

WEB TRAFFIC PREDICTION



SMART BRIDGE

**UNDER GUIDANCE:**

Ms. DEEPIKA

**GROUP MEMBERS:**

B. SASHIDHAR

M. VAMSHI REDDY

N. GURU TARUN

P. GANESH

S. SATHWIK

# Introduction

The main aim of the project is to predict the future web traffic for the particular website and to help the company to maintain the servers to give good performance of customer service. This project focuses on the problem of forecasting the future values of multiple time series. Data comes from a Kaggle website. (<https://www.kaggle.com/c/web-traffic-time-series-forecasting>)

The dataset consists of approximately 145k time series. Each of these time series represent a number of daily views of a different Wikipedia article, starting from July, 1st, 2015 up until December 31st, 2016, with missing data. Since that's the only one dataset we have, we manually split dataset into train set, validation set and test set. We take the MSE of daily values as the key metrics. We begin this project by importing dataset, dealing with missing data, simple EDA, RNN, LSTM, Time Series Model and a short conclusion.

The network traffic analysis is a significant stage for developing successful preventive congestion control schemes and to find out normal and malicious packets. These schemes target to avoid network congestion by distributing the network resources with respect to the forecasted traffic.

**Python**

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace.

**Artificial Intelligence**

In computer science, artificial intelligence, sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and animals. The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

The **modules** we used here are like:

Numpy, pandas, matplotlib, math and missingno.

# Objectives of Research

The main aim of the project is to predict the number of users visiting sites are whether increasing or decreasing day to day. For a linear graph it is easy to predict the no. of users visiting the website when compared to non linear graph to overcome this problem we are using **RNN, LSTM and SIMPLE RNN** algorithms to get **maximum accuracy**. The experiments with neural networks will achieve the best results for large data. This research opens the possibilities for the development of more efficient traffic engineering and anomaly detection tools which will result in financial gains from better network resource management.

Problem Statement

The problem of forecasting the future values of multiple time series, as it has been one of the most challenging problem in the field. More specifically, on the problem of forecasting future web traffic for approximately 145k Wikipedia articles.

Sequential or temporal observations emerge in many key real-world problems, ranging from biological data, financial markets, weather forecasting, to audio and video processing. The field of time series encapsulates many different problems, ranging from analysis and inference to classification and forecast.

# Review of literature

One of the most important activities related to traffic control is the planning for short-term forecasting, an example of which can be the prediction of daily traffic for the next few days. One of the prediction methods is the time-series forecasting. In time-series forecasting, the historical data are collected and analysed to make a model.

Many traditional methods are developed for time-series prediction. However, in recent decades, neural networks have often been used for modelling the prediction. Time-series prediction can be done using both linear and nonlinear methods. Linear method can be implemented and developed easily. Moreover, they are more understandable and interpretable than nonlinear networks. Nevertheless, this method has limitation: they cannot capture nonlinear relations and their approximation is not as easy as nonlinear ones.

Our main idea is to use Wikipedia traffic data as a core feature to evaluate the structural properties of graphs. The main asset of this method is that it represents a new insight into the data. We will study the accuracy matrices, comparing our prediction with the actual ground truth of edge existence, both on train and test data.

**Data Collection**

Our project focuses on the problem of predicting the future values of multiple time series, as it has always been one of the most challenging problems in the field. More specifically, on the problem of forecasting future web traffic for approximately 145,000 Wikipedia articles.

**Methodology:**

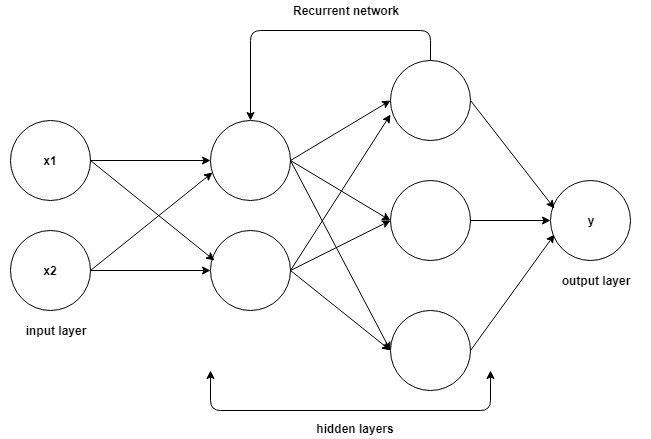
**Neural networks**

Neural networks are computing models to compute the process of information. They are useful for recognizing the pattern or relationship between data. A network is a collection of simple structures which together create a complex system.

