', loss = 'mean\_squared\_error')

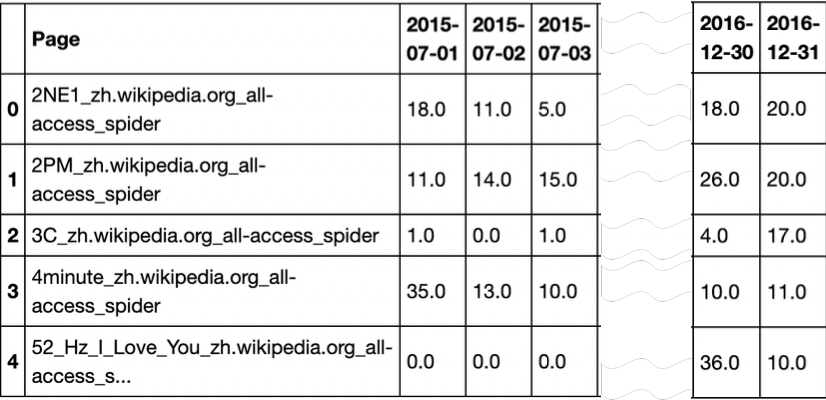
5) 모델 학습

LstmModel..fit(X\_train, y\_train, batch\_size = 10, epochs = 100, verbose = 0)

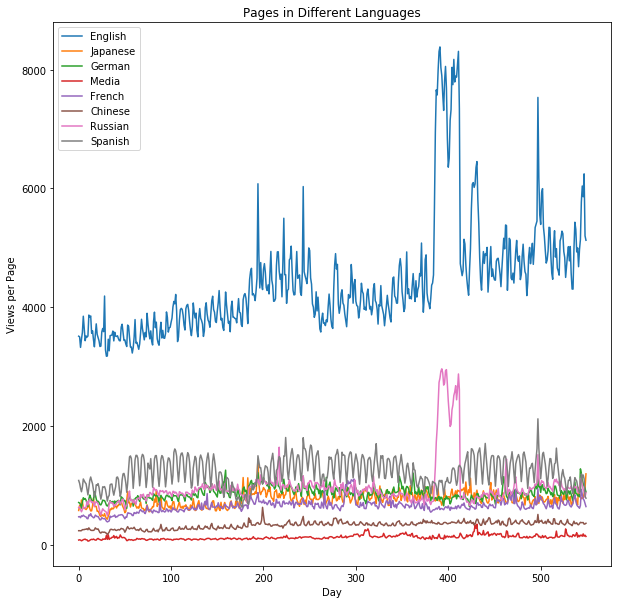
**4. 실험**

실험 데이터는 2015년 7월 1일 부터 2016년 12월 31일 까지 다양한 위키 백과 기사의 일별 조회수를 이용하였다. 실험 시 각 기사에 대한 2017년 9월 13일 부터 2017년 11월 13일 사이의 일일 뷰 예측함으로서 결과를 도출한다.

각 시계열에 대해선 기사의 이름과 이 시계열이 나타내는 트래픽 유형(모바일,데스크탑,스파이더)이 제공된다. (데이터 소스는 0의 트래픽값과 결측값을 구분 하고 있지 않다. 결측값은 트래픽이 0 이거나 해당 날짜에 데이터를 사용할 수 없음을 의미할 수 있다.) 4번씩 수행하여 그 평균값을 취한다.

****

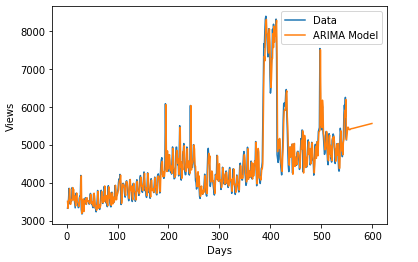
<표 1> train\_1.csv파일 (<https://www.kaggle.com/c/web-traffic-time-series-forecasting/data?select=train_1.csv.zip>)



