ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

Rapport d'étape
présenté à
Mme. Sylvie Ratté
LE 26 juin 2015

PAR:

Simon Beaulieu (BEAS27019203)
Dominique Jacques-Brissette (JACD16069105)

Dans le cadre du cours MTI830 - Forage de données texte et audiovisuelles

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ÉTÉ 2015

Titre et auteurs

Analyse économique du Canada et de ses provinces selon le parti au pouvoir et le cour du dollar Canadien depuis janvier 1991.

Auteurs principaux : Simon Beaulieu et Dominique Jacques-Brissette.
Collaborateurs : Eric Velasquez Godinez et Sylvie Ratté

Introduction

En Avril 2012, le gouvernement du Canada a décidé, à l'assemblée générale annuelle du Partenariat pour un Gouvernement Transparent[1] de présenter son plan d'action. Ce plan d'action, nécessaire pour la participation au sommet précédemment énoncé, contient des mesures concrètes à haut niveau qui ont pour but d'impliquer davantage les citoyens dans le processus gouvernemental, notamment au niveau de la lutte la corruption, en misant sur la transparence. La création d'un portail de données ouvertes accessible publiquement en ligne est un résultat de ce plan d'action et les données que nous avons utilisées pour notre projet y proviennent.

Cet article présente notre cheminement qui a abouti à la création d'un outil de visualisation permettant de présenter de façon claire et simple les données que nous avons extraites d'un ensemble de données. Nous présenterons d'abord le contexte de notre projet afin d'expliquer la solution que nous proposons, en poursuivant une liste d'objectifs catégorisés que nous voulons atteindre. Ensuite, nous allons détailler les méthodes qui ont été utilisées pour répondre à ces objectifs et nous exposerons les résultats encourus tout en portant une analyse critique de la situation. Ceci permettra d'illustrer les difficultés que nous avons rencontrées et d'effectuer une ouverture pour la suite de notre projet.

Finalement, nous concluerons brièvement en expliquant les techniques de forage de données utilisées et en effectuant un survol du travail que nous avons accompli jusqu'à maintenant.

[1] http://www.opengovpartnership.org/about

Contexte

Notre ensemble de données provient du portail de données ouverte du gouvernement du Canada et est intitulé “Statistiques de finances publiques, situation des opérations des administrations publiques et bilan”[2]. Le contenu de cet ensemble de données représente les revenus et les dépenses du domaine de l’administration publique au Canada, comptabilisé à la fin de chaque quartile depuis l’année 1991, le tout organisé par secteur gouvernemental. Nous avons donc gratuitement accès à une très grande quantitée de données qui proviennent d’une source relativement très fiable.

Là où la situation se complique est que le format dans lequel les données sont présentées n’est pas des plus conviviaux. Les “datasets”, comme le nôtre, sont souvent sous la forme de grands tableaux comportant peu d’informations sur le contexte. Certes, il y a une grande quantité de données mais il est difficile, voire impossible de les interpréter facilement dès le premier coup d’oeil tant qu’elles restent présentées sous ce format. La quantité d’information qu’on y trouve est phénoménale mais les moyens accessibles et utilisables selon les technologies disponibles de nos jours n’existent tout simplement pas.

Nous possédons donc une très grande quantité d’informations représentant les dépenses et les revenus par quartiles depuis janvier 1991 pour le gouvernement fédéral du Canada. Ces données ont été compilées par statistiques Canada.

Ce projet permet de porter une analyse plus objective sur la situation économique du Canada ainsi que ses provinces dépendant du parti politique fédéral au pouvoir ainsi que du cour du dollar Canadien.

[2] http://ouvert.canada.ca/data/fr/dataset/9d38820b-41b1-442e-822f-f4c6d94f9ad6

Objectifs

Nous voulons donc créer un outil de visualisation web à partir de ce dataset afin de représenter graphiquement les tendances entre les montants alloués et perçus pour chaque quartile ainsi que le gouvernement au pouvoir. Nous voulons également prendre en compte les données financières externes telles que le cour du dollar Canadien ainsi que les indices de croissance économique donnés par des organismes tel que le Fond Monétaire International. Ceci permettra de mieux mettre en perspective les changements financiers perçus au fil du temps. Notre outil permettra également de comparer toutes les variables différentes présentées dans le dataset pour lequel certaines comparaisons présentent des tendances intéressantes.

Nous cherchons à établir un lien entre les différents domaines de dépenses, le montant de ces dépenses et le gouvernement au pouvoir. En utilisant un dataset disponible publiquement sur le site du gouvernement, nous voulons classifier les données, notamment par secteur économique afin de les relier aux valeurs du parti politique au pouvoir à ce moment.

Classifier les données fournies pour pouvoir les regrouper par secteurs économiques.
Décrire les différents secteurs économique dans le contexte de notre analyse.
Établir des mesures statistiques fiables pour établir une tendance générale entre la situation économique du Canada et le parti politique au pouvoir.

