LIM Annie

MARCOCCIA Félix

VICHAIKIT Jean-Philippe

ZHOU Sébastien

## **Compétition Kaggle**

## **Web Traffic Time Series Forecasting**

## **Forecast future traffic to Wikipedia pages**

Julien Krywyk, Jean-François Bercher

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

<https://www.kaggle.com/c/web-traffic-time-series-forecasting/discussion/43795>

**Data Description**

The training 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. The leaderboard during the training stage is based on traffic from January, 1st, 2017 up until March 1st, 2017.

The second stage will use training data up until September 1st, 2017. The final ranking of the competition will be based on predictions of daily views between September 13th, 2017 and November 13th, 2017 for each article in the dataset. You will submit your forecasts for these dates by September 12th.

For each time series, you are provided the name of the article as well as the type of traffic that this time series represent (all, mobile, desktop, spider). You may use this metadata and any other publicly available data to make predictions. Unfortunately, the data source for this dataset does not distinguish between traffic values of zero and missing values. A missing value may mean the traffic was zero or that the data is not available for that day.

To reduce the submission file size, each page and date combination has been given a shorter Id. The mapping between page names and the submission Id is given in the key files.

## **File descriptions**

Files used for the first stage will end in '\_1'. Files used for the second stage will end in '\_2'. Both will have identical formats. The complete training data for the second stage will be made available prior to the second stage.

* **train\_\*.csv** - contains traffic data. This a csv file where each row corresponds to a particular article and each column correspond to a particular date. Some entries are missing data. The page names contain the Wikipedia project (e.g. en.wikipedia.org), type of access (e.g. desktop) and type of agent (e.g. spider). In other words, each article name has the following format: 'name\_project\_access\_agent' (e.g. 'AKB48\_zh.wikipedia.org\_all-access\_spider').
* **key\_\*.csv** - gives the mapping between the page names and the shortened Id column used for prediction
* **sample\_submission\_\*.csv** - a submission file showing the correct format

La compétition Kaggle “Web Traffic Time Series Forecasting : Forecast future traffic to Wikipedia pages” est disponible sur ce lien : <https://www.kaggle.com/c/web-traffic-time-series-forecasting/>.

Le but de cette compétition est de prédire le nombre de visites sur des articles de pages Wikipédia.

Nous avons à notre disposition différents fichiers :

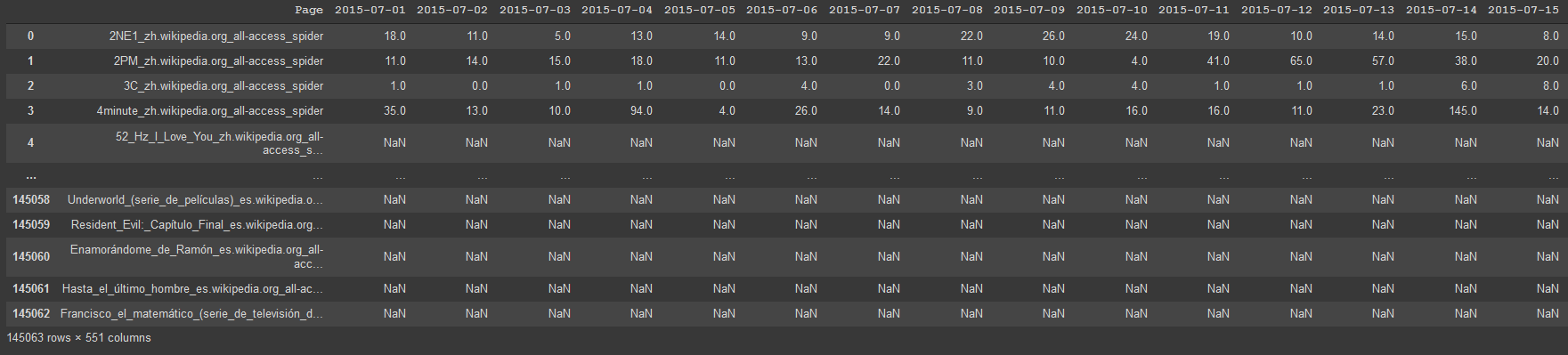
* train\_1.csv et train\_2.csv : contiennent les données de fréquentation avec un article par ligne et une date par colonne. Le nom de la page Wikipédia est sous la forme “nom”\_“projet”\_”type d’accès”\_”type d’agent”.
* key\_1.csv et key\_2.csv : donnent la correspondance entre le nom des pages et leur id, qui nous sera utile pour générer nos soumissions
* sample\_submission\_1.csv et sample\_submission\_2.csv : montrent le format des soumissions pour la compétition Kaggle

Les données d’entraînement ont environ 145k séries temporelles. Chacune correspond à un nombre de visites par jour sur un article Wikipédia.

