DATA 598 PROJECT PROPOSAL

Anuhya Bhagavatula

TOPIC: Web Traffic Time Series Analysis and Forecasting

DATASET LINK:

https://www.kaggle.com/competitions/web-traffic-time-series-forecasting/data

DATASET SIZE:

The dataset consists of approximately 145k time series.

DATASET DESCRIPTION:

Each of these time series represent a number of **daily** views of different **Wikipedia** articles, starting from **July, 1st, 2015** up until **December 31st, 2016**. The dataset has 804 columns – except the first column, each column represents a date and the daily traffic for that particular Wikipedia page. The first column contains the name of the page, the language of the page, type of access and agent

CONTEXT AND BACKGROUND:

Analysis and forecasting web traffic has many applications in various areas. It is a proactive approach to provide secure, reliable and qualitative web communication. Web traffic can be defined as the amount of data sent and received by visitors to a website. In recent years, emphasis on how to predict traffic of web pages has increased significantly. Predicting web traffic can help web site owners in many ways including:

(a) determining an effective strategy for load balancing of web pages residing in the cloud
(b) forecasting future trends based on historical data
(c) understanding the user behaviour.

Wikipedia is a multilingual free content online encyclopaedia written and maintained by a community of volunteers through a model of open collaboration, using a wiki-based editing system. Wikipedia grants open access to all traffic data and provides lots of additional (semantic) information in a context network besides single keywords. Wikipedia is often used for deep topical reading.

GOALS:

1. Grouping the data based on the language of the page and seeing if there exist any interesting patterns in web traffic based on language patterns. (ex: English, French, Chinese)
2. Forecasting future traffic for each page as well as for each language of the web pages as a group.

I am interested in this project as it helps me understand the underlying principles of timeseries forecasting by applying them on a real world web traffic model. I believe that by understanding this I can also use such models in various other applications such as vehicle traffic forecasting, network packet forecasting etc.

CITATIONS:

Source where data was found:

1. https://forecastingdata.org/
2. https://zenodo.org/record/4656075#.YlikZsjMLb0

Original Source:

https://www.kaggle.com/c/web-traffic-time-series-forecasting

Other Resources:

1. N. Petluri and E. Al-Masri, "Web Traffic Prediction of Wikipedia Pages," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 5427-5429, doi: 10.1109/BigData.2018.8622207.
2. Kämpf M, Tessenow E, Kenett DY, Kantelhardt JW. The Detection of Emerging Trends Using Wikipedia Traffic Data and Context Networks. *PLoS One*. 2015;10(12):e0141892. Published 2015 Dec 31. doi:10.1371/journal.pone.0141892