* The set of nodes, each node in the network computing unit which performs processing on the input and output obtained.
* Links between nodes which define the connections and the transition of the data between nodes. These connections can be unidirectional or bidirectional.

Interaction between nodes through the network connections are presented by a general behaviour. Such behaviour does not occur alone in any of the network elements. Extensive application of the general behaviour on each node makes a network a powerful tool. The neural networks have three segments input layer, hidden layer and output layer.

Error back propagation algorithm is a method of network training which modifies the network weights. This method is used in the Recurrent Neural Networks. Recurrent means that the outputs of the neurons in each layer are transferred to the next layer. To implement this approach, the network weights are randomly selected then based on input and output data, errors are calculated. Weights are updated in the sense that the calculated errors are minimized.



**Using neural network for time-series forecasting**

For solving different problems, many type of neural network are developed. For time-series forecasting, the recurrent neural network is mostly used. The following figure represents a recurrent neural network. This kind of neural network is composed of several layers called nodes or neurons. In this network model, there is one input layer and one output layer. The input nodes are used for receiving the data of the past. The most important part of the network is the hidden layer. Hidden layer is composed of nodes which are connected to both output and input layers. The output nodes are used for prediction of the future value of the time-series. In the recurrent neural network, the information flow is in one and bi directional.

**LSTM**

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behaviour. All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.

All RNNs have feedback loops in the recurrent layer. This lets them maintain information in 'memory' over time. But it can be difficult to train standard RNNs to solve problems that require learning long-term temporal dependencies. This is because the gradient of the loss function decays exponentially with time (called the vanishing gradient problem).LSTM networks are a type of RNN that uses special units in addition to standard units . LSTM units include a 'memory cell' that can maintain information in memory for long periods of time. A set of gates is used to control when information enters the memory, when it's output, and when it's forgotten.This architecture lets them learn longer-term dependencies. GRUs are similar to LSTMs, but use a simplified structure. They also use a set of gates to control the flow of information, but they don't use separate memory cells, and they use fewer gates.

**Output**

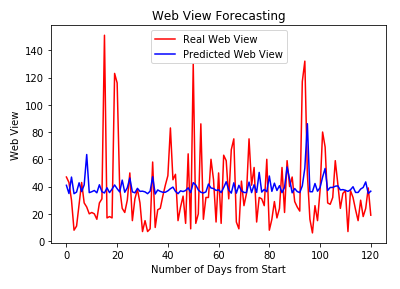
**Hidden**

**Input**

**Exploratory Data Analysis**

**Figures and Tables**

Web Traffic Time series dataset consists of two data-source. One is the features space we want to explore, that is the traffic time series for each article of Wikipedia. This data can be accessed by interfacing with an API, where one can query the traffic time series of a given article over a given time window, down to a given precision. However, in order to prototype our project as quickly as possible, we used as a working basis the data made available in the Kaggle competition “Web Traffic Time Series Forecasting” (https://www.kaggle.com/c/web-traffic-time-series-forecasting). It only spans over a limited selection of articles (Wikipedia 1,45,000 articles), and is limited to daily traffic recordings, from January 2015 to December 2016. Exploring some of these time series yields the following figure:



**Data Modelling**

**1) RNN**

**2) LSTM**

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behaviour. All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.

All RNNs have feedback loops in the recurrent layer. This lets them maintain information in 'memory' over time. But it can be difficult to train standard RNNs to solve problems that require learning long-term temporal dependencies. This is because the gradient of the loss function decays exponentially with time (called the vanishing gradient problem).LSTM networks are a type of RNN that uses special units in addition to standard units . LSTM units include a 'memory cell' that can maintain information in memory for long periods of time. A set of gates is used to control when information enters the memory, when it's output, and when it's forgot. This architecture lets them learn longer-term dependencies. GRUs are similar to LSTMs, but use a simplified structure. They also use a set of gates to control the flow of information, but they don't use separate memory cells, and they use fewer gates.

**Findings and Suggestions**

1) You can also use this for ‘N’ no. of website prediction for tracking of the no. of servers you need for the future use.

2) You can even implement for price estimation of gold price for next few months or years depends on how you process your model and dataset.

3) Similarly, for shopping also. In sense in which month and season the sales can be more.

**Conclusion:**

This project is all about the estimation of the servers to be maintained by a company to provide a better performance of the websites without being crashed due to over load of users visiting the website. Here, LSTM and RNN gives the actual graph for the next few months that the user visits the particular website in terms of percentage. So, that the company will have good cost reduction. In the last decade, the analysis and prediction of network traffic has become a subject of continuous research in various sub-fields of computer networks. Innumerable number of researchers have been implemented an effective network traffic algorithm for the analysis and the prediction of network traffic.