Les données sont séparées en deux étapes :

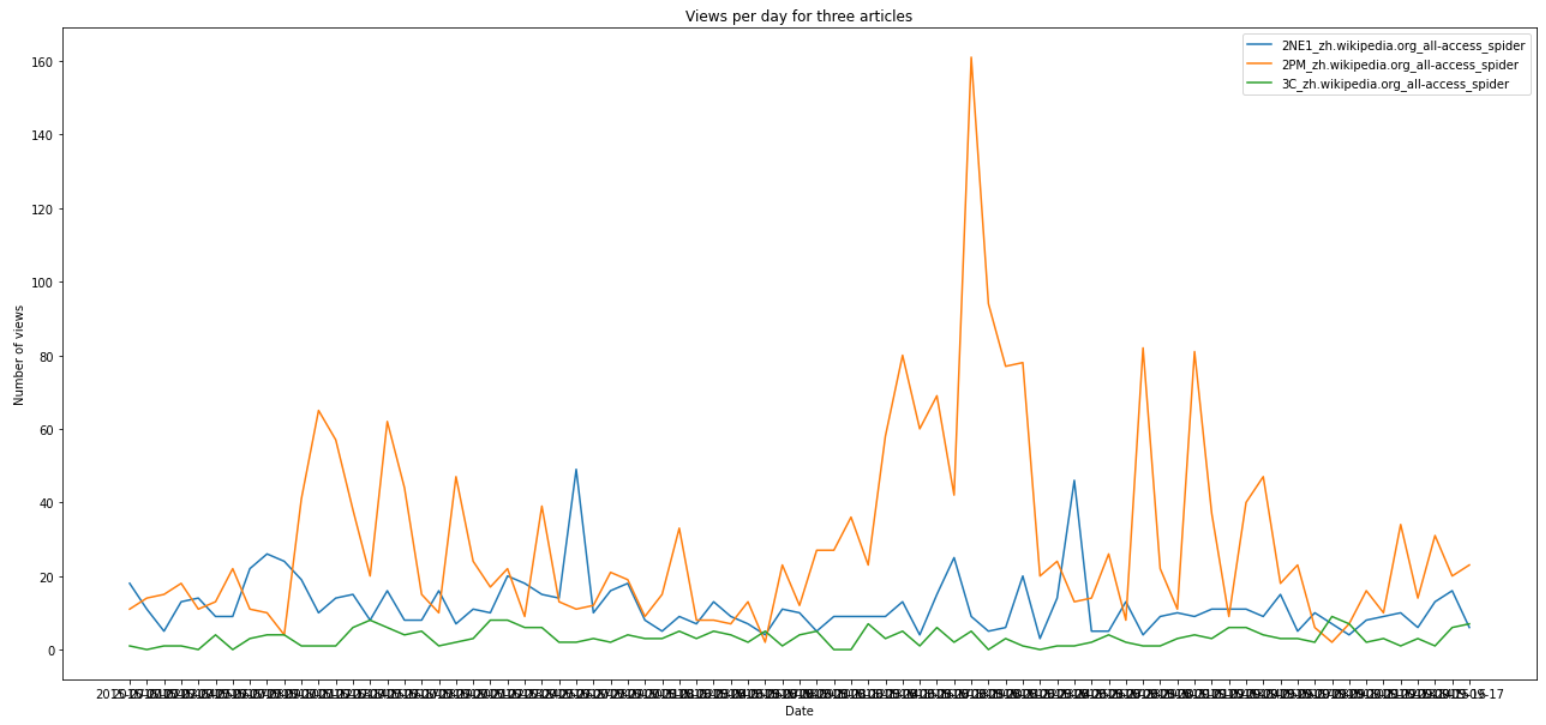
* La première étape est du 1er Juillet 2015 au 31 Décembre 2016. Un classement était fait pour prédire le nombre de vues du 1er Janvier 2017 au 1er Mars 2017.
* La deuxième étape est du 1er Juillet 2015 au 10 Septembre 2017. Ces données devaient servir pour les prédictions de vues du 13 Septembre 2017 au 13 Novembre 2017.

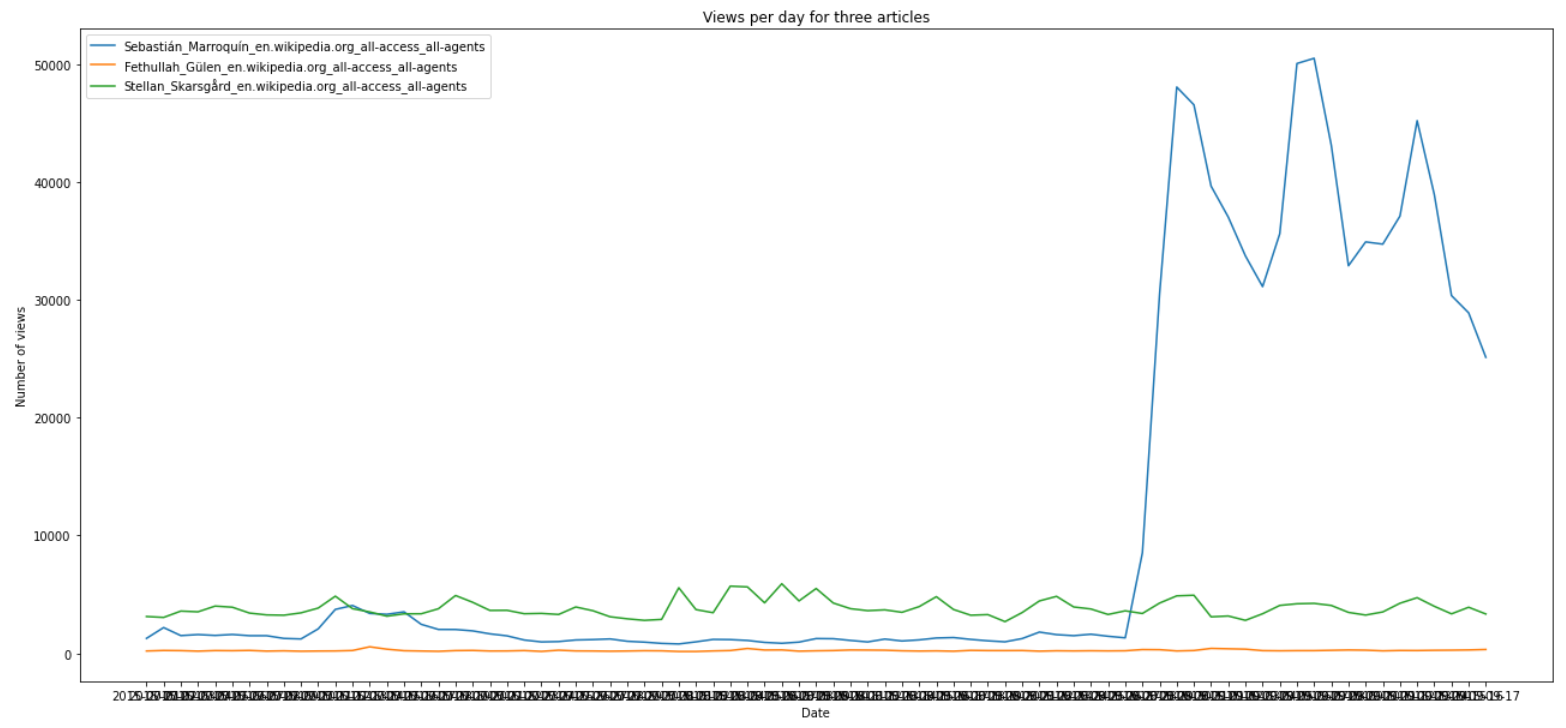
Nous n’allons traiter que le fichier train\_1.csv qui nous fournit suffisamment de données pour faire notre entraînement.

