La donnée et son traitement

Dictionnaire de données :

Nom de la feature	Description
id	Identifiant du post de l'utilisateur
user	Adresse mail anonymisée de l'utilisateur
lat	Latitude
long	Longitude
tags	Tags du post
title	Titre du post
date_taken_{minute, hour, day, month, year}	Date de prise de la photo (en 5 features)
date_upload_{minute, hour, day, month, year}	Date de publication de la photo (en 5 features)

Nous disposons d'un jeu de données de dimensions 83837x16 provenant d'une extraction de base de données du réseau social Flickr. Ce *dataset* contient un certain nombre de défauts qui seront explicités par la suite.

Des données dupliquées

Effectivement, bien que le nombre de posts soit de 83837, seul 15175 d'entre eux sont uniques. Il y a donc une duplication de la donnée de l'ordre de 68635 posts, ce qui représente tout de même 81.89% de l'ensemble des données.
En plus d'avoir un impact sur la redondance de l'information, ces duplications entrainent des statistiques erronées. En effet, avant la suppression de ces données, le nombre de poste moyen était cinq fois plus élevé qu'après, il en va de même pour la médiane du nombre de posts qui perd 25%.

Des données temporelles non standardisées

L'ensemble de nos entrées contiennent des informations sur la date et l'heure de prise et de post. Néanmoins, ces données sont stockées dans des colonnes séparées et ne sont donc pas sous la forme d'objet *datetime*, ce qui complexifie leurs manipulations. La première chose qui a donc été faite a été de 'réunir' ces données dans deux nouvelles features correspondant tout deux au format %d/%m/%Y %H:%M:%S.
Cette transformation a fait apparaitre des incohérences quant aux dates et horaires de certaines données. En effet, certaines d'entre elles, ont une date de publication inférieure à la date de prise, ce qui est théoriquement impossible. Il apparaît néanmoins que ces métadonnées peuvent être manuellement modifiées par l'utilisateur sans qu'aucune vérification ne soit faite par la plateforme, ce qui explique ces données. Celles-ci ne représentant qu'un volume d'environ 1%, nous avons décidé de supprimer les lignes correspondantes.

Enfin, en vue de son exploitation via des techniques de Machine Learning, les champs de dates ont été ‘catégorisés’ en fonction de leurs valeurs. Ainsi, deux nouvelles features ont été crée à partir de ces éléments.

Des données textuelles à traiter

Ce dataset contient deux features textuelles. La première concerne les tags, la seconde, le titre du post. Très vite, on peut remarquer que la forme de ces données est très fluctuante. En effet, il s’agit là encore d’une entrée utilisateur, ce qui suppose une infinité de forme de contenu. Néanmoins, certains patterns ou caractères spéciaux nous permettent d’isoler certaines portions du texte pour ainsi mieux le traiter par la suite. C’est ainsi qu’à partir de ces deux features, un total de cinq nouvelles features vont être crée comme suit :

- uploaded_via : nouvelle feature contenant la plateforme de post original, cette donnée est extraite depuis la feature tags.
- foursquare_venue : un id de foursquare
- links : l’ensemble des liens présents dans le texte
- file_title : le nom des fichiers uploadé
- people_tag : le nom des personnes taguées dans un post

Les données spatiales

Le jeu de données contient également des données spatiales. En effet, nous retrouvons la latitude et longitude pour l’ensemble de nos entrées. Aucune manipulation particulière n’a été nécessaire pour ces données.