2020

LIFPROJET – RC1 : Data Mining

AVALANCHES

Aymeric TOUCHE & Gérome FERRAND

Enseignant référent Mr Cazabet

Université Lyon I

Table des matières

I. Le sujet

A. Le choix

Nous avons choisi le sujet RC1 – Data Mining.

La recherche et l'exploitation de données sont des sujets qui nous ont tout de suite intéressés.

Nous avons choisi un support web pour notre application, car nous voulions progresser dans ce domaine.

L'un d'entre nous étant passionné de montagne, nous nous sommes très vite orientés vers la récupération de données concernant les avalanches, principalement dans les Alpes. Ce sujet n'ayant jamais été traité lors des sessions précédentes, cela nous a confortés dans notre choix.

L'idée est ici de donner un maximum d'informations sur les avalanches recensées de la manière la plus simple possible et d'en ressortir leurs principales caractéristiques grâce un une base de données fournie.

B. Nos objectifs

Nos objectifs de départs étaient les suivants :

- Collecter des données sur le net
- Analyser les caractéristiques principales d'une avalanche
- Cartographier les avalanches en fonction de ces caractéristiques
- Établir une analyse graphique via des statistiques
- Faire ressortir des indicateurs et en dégager un sens
- Rester ouverts aux possibilités offertes par les jeux de données collectées

II. Le travail réalisé

A. La recherche de données

Dans un premier temps, nous avons recherché des jeux de données sur internet et notamment sur Google DataSearch.

- 1er jeu de données : un fichier geojson rapportant une photo-interprétation des avalanches dans les Alpes. Des personnes ont étudié la forme des terrains pour en déduire si une avalanche y avait eu lieu.

- 2e jeu de données : Fichier csv par année recensant les accidents mortels ou non dus aux avalanches, toujours dans les Alpes.

- 3e jeu de données : Un site internet qu'il a fallu parcourir via un programme afin de récupérer les données intéressantes et similaires au jeu de données précédent et de créer un fichier geojson.

Nous avons pris contact avec le propriétaire du premier jeu de données (l'institut régional de recherche pour l'agriculture, l'alimentation et l'environnement) afin d'obtenir des informations complémentaires, malheureusement les délais pour obtenir ces données et signer une licence de mises à disposition étaient trop longs.

B. La création d'un site web

Le site web a été développé via le Framework Slim PHP qui permet une architecture MVC et une gestion simple et efficace des requêtes HTTP.

Son utilisation n'est pas essentielle à un tel projet, néanmoins nous voulions aller plus loin dans la programmation web.

Le choix du moteur de templates twig nous a permis de factoriser notre code afin d'éviter toutes formes de redondances, c'est un outil qui s'utilise naturellement avec Slim PHP.

Le site web a par la suite été déployé sur un serveur mis à disposition par Mr Cazabet.

C. La carte interactive

Le site MapBox.com permet de créer un fond de carte que nous avons adapté au sujet, avec les reliefs. La carte interactive a été créée grâce à la bibliothèque MAPBOX GL JS qui est une librairie JavaScript.

Cet outil permet de créer différents layers traduisant les fichiers geojson préalablement créés ou récupérés. Ces layers sont ajoutés à la carte, les avalanches y sont alors représentées.

Des scripts en JavaScript ou JQuery sont implémentés afin d'afficher les différents layers selon ce que souhaite l'utilisateur.

D. Le scrapper

Toutes les parties collecte, formatage et analyse de données ont été réalisées en python.

Le scrapper a, dans un premier temps, été développé avec la bibliothèque Scrapy avec l'IDE Spyder. Or à la suite des conseils de notre entourage et dans un souci de découverte, nous avons migré le projet sur PyCharm. Par ailleurs, afin de pouvoir parcourir des sites web dont le contenu est géré dynamiquement, nous avons dû recourir aux solutions offertes par la lib. Selenium.

Ce programme nous a permis de réunir 2035 avalanches supplémentaires à notre base de données.

Le programme passe de page en page sur le site data-avalanche.org, et procède par étape :

- Il récupère la page HTML
- Sélectionne la zone contenant les données nécessaires à la formation du dataset
- Décode la chaine de caractère
- Charge les données dans un dictionnaire
- Ajoute les cases nécessaires à l'uniformisation de nos jeux de données
- Retire les propriétés qui portent peu d'intérêt pour notre utilisation des données
- Nettoie la propriété description de tous les caractères parasites (balise HTML, \n, \r, \t, etc.)
- Ajoute la ligne à une liste cache
- Transfert la liste dans un fichier json.

E. L'Analyse

Cette étude a été élaborée essentiellement en Python, grâce notamment aux modules pandas et matplotlib.

