MASTER OF SCIENCE IN DATA SCIENCE
UNIVERSITY of WASHINGTON

Wikipedia Article Web Traffic Time Series Analysis

Spring 2022 Data 598 Project Proposal

Submitted To: Kellie Willis, Ph.D.

Submit Date: April 14, 2022

Submitted By: Aaliyah Hänni

Table of Contents

Overview

Wikipedia is a free online encyclopedia that has over 6 million articles accessed by users all over the world (Wikimedia Foundation, 2022). This project proposal outlines the exploration of web traffic for a sample of approximately 145,000 Wikipedia articles over a two year period. The goal for this project is to forecast the number of page views over time, and to explore potential relationships between page views, traffic types and article topics.

Project Description

1. ### Wikipedia Article Page Views Over Time

   Using daily web traffic information for a sample of Wikipedia articles, we will explore several key properties, such as identifying potential trends or seasonality in the total number of page views of varying articles over time. We will then generate forecast predictions of web traffic for the different articles, and compare predictions against the true number of page views.

2. ### Wikipedia Article Page Views Over Time by Traffic Type

   For a given Wikipedia article site visit there are three types of traffic that Wikipedia measures: mobile, desktop, and spider. Mobile and desktop traffic are site visits from individual users through different interfaces, while spider traffic is access to a site from a web-crawler that was developed to collect data. For this analysis, we will explore differences between the types of traffic in Wikipedia articles over time, and identify the types of relationships that exist between them.

3. ### Wikipedia Topics Page Views Over Time

   There are natural groupings of Wikipedia articles by topic, such as articles about flowers, sports, or historical events. For this analysis, we will explore the page traffic over time for varying Wikipedia topics. To determine a Wikipedia article topic, we will develop a simple classification model to group articles based on tokenized article titles. Then we will explore relationships in the aggregated sum of page views over time for the varying topics.

Data

The data set used for this analysis will be the Web Traffic Time Series Forecasting provided by Google Inc, and hosted on Kaggle (Google Inc., 2017). This dataset contains the daily count of unique user visits to 145,063 Wikipedia articles from July 1st, 2015 to September 10th, 2017. These articles are in multiple languages, have varying grades (i.e article content assessment score), and contain multiple topics. The dataset also groups the counts of the number of articles accessed by each traffic type: all, mobile, desktop, and spider.

Context and Background

Personal Motivation

The initial motivation for this analysis was to obtain more experience working with large datasets and to compete in my first Kaggle Competition. This specific Kaggle Competition is closed, but Kaggle members are still able to create their own submission to obtain a personal score on how well their forecasting predictions compare to others. Experience competing in a competition using a large dataset, will allow me to grow my professional portfolio of projects that I have completed.

Forecasting and Modeling

The ability to develop a model that forecasts web traffic for a website has many benefits related to networking and resource allocation. Knowing if there are trends or seasonality in website traffic, allows web administrators to allocate the needed bandwidth to optimize performance, while minimizing costs.

Identifying Relationships

Identifying patterns in page views can provide interesting sociological information about potential trends or seasonality in the topics that people are interested in. This can be the beginning of motivation for additional research in why people are interested in specific topics. For example, in the article, ‘Why do people search Wikipedia for information on multiple sclerosis?’, researchers found peaks in google searches for multiple sclerosis correlated with celebrities mentioning it on television, or when new treatments were released (Brigo, et. al, 2018).

Limitations

The dataset used for this analysis does not differentiate between missing data and zero page views. To overcome this limitation, careful considerations need to be made for how to handle missing data, such as by performing missing value imputations or removing specific rows of data with too many missing values, as it could skew analysis.

Another concern about the data is the overall size of the dataset and lack of needed computation power. Opening the dataset in excel caused the program to crash several times. If this is an issue for analysis, a potential solution is to limit the number of websites used by random selection.

Generating the Wikipedia article topics by using the title generates another set of potential biases and issues. GIven that the titles are in different languages, it will make generalizations of the topic difficult to verify. Also, not all groupings of topics may be intuitive or ideal. A potential way to overcome this problem, is to incorporate an additional dataset that includes article summaries (Scheepers, 2018), and perform the classification algorithm on article summary rather than title.

Cited Sources

- Brigo, F., Lattanzi, S., Bragazzi, N., Nardone, R., Moccia, M., & Lavorgna, L. (2018, February 5). *Why do people search Wikipedia for information on multiple sclerosis?* Multiple Sclerosis and Related Disorders. Retrieved April 9, 2022, from https://www.sciencedirect.com/science/article/pii/S2211034818300476?casa_token=12JD_y_NnVIAAAAA%3AHs6WgvUlGzhk3kSQToSki32Pz5ksktGCqrNUKUZRe1bnbwqOT9TG_MaB_ZZaNGrd_T7SAgCHlw

- Google Inc. (2017, September 12). *Web traffic time series forecasting*. Kaggle. Retrieved April 9, 2022, from https://www.kaggle.com/competitions/web-traffic-time-series-forecasting/overview

- Scheepers, T. (2018, January 16). *Wikipedia Summary Dataset*. GitHub. Retrieved April 9, 2022, from https://github.com/tscheepers/Wikipedia-Summary-Dataset

- Wikimedia Foundation. (2022, April 5). *Size of Wikipedia*. Wikipedia. Retrieved April 9, 2022, from https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia