IMT Mines Alès
École Mines-Télécom

Rapport : Collecte et Stockage de données

Analyse de données météorologiques

Oumaïma BOUNHAR, Amine HADDAD, Jérémy LUCAS et Matthieu SFEIR
FIG 172

IMT Mines Alès
École Mines-Télécom

Table des matières

1. Introduction

L'étude des conditions météorologiques est un domaine d'intérêt pour de nombreuses personnes, notamment les professionnels de l'agriculture, de la navigation, du transport, et de l'énergie. Dans cette étude, nous nous intéressons aux écarts météorologiques observés au niveau des préfectures françaises. Pour cela, nous allons utiliser le site Météo Blue [1] qui fournit différents paramètres (température, vent, précipitations…) ainsi que les techniques du Web Scraping en utilisant le langage Python pour collecter les données nécessaires.

L'objectif de cette étude est de fournir une analyse comparative entre les données météorologiques collectées et les données en temps réel récupérées depuis différentes régions à partir des différents indicateurs choisis : Température (°C), température ressentie, direction du vent (Nord, Est…), vitesse du vent (km/h), précipitations (mm/3h) et probabilités des précipitations.

2. Design Of Experiment

a. Question de recherche

Quels sont les écarts météorologiques observés au niveau des préfectures à l'échelle de la France ?

b. Facteurs expérimentaux

Les facteurs indépendants sont les préfectures françaises qui seront utilisées pour extraire les données météorologiques du site Météo Blue en utilisant différentes techniques de Web Scraping. Les facteurs dépendants sont les différents paramètres météorologiques tels que la température (°C), la température ressentie, la direction du vent (Nord, Est…), la vitesse du vent (km/h), les précipitations (mm/3h) et les probabilités des précipitations.

c. Plan d'expérience

Les données météorologiques seront collectées sous différentes heures pour une même journée à partir du site de Météo Blue en utilisant différentes techniques de Web Scraping. Nous aurons des prévisions avec un horizon de 7 jours à un horizon d'un jour par rapport à la valeur réelle mesurée. Aucun échantillon humain ne sera utilisé dans cette expérience.

d. Analyse des données

Les données météorologiques extraites seraient comparées à la véritable valeur mesurée et ce par le biais du calcul du pourcentage d'erreur de la prévision par rapport à la mesure réelle.

Nous pourrons aussi comparer la fiabilité des indicateurs et suivre leurs évolutions durant la collecte des données.

e. Résultats attendus

Il est attendu que plus les prévisions se rapprochent de la date voulue, plus leur valeur est proche de celle de la valeur mesurée. Cette étude pourrait révéler des différences significatives dans les paramètres météorologiques extraits.

3. Méthodologie de la Collecte et du Stockage des Données

a. Choix du site

Pour la collecte de données par la méthode du Web Scraping, nous devons obtenir un site qui comporte le maximum des indicateurs météorologiques souhaités.

Pour cela, nous avons observé la plupart des sites météorologiques de France. L'un d'entre eux, Météo France nous semblait très pertinent puisque c'est une base de données stockant la majorité des données météorologiques de France jusqu'à plus de 10 années. Or, il nous donne accès aux valeurs réelles de la météo à l'instant sauvegardé, il n'y a pas de stockage des données de prévisions une fois la date passée.

C'est pourquoi, nous avons cherché des sites avec des prévisions météorologiques pour les jours à venir.

Dans un premier temps, nous avons trouvé le site de La Météo Agricole qui avait des prévisions jusqu'à deux semaines, avec un grand panel d'indicateurs, mais ce service est disponible uniquement aux abonnés. Dans un second temps, nous avons trouvé un Météo Blue qui permet d'obtenir les prévisions jusqu'à une semaine avec tous les indicateurs qui nous convenaient.

C'est donc ce dernier que nous avons sélectionné pour réaliser le Web Scraping.

b. Étude du code source de la page

Puis, nous avons étudié le code HTML de la page de Météo Blue pour connaître le chemin jusqu'aux indicateurs. Sur le site de Météo Blue, les données sont affichées dans une balise tableaux "table", donc avec des balises "tr" et des balises "td" pour les données, ce qui est simple à obtenir en Web Scraping.

Enfin, nous pouvons collecter les données grâce aux packages Request et BeautifullSoup de python. Suite au programme de Web Scraping, les données sont contenues dans des listes.

c. Étude de l'URL

Ensuite, nous avons réalisé une étude de l'url. Le but de cette étude était de percevoir comment utiliser l'url pour passer d'une ville à une autre sur le site, afin d'automatiser la collecte des données.