Pandas répondait parfaitement à nos besoins. En effet, cette bibliothèque python a été conçue pour la manipulation et l'analyse de données. Les dataframes et les séries se sont révélées particulièrement utiles pour exploiter les données. De plus, Pandas offre la possibilité d'extraire et de formater directement les données d'un fichier vers une dataframe ; il est également très efficace pour la gestion d'un important volume de donnée en procédant morceau par morceau.

MatPlotLib est le module python qui nous a permis de représenter les données sur des courbes, 'pie' et histogrammes en barres. Inspirée de MatLab, il s'agit probablement de la lib la plus célèbre de python dans la représentation de données.

Au travers de ce document, nous tentons d'extraire des statistiques les plus représentatives de la Réalité. Et ce en proposant dans un premier temps la répartition des avalanches en fonction de diverses caractéristiques : orientation, déclenchement, répartition géographique, temporelle, etc.

III. Organisation du travail

A. Répartition

Trois tâches ont rapidement été identifiées.

La première étant de trouver des données, la seconde de créer une application web.

La recherche de données a été réalisée en binôme, nous avons passé du temps à deux là-dessus.

Par la suite l'un de nous s'est occupé principalement de l'application web en elle-même et du traitement des données récupérées afin de les afficher correctement sur la carte et de rendre l'application interactive.

L'autre s'est chargé de créer un scrapper en python afin de parcourir le site data-avalanche.org, de récupérer les données concernant les avalanches cataloguées et de les traduire dans des fichiers geojson.

Ces fichiers ont ensuite servi de base de données pour l'application web, en plus de celles déjà mises à disposition directement sur d'autres sites.

B. Synchronisation

Un dépôt GitHub a été créé, chacun pouvait donc avancer sur l'implémentation des fonctionnalités qu'il avait à réaliser.

Au début du projet nous travaillions souvent ensemble à l'université ou chez nous. Tout le travail était réalisé en lors de ces séances fréquentes afin que chacun puisse participer à toutes les fonctionnalités.
En revanche depuis le début du confinement, chacun les deux aspects du projet est devenu assez hermétiques et nous travaillions sur nos parties respectives, la communication étant devenue plus difficile.

IV. Les défis

Nous avons dû nous former sur de nouvelles technologies connues ou non pour mener à bien ce projet.

Nous ne voulions pas faire un site web classique en tout HTML bien que cela aurait été possible et n'aurait rien changé au rendu, mais nous n'aurions pas progressé. Pour cela il a fallu se renseigner sur les Frameworks PHP existant et en sélectionner un adapté à la structure et l'envergure du projet.

Plusieurs jours ont été nécessaires afin de nous familiariser avec Slim PHP et ses techniques.

Au dernier moment nous avons découvert un conflit entre le serveur partagé hébergeant notre site et Slim PHP concernant les requêtes HTTP. Une solution de secours tout HTML a été conçue en quelques dizaines de minutes. Finalement nous avons pu gérer ce problème et conserver une architecture plus avancée.

La méthode d'affichage des données sur une carte interactive nous était totalement étrangère, et nos compétences en JavaScript et JQuery limitées. Nous nous sommes donc formés pendant des jours sur ces technologies. L'outil leaflet proposé a été remplacé par mapbox qui est plus complet et possède une documentation très fournie ainsi qu'un système d'hébergement.
Cette solution d'hébergement nous a été très utile concernant la partie photo-interprétation, les données étant trop volumineuses nous avons pu ainsi en déléguer le traitement ce qui a permis un gain de temps considérable au chargement de la carte.

Concernant le robot de collecte, il s'est avéré que l'une des librairies ne pouvait résoudre le problème. Dans un premier, il a fallu comprendre l'origine de la problématique. En tant que néophytes, nous avons longtemps pensé qu'il s'agissait d'une erreur de notre part. Or, en réalité la bibliothèque que nous utilisions ne pouvait pas gérer ce type de gestion de contenu web. Il a fallu ensuite rechercher une nouvelle technologie qui répondait à notre besoin.

Un autre aspect qui nous a posé un problème était la qualité de données à laquelle nous avons eu accès. Cela a eu tendance à nous freiner dans le développement du site. Bien qu'avoir tenté des organismes spécialisés, les temps de réponse furent particulièrement longs, voire inexistants (et dans tous les cas infructueux). Nous avons cependant bien agi en commençant à développer le crawler immédiatement après avoir demandé les données à l'association. Nous avons fait preuve d'anticipation, et cela nous a fait gagner un temps certain.

Enfin, comme cité précédemment, la crise sanitaire et le confinement que nous vivons ont été de réels freins à l'avancement de notre projet. Collaborer à distance aura constitué un défi majeur que nous avons relevé en modifiant notre stratégie de fonctionnement.