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.

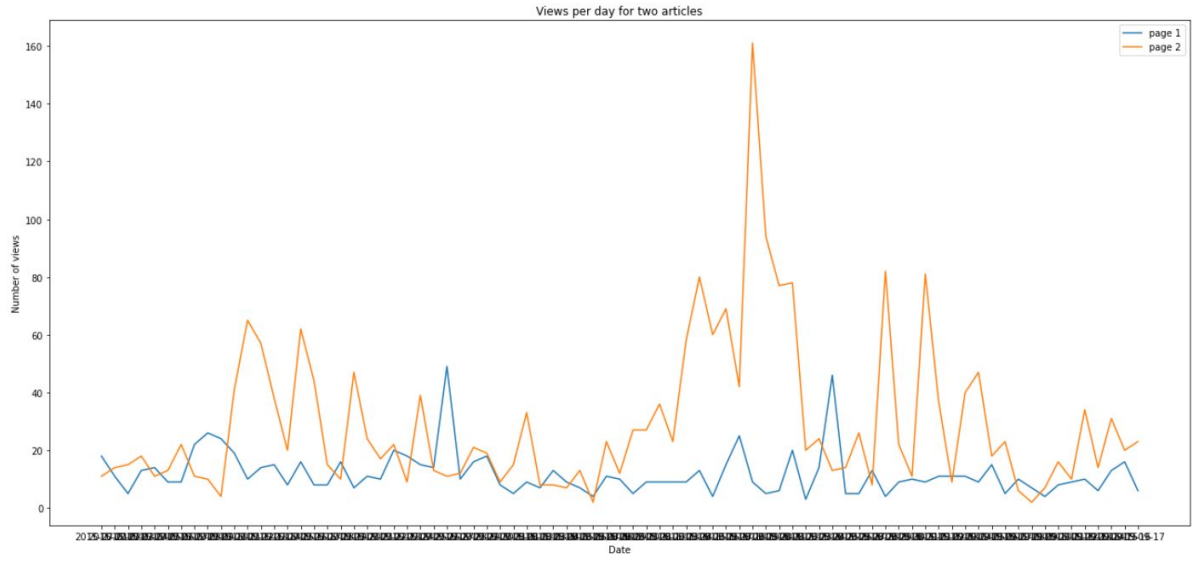
Nous n’allons traiter que le fichier train_1.csv qui nous fournit suffisamment de données pour faire notre entraînement.