Exemple d'URL : https://www.meteoblue.com/fr/meteo/semaine/alès _france_3038224

Premièrement, si l'on modifie la ville dans l'url alors la page reste sur la ville précédente, donc ce n'est pas ce paramètre qui nous intéresse pour passer d'une ville à une autre.

Mais en modifiant l'identifiant qui se trouve à la fin de l'url, on arrive à obtenir la page associée à cet identifiant donné.

Donc pour automatiser la collecte de données, il faut obtenir ces différents codes numériques associés aux différentes villes, ici les préfectures des départements de France. De plus, nous avons remarqué que ces codes sont générés aléatoirement pour toutes les villes de France, il n'y a pas de modèle pour les retrouver simplement.

Pour les collecter, nous avons deux solutions. La première est de collecter ces identifiants un par un en recherchant la ville dans la barre de recherche du site. La seconde est de trouver dans l'identifiant dans le script qui permet la liaison entre le nom de la ville souhaité et le code associé. Par la suite, nous avons stocké ces données dans un document csv.

d. Stockage des Données

Après la collecte des données, il faut stocker celles-ci. Pour le stockage, deux options s'offrent à nous. La première option est un stockage des données sur une base de données, la seconde est un stockage de données sur un document csv. Nous avons choisi de réaliser le stockage de données sur un serveur local MongoDB grâce au package pymongo de python.

Pour finir, nous avons converti les données dans un fichier JSON, grâce à la commande suivante : `mongoexport --db=MeteoWebScrapingDB --collection=Indicator --out=data.json`

4. Méthode d'analyse

Une fois les données récupérées, ici plus de 45 000 données, il est nécessaire de passer par une phase de calcul pour en tirer des informations pertinentes. Nous allons réaliser différents calculs pour comparer les données entre elles.

Nous avons stocké des données de prévisions prises à différents horizons : de 7 jours en avance à 1 jour en avance. Nous avons également pris des prévisions à différentes heures pour une même journée.

Le test va être de comparer toutes ces valeurs à la véritable valeur mesurée.

Tous nos indicateurs sont numériques, nous utiliserons donc la même méthode pour tous :

$$\Delta = \left| \frac{V_{prévisions} - V_{mesure}}{V_{mesure}} \right|$$

Ce calcul nous informera sur le pourcentage d'erreur de la prévision par rapport à la mesure réelle.

Une fois ces écarts déterminés, nous pourrons en faire la moyenne pour chaque indicateur et pour horizon de prévision, c'est-à-dire la moyenne des prévisions à 7 jours, 6 jours jusqu'à 1 jour.

Le résultat auquel nous nous attendons est de voir que plus les prévisions sont réalisées proches de la date voulue, plus elles seront proches de la valeur mesurée.

Une fois ces moyennes en notre possession, nous pouvons réaliser différentes comparaisons.

a. Comparer la fiabilité des indicateurs

Nous pouvons comparer les indicateurs entre eux, pour voir s'il se dégage des différences dans leur prévision. Est-ce que tous les indicateurs sont prédits avec le même niveau de fiabilité ? Est-ce qu'il y en a un plus difficile que les autres à prédire ?

b. Analyser l'évolution des prévisions des indicateurs

Nous pouvons nous intéresser à la façon dont évoluent les prévisions pour un même indicateur. L'amélioration des prévisions est-elle linéaire ? Un seuil est-il visible ?

c. Comparer l'évolution des prévisions entre indicateurs

Nous pouvons comparer l'évolution des prévisions des indicateurs, c'est-à-dire analyser si les prévisions évoluent de la même façon pour tous les indicateurs. Est-ce que toutes les prévisions s'améliorent jusqu'à J-1 ? Est-ce que toutes les prévisions s'améliorent dès J-6 ou faut-il encore plus se rapprocher du jour J pour commencer à voir une amélioration des prévisions ?

d. Comparer les prévisions en fonction des départements

Nous récupérons les données pour toutes les préfectures françaises. Nous pourrons donc comparer les écarts en fonction des préfectures pour voir si les prévisions sont plus difficiles à réaliser dans certains départements.

5. Conclusion

Nous avons utilisé le Web Scraping pour récupérer des prévisions météorologiques et les données réelles. Nous avons ensuite réfléchi à une méthode pour utiliser et analyser ces données. Nous sommes conscients que ce jeu de données n'est pas suffisant pour tirer des conclusions pertinentes, qui plus est pour un phénomène aussi complexe que la météo. Nous avons recherché un historique de prévisions météorologiques pour avoir un jeu de données conséquent mais nous n'en avons pas trouvé.

6. Source

[1] Météo Blue, https://www.meteoblue.com/