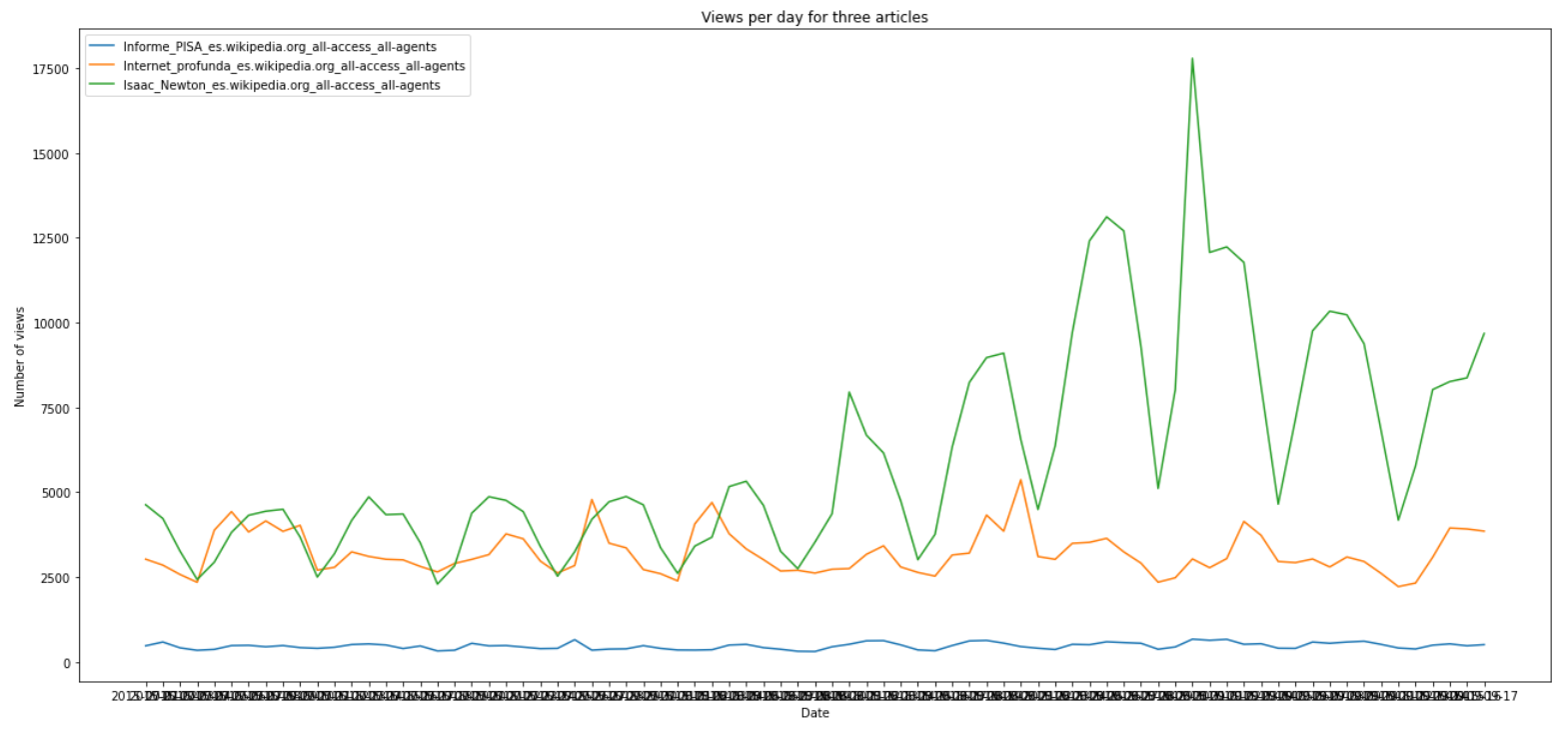


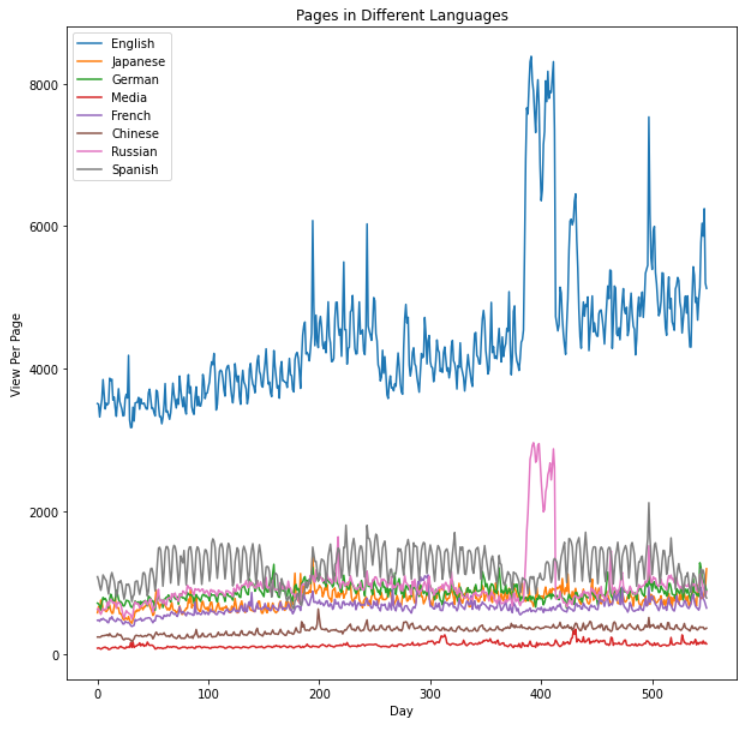
Nous avons donc le nombre de vues sur 145063 articles Wikipédia sur 550 jours, du 1er Juillet 2015 au 31 Décembre 2016.

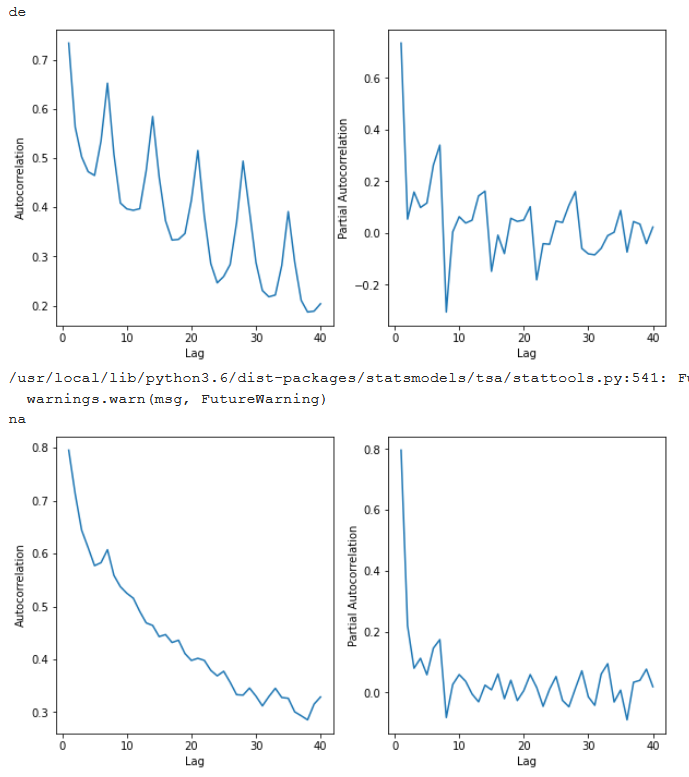
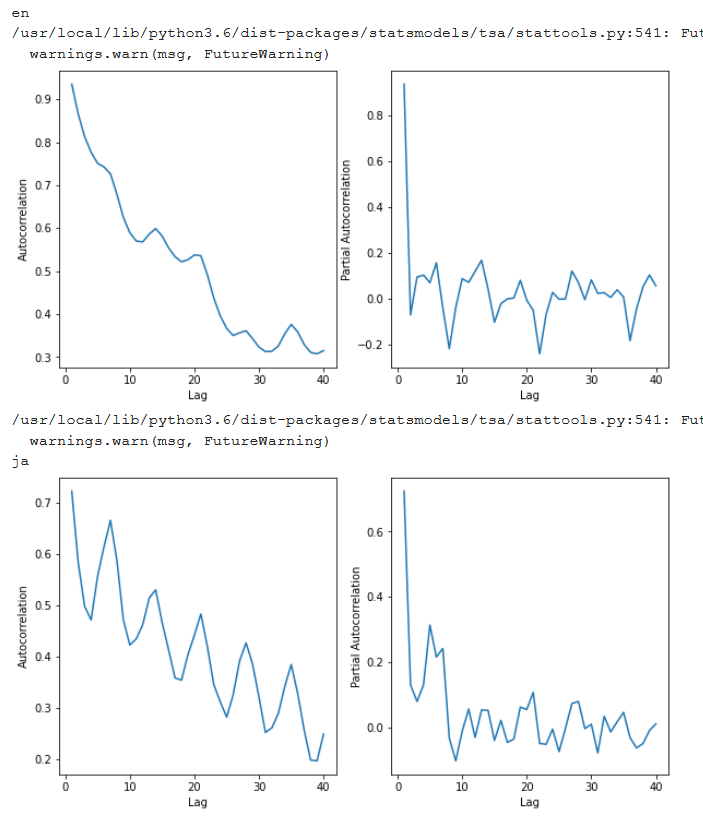
Nous pouvons observer qu’il y a de nombreuses données manquantes “NaN”. Malheureusement la source du dataset n’a pas fait de distinction et les données en “NaN” peuvent à la fois correspondre à des données dont le trafic est à 0 ou alors des données qui n’étaient pas disponibles à ce jour.