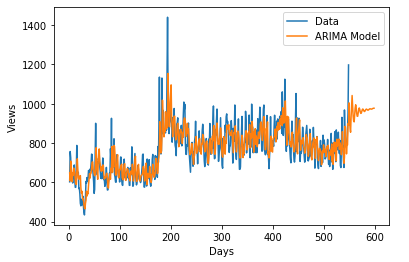
<그림 1> 언어별 일별 평균 조회수

**실험 4.1**

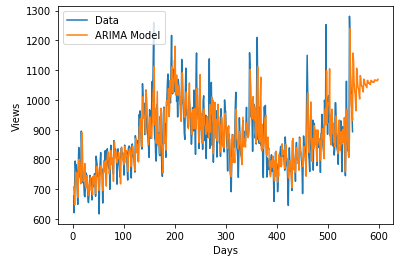
**ARIMA 사용 결과 그래프**



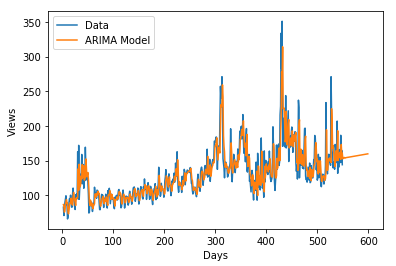
<그림 1> 영어 일일 조회수 예측값



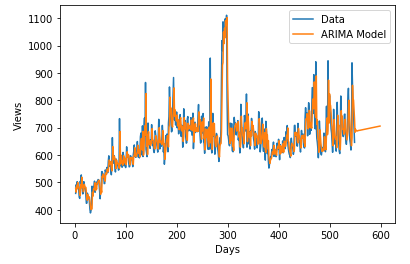
<그림2> 일본어 일일 조회수 예측값



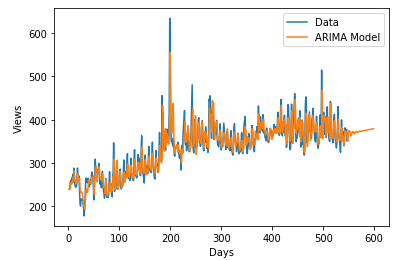
<그림3> 독어 일일 조회수 예측값



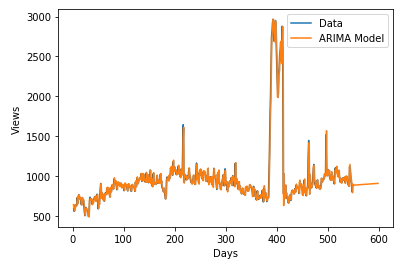
<그림4> 미디어 일일 조회수 예측값



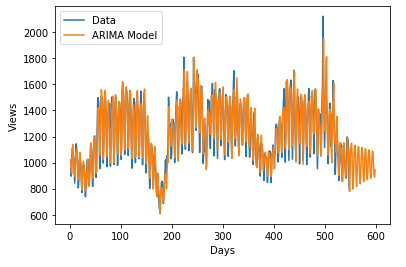
<그림5> 불어 일일 조회수 예측값



<그림6> 중국어 일일 조회수 예측값



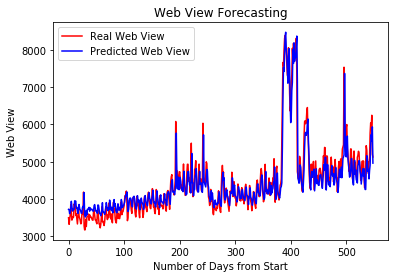
<그림7> 노어 일일 조회수 예측값



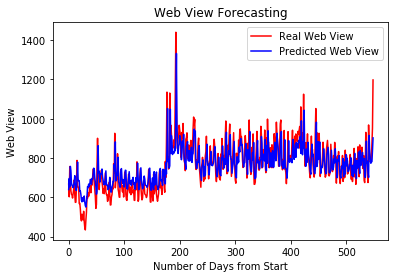
<그림8> 서어 일일 조회수 예측값

**실험 4.2**

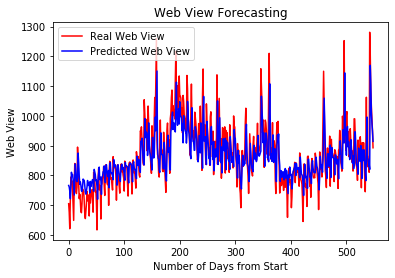
**LSTM 결과 그래프**



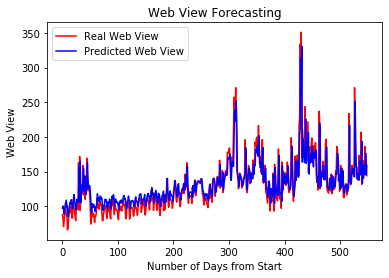
<그림 1> 영어 일일 조회수 예측값



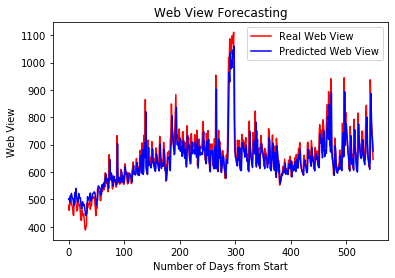
<그림 2> 일본어 일일 조회수 예측값



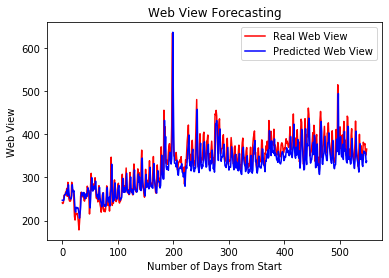
<그림 3> 독어 일일 조회수 예측값



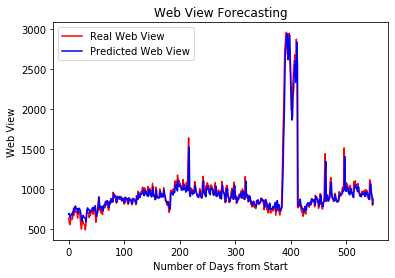
