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

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 :

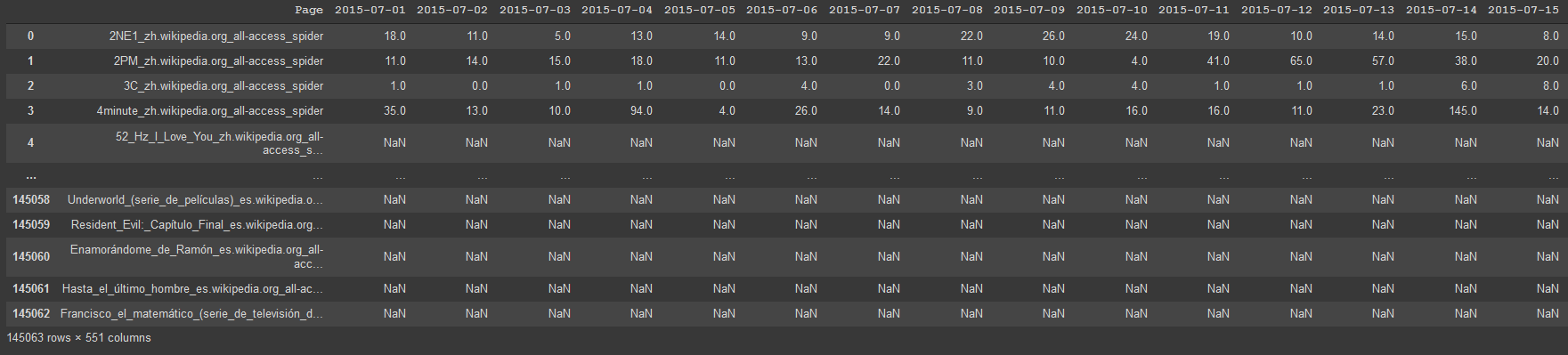
* train\_1.csv et train\_2.csv : contiennent les données de fréquentation avec un article par ligne et une date par colonne. Le nom de la page Wikipédia est sous la forme “projet Wikipédia”\_”type d’accès”\_”type d’agent”.
* key\_1.csv et key\_2.csv : donnent la correspondance entre le nom des pages 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 :