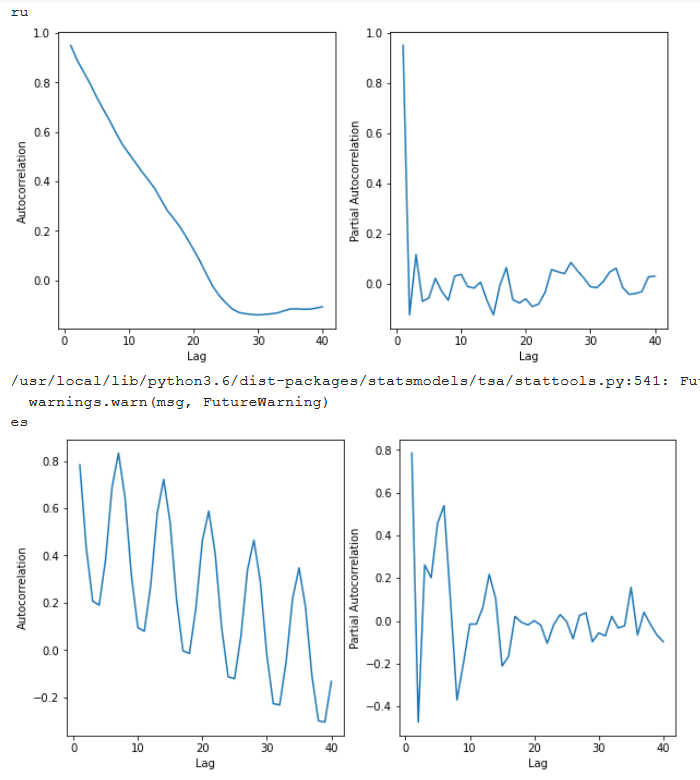
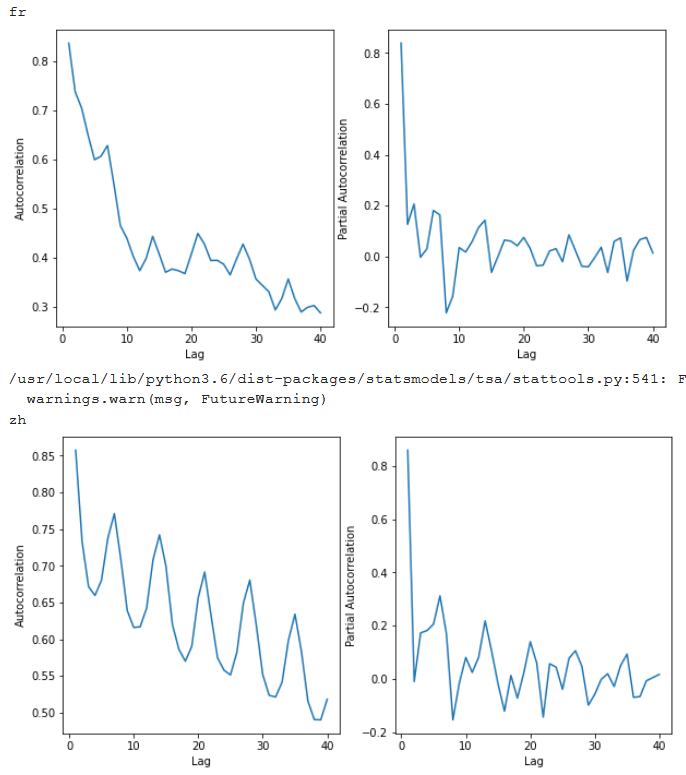




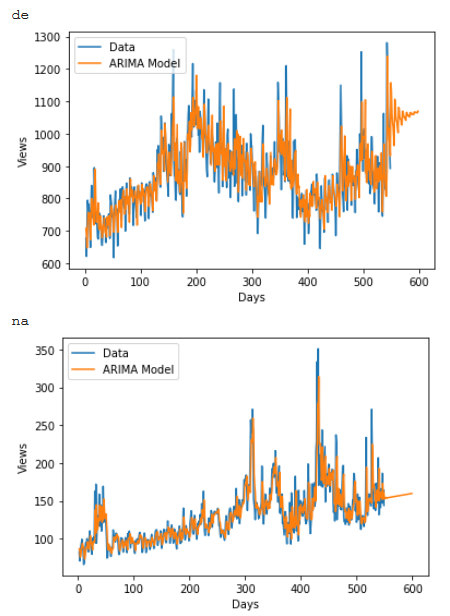
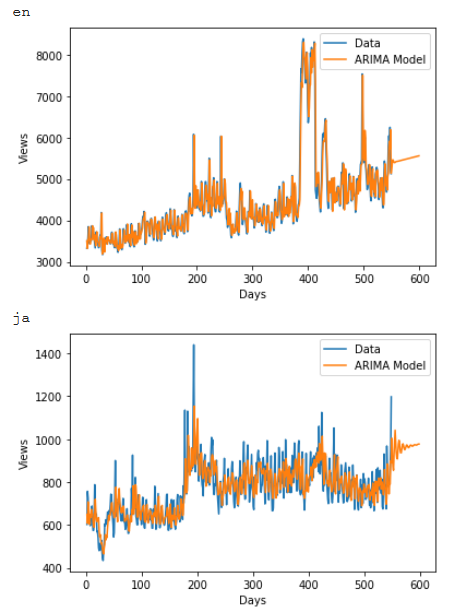