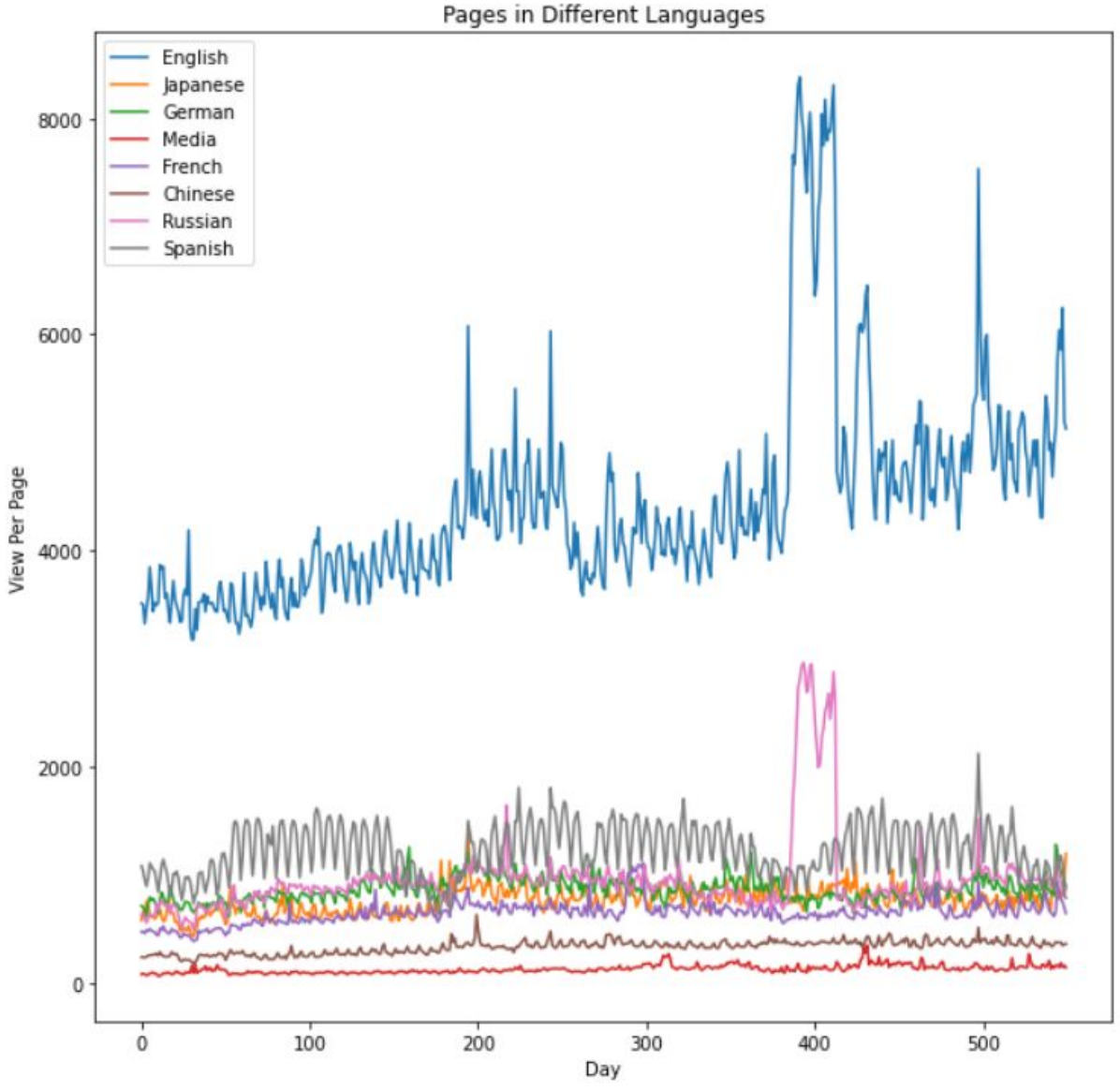
	Page	2015-07-01	2015-07-02	2015-07-03	2015-07-04	2015-07-05	2015-07-06	2015-07-07	2015-07-08	2015-07-09	2015-07-10	2015-07-11	2015-07-12	2015-07-13	2015-07-14	2015-07-15
0	2NE1_zh.wikipedia.org_all-access_spider	18.0	11.0	5.0	13.0	14.0	9.0	9.0	22.0	26.0	24.0	19.0	10.0	14.0	15.0	8.0
1	2PM_zh.wikipedia.org_all-access_spider	11.0	14.0	15.0	18.0	11.0	13.0	22.0	11.0	10.0	4.0	41.0	65.0	57.0	38.0	20.0
2	3C_zh.wikipedia.org_all-access_spider	1.0	0.0	1.0	1.0	0.0	4.0	0.0	3.0	4.0	4.0	1.0	1.0	1.0	6.0	8.0
3	4minute_zh.wikipedia.org_all-access_spider	35.0	13.0	10.0	94.0	4.0	26.0	14.0	9.0	11.0	16.0	16.0	11.0	23.0	145.0	14.0
4	52_Hz_I_Love_You_zh.wikipedia.org_all-access_s...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
145058	Underworld_(serie_de_películas)_es.wikipedia.o...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
145059	Resident_Evil:_Capítulo_Final_es.wikipedia.org...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
145060	Enamorándome_de_Ramón_es.wikipedia.org_all-acc...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
145061	Hasta_el_último_hombre_es.wikipedia.org_all-ac...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
145062	Francisco_el_matemático_(serie_de_televisión_d...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

145063 rows × 551 columns

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.


Views per day for two articles
page 1
page 2
Number of views
Date


Pages in Different Languages
English
Japanese
German
Media
French
Chinese
Russian
Spanish
View Per Page
Day

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 :

$$SMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{F_t + A_t}$$

avec F_t la valeur prédite, et A_t la valeur réelle