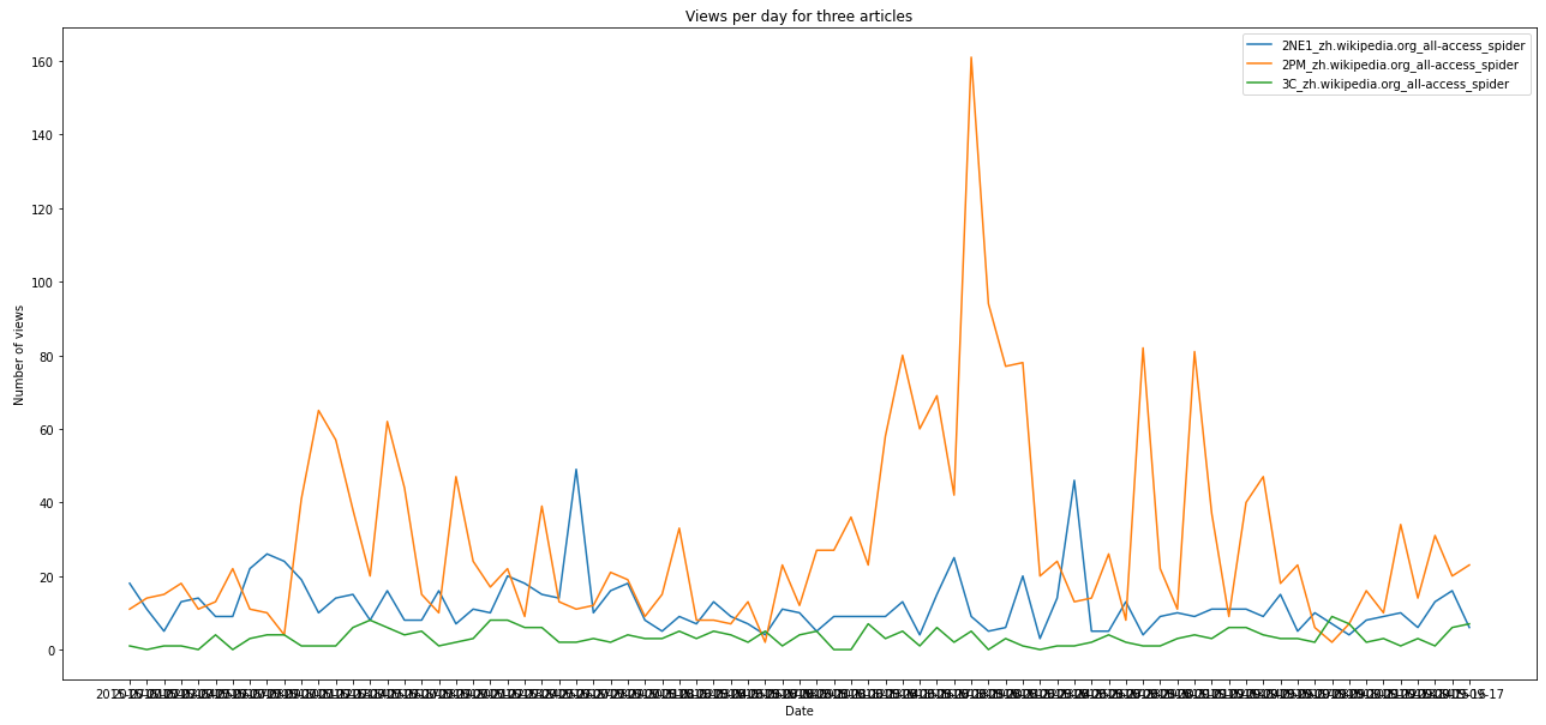
* La première étape est du 1er Juillet 2015 au 31 Décembre 2016. Un classement était fait pour prédire le nombre de vues du 1er Janvier 2017 au 1er Mars 2017.
* La deuxième étape est du 1er Juillet 2015 au 10 Septembre 2017. Ces données