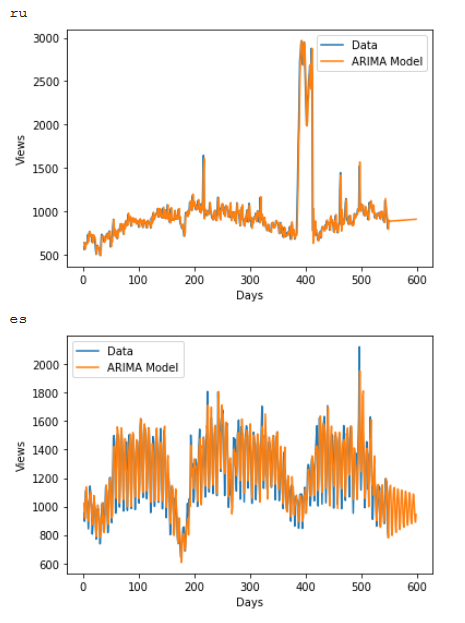
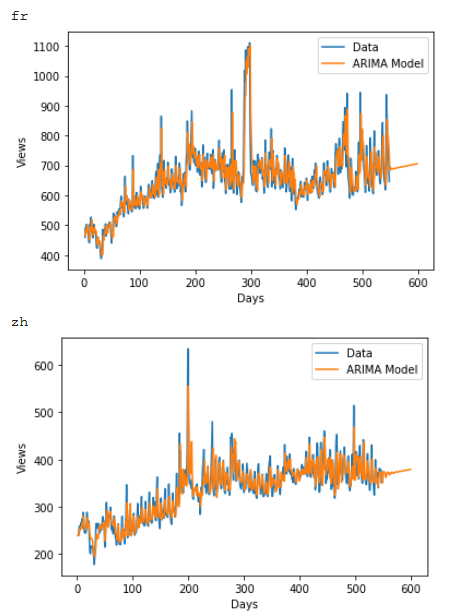


