
Making-of du travail final

Pour mon travail final, j'ai décidé de créer un code qui permettrait de répertorier dans un fichier CSV tous les produits disponibles vendus à la Société des alcools du Québec (SAQ) et plusieurs de leurs caractéristiques. Plus précisément, la liste présente le nom du produit, son prix, sa catégorie, le nom du pays de production, le format, le degré d'alcool, le code SAQ, le code CUP et le site internet lié au produit. Je suis partie de l'URL suivant : <https://www.saq.com/fr/produits> . Cette URL présente tous les produits du site de la SAQ. Afin d'avoir accès à tous ces produits, il fallait créer une nouvelle URL (*urlproduits*), en ajoutant « ?p= », le numéro de page, et « &product_list_limit=96 », qui permettait d'avoir 96 produits par page. J'ai également utilisé le lien d'un produit en particulier afin d'inspecter l'élément: <https://www.saq.com/fr/12708428> . Pour m'aider à construire mon code, j'ai utilisé mes notes de cours, mes devoirs précédents, le guide simplifié de *BeautifulSoup* (<http://jhroy.ca/uqam/edm5240/BeautifulSoup-DocAbregée.pdf>) et la documentation complète (<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>). J'ai également visité plusieurs forums qui m'ont aidé lorsque j'étais face à une impasse. Voici le lien de ces forums :

- <https://stackoverflow.com/questions/18994334/make-python-code-continue-after-exception>
- <https://www.w3schools.com/python/default.asp>
- <https://docs.python.org/fr/3.5/tutorial/errors.html>

Je trouvais ce sujet pertinent puisqu'il s'intéresse à une société d'État très importante au Québec, qui rejoint une grande partie de la population. En effet, lors de son rapport annuel de 2019, la SAQ a révélé avoir réalisé 3,294 G\$ de ventes des réseaux de vente et de distribution¹. Puisqu'il s'agit d'un monopole, peu de gens vont se questionner par rapport aux prix des produits offerts en les comparant avec d'autres, d'où l'utilité de mon fichier CSV. En répertoriant tous les produits disponibles à la SAQ, nous pourrions

¹ <https://www.saq.com/fr/rapport-annuel-2019>

comparer les prix de vente avec ceux des autres provinces canadiennes, voir même des autres pays (en ramenant tous les prix selon la devise canadienne). Plusieurs autres usages pourraient être faits avec ce fichier. Nous pourrions établir les pays qui produisent le plus de boissons alcoolisées pour la SAQ, et même évaluer le nombre de produits canadiens vendus. Avec ces résultats, nous pourrions constater où en est le Québec dans la promotion des produits canadiens, en comparaison avec les autres provinces du pays. Nous pourrions également trouver quelle est la catégorie de produit la plus présente sur le site de la SAQ, et encore une fois, comparer ces résultats avec les autres provinces canadiennes. D'une plus générale, ce fichier pourrait donc être utilisé pour étudier les habitudes de consommation d'alcool des Québécois.

Pour faire mon fichier CSV, j'ai utilisé le logiciel Visual Studio Code et l'extension Python. Puisque je devais effectuer un moissonnage de données, il s'agissait des outils les plus appropriés et ceux avec lesquels j'étais le plus à l'aise. J'ai également utilisé la bibliothèque python *BeautifulSoup*, qui m'a permis d'extraire des données spécifiques (nom du produit, prix, catégorie, nom du pays de production, format, degré d'alcool, code SAQ, code CUP et site internet) sur le site de la SAQ et d'en faire un fichier CSV (que j'avais préalablement importé). J'ai finalement utilisé le module *requests*, qui m'a permis de faire des requêtes sur le site internet de la SAQ.

J'ai eu plusieurs petits pépins qui ont facilement pu être réglés en relisant mon script ou mes notes de cours, mais j'ai eu deux plus gros problèmes qui ont été plus complexes à régler. Tout d'abord, je devais isoler les données concernant le nom du pays de production, le format, le degré d'alcool, le code SAQ et le code CUP (les autres données étaient dans des balises *meta* et plus faciles à isoler). Ces données étaient dans des balises *li*, qui elles étaient dans des balises *ul*. Au début, j'ai isolé tout le texte qui se trouvait dans les balises *ul*, mais il y avait trop d'information et ce n'était pas assez précis. Avec l'aide de mon enseignant et des recherches sur les forums, j'ai réussi à préciser mon extraction de données en ne sélectionnant que le texte se trouvant dans les balises *strong*. J'ai ainsi fait un *find* sur les balises *ul* et un *find_all* sur toutes les balises *li* se trouvant à l'intérieur. Mon deuxième problème a fait surface à la toute fin de mon travail, lorsque j'ai fait rouler mon

script. Tout se passait très bien jusqu'à ce que j'arrive au produit 1148, *La Massa Carla 6 2015*. Un message d'erreur de type *NameError* apparaissait en lien avec le prix du produit. Je suis donc allée voir la page de ce vin afin de comprendre pourquoi il y avait ce message d'erreur. Je me suis rendu compte que le produit était épuisé, et qu'il n'y avait pas de prix associé. C'est pourquoi le script ne trouvait pas de données reliées au prix pour ce produit. J'ai donc utilisé le code `try...except(NameError,TypeError)`. Ainsi, si mon script fait face à l'une de ces erreurs, au lieu de s'arrêter, il continue et passe au produit suivant. Au début, j'avais utilisé le code `pass` après le `except`, mais pour une raison qui m'échappe, le produit épuisé apparaissait en double et le prix affiché était celui du produit qui le précédait. J'ai donc réglé ce problème en utilisant le code `continue`. Avec ce code, le produit est donc omis de la liste. Puisqu'il s'agit de produits qui ne sont plus offerts à la SAQ, ce n'est pas problématique qu'ils ne soient pas sur la liste. Après avoir fait cet ajustement, mon script roulait sans problème.

Puisqu'il est plutôt long de rassembler plus de 14142 données (nombre de produits sur le site aujourd'hui), j'ai décidé de m'arrêter à 2174 produits. Par contre, le script était programmé pour se rendre jusqu'à la page 148, donc jusqu'à un peu plus de 14200 produits (avec 96 produits par page). Il aurait donc été en mesure de ramasser les données de tous les produits disponibles sur le site de la SAQ.