devaient servir pour les prédictions de vues du 13 Septembre 2017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

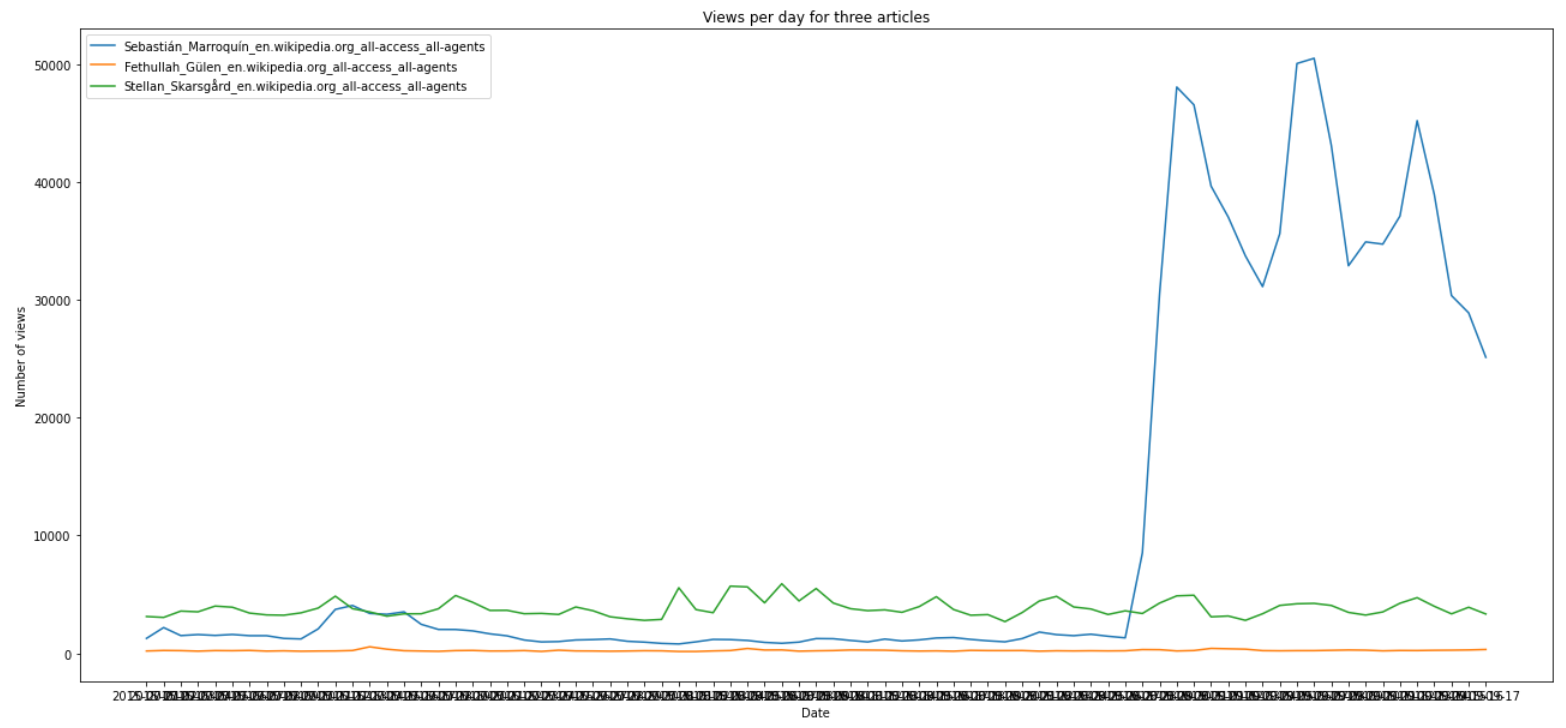
Nous n’allons traiter que le fichier train\_1.csv qui nous fournit suffisamment de données pour faire notre entraînement.