<그림 4> 미디어 일일 조회수 예측값



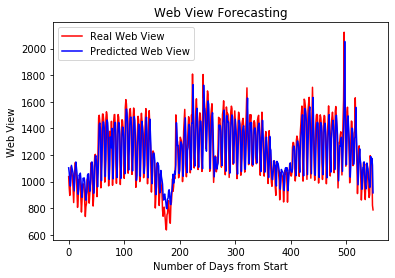
<그림 5> 불어 일일 조회수 예측값



<그림 6> 중국어 일일 조회수 예측값



<그림 7> 노어 일일 조회수 예측값



<그림 8> 서어 일일 조회수 예측값

**5. 결론**

본 논문은 여러가지 시계열 예측 인공지능 알고리즘을 이용해 웹 트래픽 예측하는 것이다. 우리는 가장 대표적인 알고리즘 LSTM과 ARIMA를 이용해 주어진 데이터를 가지고 학습시킨후 학습한 모델을 바탕으로 예측해보았다. 예측 결과를 보면 LSTM모델이 ARIMA모델보다 더 뛰어난 예측 결과를 보여 준다. 따라서 시계열 분석 모델을 선택할 시 ARIMA보다는 LSTM 모델을 선택하는 것을 추천해 본다.

**6. 참고문헌**

1. [Su, Yang](https://library.sogang.ac.kr/eds/brief/discoveryResult?st=KWRD&service_type=brief&si=AU&q=%22Su%2C+Yang%22) (2020) : Research on Website Phishing Detection Based on LSTM RNN
2. [Lee, Dongwon](https://library.sogang.ac.kr/eds/brief/discoveryResult?st=KWRD&service_type=brief&si=AU&q=%22Lee%2C+Dongwon%22)(2020) : Prediction of Network Throughput using ARIMA

**7. 팀원 공헌도**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 장승민 | 최규재 | 홍순범 |
| 공헌도 | 34% | 33% | 33% |