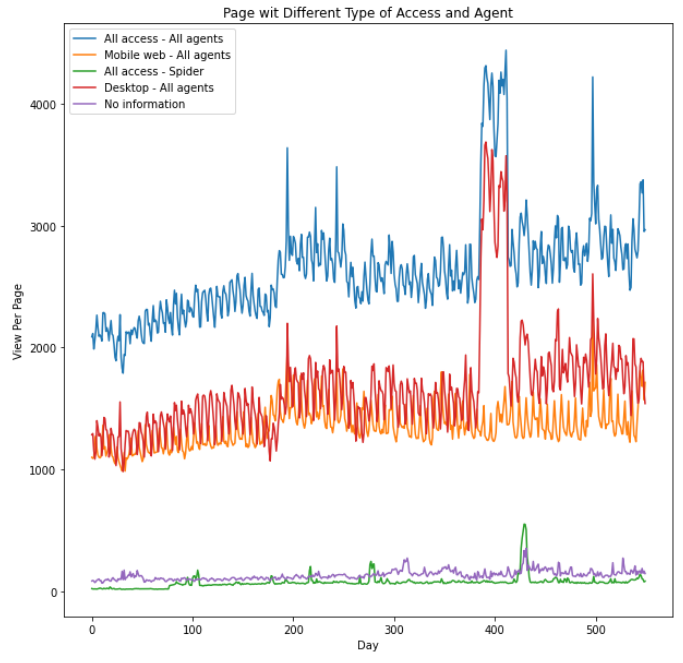


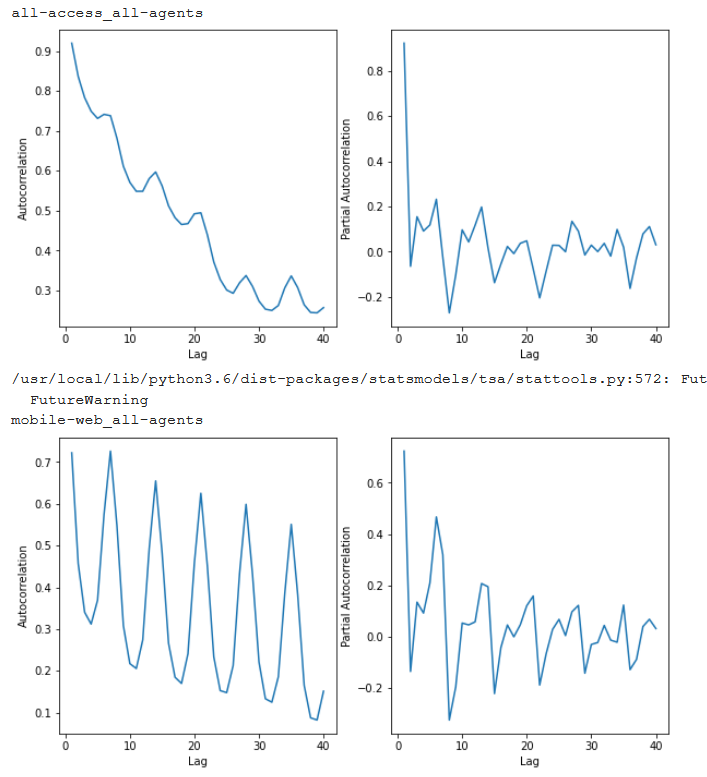


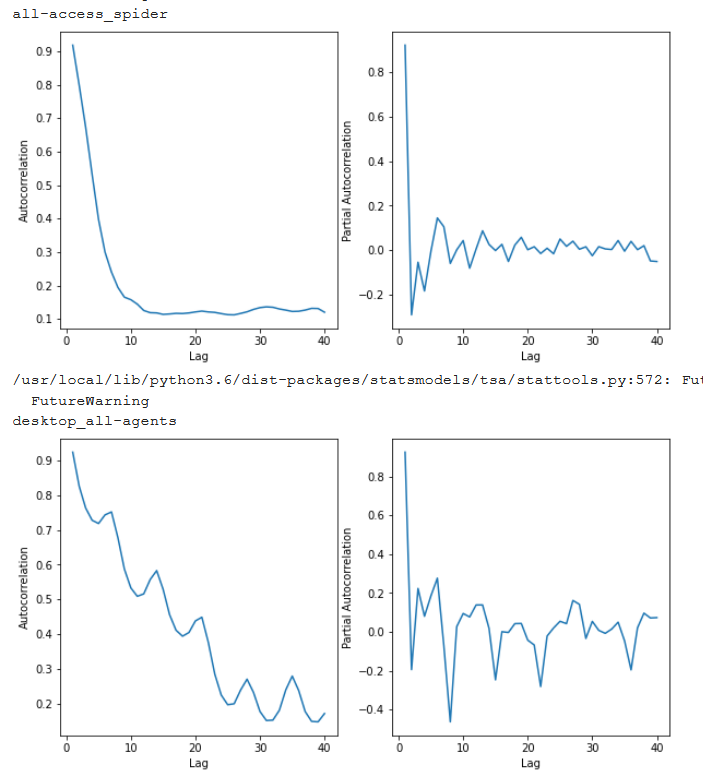
ARIMA model

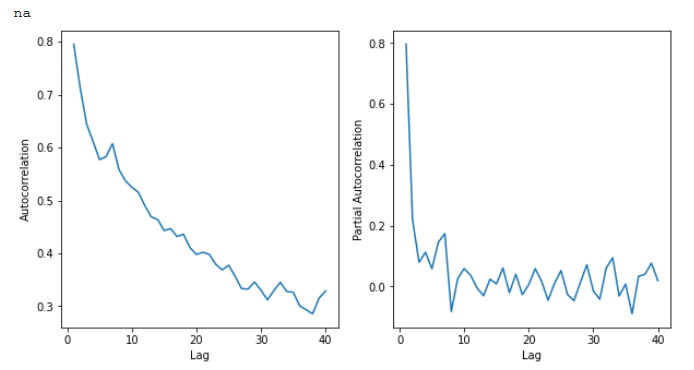




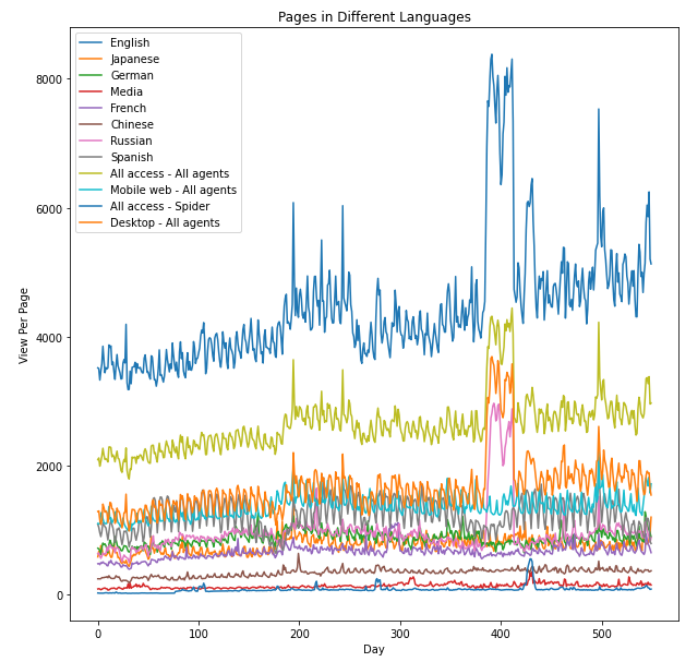




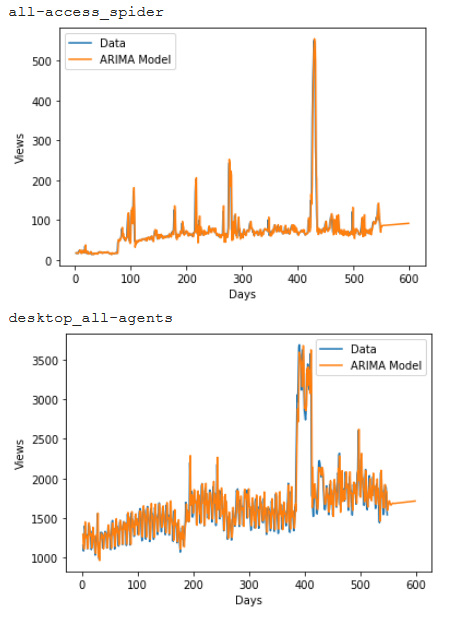
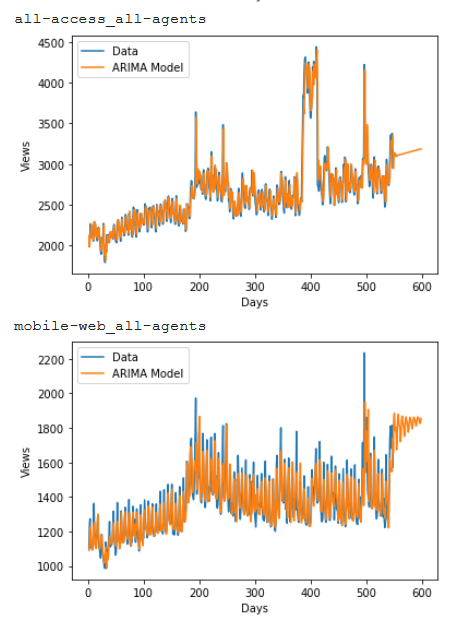