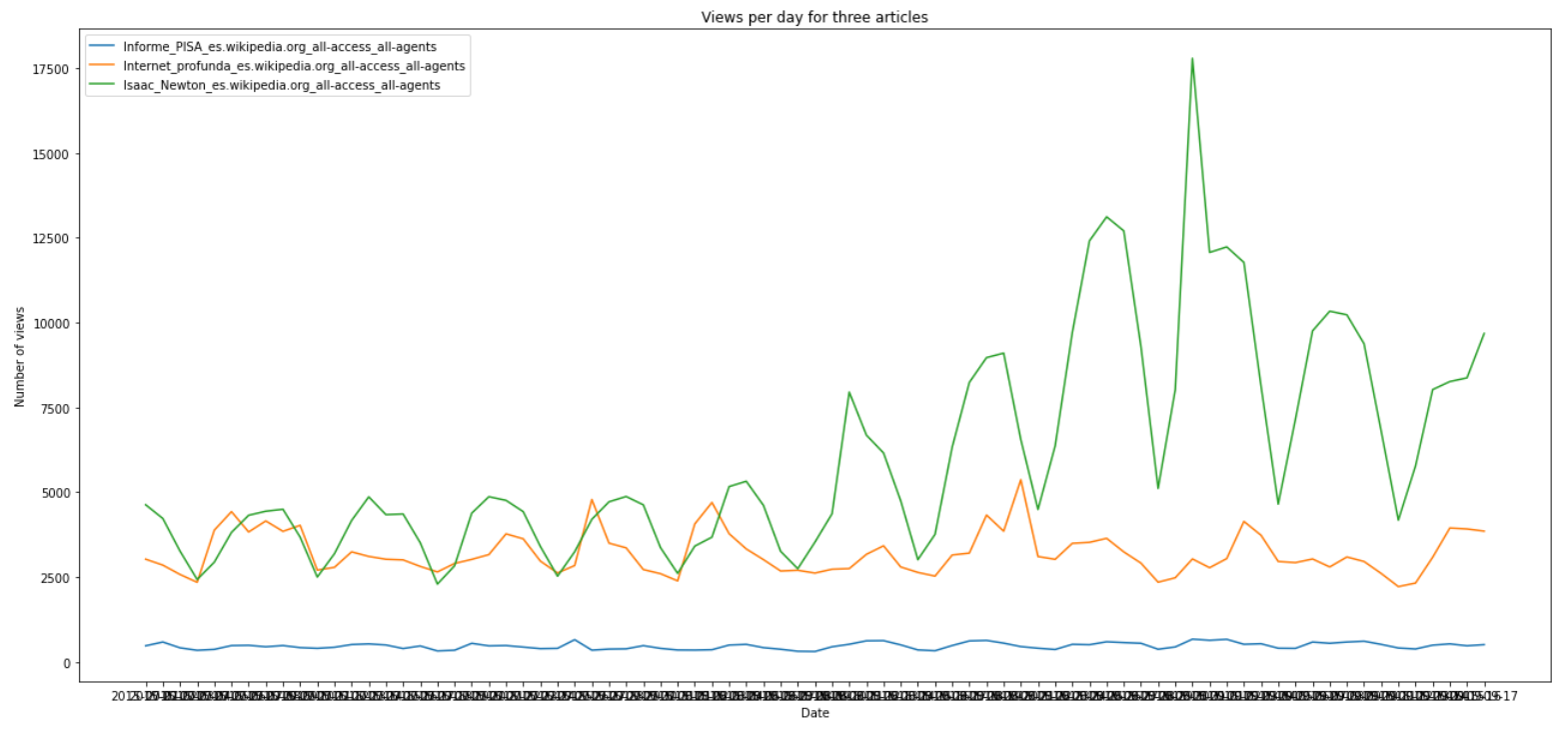


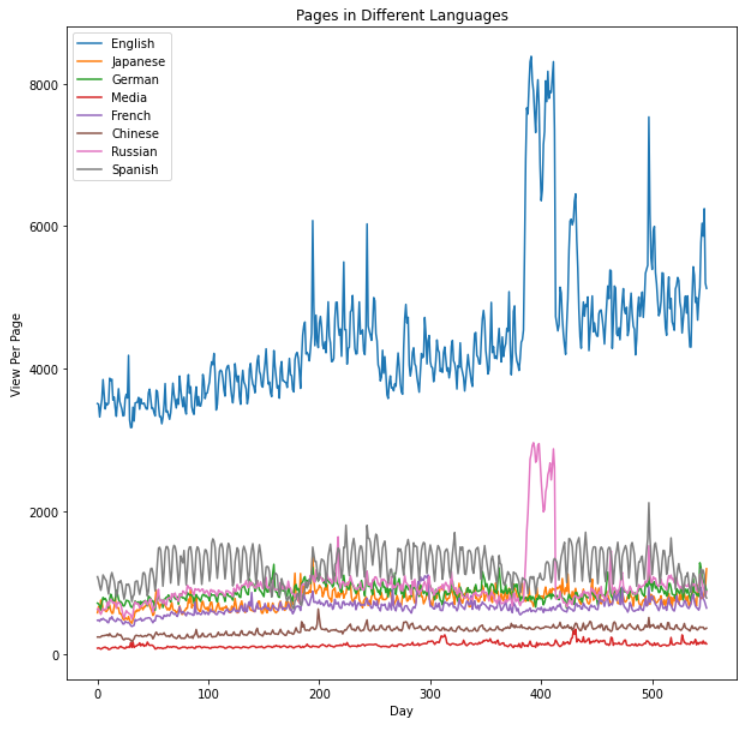
Nous avons donc le nombre de vues sur 145063 articles Wikipédia sur 550 jours, du 1er Juillet 2015 au 31 Décembre 2016.

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.



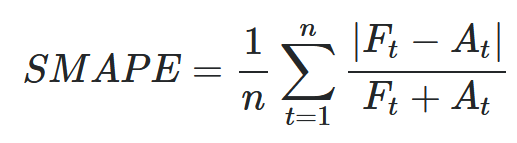






La mesure de performance utilisée pour calculer l’erreur dans cette compétition Kaggle est le SMAPE (Symmetric Mean Asolute Percentage Error). Cette métrique a pour particularité d’ignorer les cas particuliers et est invariant si l’on redimensionne linéairement les données.

Elle se calcule telle que :



avec Ft la valeur prédite, et At la valeur réelle