En combinant les langues et les types d’accès-agent :



on a en plus :



<https://www.imo.universite-paris-saclay.fr/~goude/Materials/time_series/cours6_ARIMA.pdf>

Modèle ARIMA (Autoregressive Integrated Moving Average)

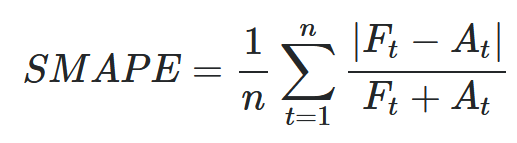
Le modèle ARIMA, formalisé par Box et Jenkins en 1976, cherche à déterminer chaque valeur de la série en fonction des valeurs qui la précède.

Le modèle est composé de trois parties paramétrées par les indices p, d, q tel que ARIMA (p, d, q) :

* Autoregression (AR) / p : Les données passées sont utilisées pour calculer un modèle de régression pour les données futures. Le paramètre p indique le nombre de termes auto-régressifs c’est-à-dire le décalage (lags).
* Integrated (I) / d : Il correspond au paramètre de différenciation pour permettre à la série temporelle de devenir stationnaire, c’est-à-dire que les valeurs de données sont remplacées par la différence entre les valeurs actuelles et les précédentes. Il est rare de rencontrer d>2, en effet il est dangereux de sur-différencié un processus.
* Moving Average (MA) / q : C’est le nombre de moyennes mobiles. On considère que la meilleure estimation est représentée par la moyenne pondérée d’un certain nombre de valeurs antérieures, permettant de lisser les données.

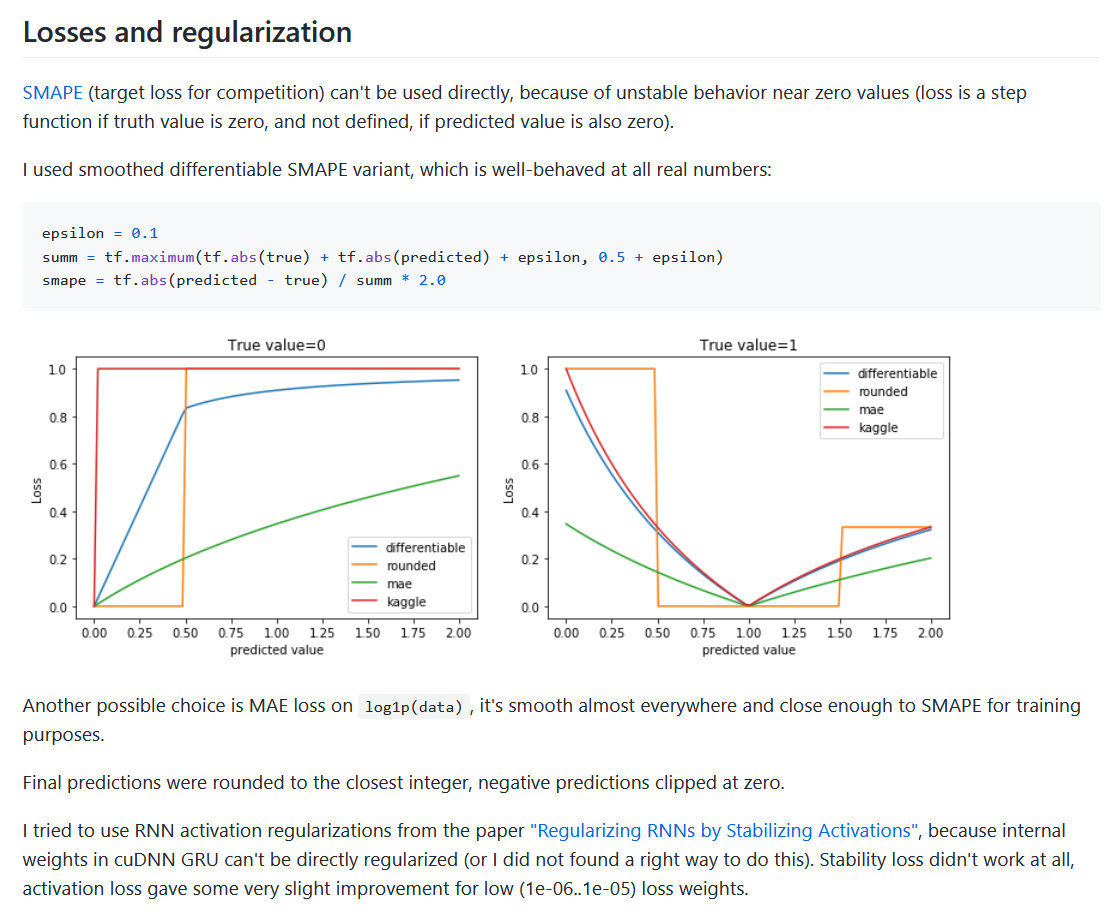
La mesure de performance utilisée pour calculer l’erreur dans cette compétition Kaggle est le SMAPE (Symmetric Mean Asolute Percentage Error). Cette métrique a pour particularité d’ignorer les cas particuliers et est invariant si l’on redimensionne linéairement les données.

Elle se calcule telle que :



avec Ft la valeur prédite, et At la valeur réelle

Solution gagnante : <https://github.com/Arturus/kaggle-web-traffic/blob/master/how_it_works.md>



## **Seq2Seq Model for Temperature Prediction in Firing Furnace Process**