8

Méthodes

Données :

Le Dataset que nous avons en notre possession représente les revenus et les dépenses du gouvernement fédéral pour chaque quartile, le tout divisé par secteur. Pour chaque année, on peut y retrouver 53 directives uniques réparties en 5 catégories gouvernementales.

Nous allons également manuellement ajouter à notre corpus d'autres informations externes qui sont cohérentes pour notre analyse. Premièrement, nous ajouterons la liste des partis politiques au pouvoir et le statut de leur gouvernement (minoritaire, majoritaire) pour la période couvrant de janvier 1991 à aujourd'hui. Deuxièmement, nous voulons prendre en compte certains indices financiers relatifs au Canada tel que le cour du dollar canadien (CAD) ainsi que les taux de croissances correspondantes. En effet, plusieurs organismes tel que le Fond Monétaire International donnent ces données publiquement, alors nous les ajouterons à notre projet.

Outils :

La base de données utilisée se nomme MongoDB et est un projet de logiciel libre qui gagne beaucoup en popularité dans les solutions de bases de données n'utilisant pas de SQL (c'est-à-dire qu'il ne s'agit pas d'une base de données relationnelle). L'utilisation d'une telle base de donnée nous permet de rendre notre corpus accessible à notre outil de visualisation et de manipuler les données facilement à l'aide de Javascript sous la forme d'objets JSON.

Nous allons utiliser le cadriciel `AngularJS` pour bâtir notre application, en collaboration avec du HTML5 / CSS3 et d'autres librairies utilitaires en Javascript nous permettant de modéliser des graphiques complexes et visuellement attirants pour les utilisateurs.

Pour l'extraction initiale de notre corpus depuis le format original fourni par le gouvernement (fichier CSV) vers le format choisi pour l'intégrer dans notre base de données, un script Python sera conçu. Toutes les librairies nécessaires à l'extraction et l'écriture des nouvelles données sont disponibles en ligne et cette méthode est l'une des plus facile à implémenter.

Python est un langage facile à utiliser, très performant et comportant beaucoup de librairies fiables et variées nous permettant de faire à peu près n'importe quoi. La syntaxe du langage est basée sur l'indentation, ce qui nous oblige à structurer notre code toujours de la même façon.

Prétraitement

Le dataset original est présenté dans un fichier CSV qui contient une liste de mots clés séparés par des virgules. Nous allons extraire les données de ce fichier et les stocker dans une base de données non relationnelle utilisant des collections d'objets JSON. La notation JSON signifie Javascript Object Notation, et peut être représentée comme suit :

```
{
    "menu": {
        "id": "file",
        "value": "File",
        "popup": {
            "menuitem": [
                {
                    "value": "New",
                    "onclick": "CreateNewDoc()"
                },
                {
                    "value": "Open",
                    "onclick": "OpenDoc()"
                },
                {
                    "value": "Close",
                    "onclick": "CloseDoc()"
                }
            ]
        }
    }
}
```

L'équivalent en CSV, en utilisant la virgule comme séparateur, serait représentée comme suit, les clés dans la première section avec les valeurs après le saut de ligne :

```
"id";"value";"popup.menuitem.0.value";"popup.menuitem.0.onclick";"popup.menuitem.1.value";"popup.menuitem.1.onclick";"popup.menuitem.2.value";"popup.menuitem.2.onclick"
"file";"File";"New";"CreateNewDoc()";"Open";"OpenDoc()";"Close";"CloseDoc()"
```

L'interface Web permet de présenter les résultats de façon visuelle et interactive ce qui rends le résultat de notre analyse accessible à un grand nombre d'utilisateurs, qu'ils soient connaisseurs dans le domaine ou non. De plus, il s'agit d'un format compatible pour pratiquement tous les dispositifs électroniques connectés à l'Internet, ce qui croît encore plus l'accessibilité du grand public à notre application d'analyse économique.

Résultats

Au départ, nous voulions faire ce projet en utilisant les langages Python et R, mais après avoir regardé ces technologies nous en sommes venus à la conclusion qu'il n'y a pas de raison particulière où R nous avantagerait par rapport à d'autres technologies dont nous sommes déjà à l'aise avec (tel que les technologies Web que nous utiliserons). En effet, nous ne connaissons presque pas ce langage, et nos connaissances en Javascript nous permettront de faire ce dont nous avons besoin de faire et cela beaucoup plus rapidement que nous aurions pu en utilisant R.

Le dataset contient deux colonnes ("vector" et "coordinates"), dont nous ignorerons dans notre analyse, car elles n'existent que pour des raisons techniques lors de la construction du dataset par le gouvernement.

Un script Python a été conçu pour lire le fichier CSV du gouvernement et le convertir en format JSON. Il reste par contre à l'utiliser pour populer la base de données MongoDB.

Nous savons exactement quelles données supplémentaires nous voulons (cour du dollar CAD, indices de croissances du Canada, ratio CAD/USD tout cela au fil du temps depuis 1991), mais nous ne savons pas encore de quelle source nous allons les prendre (il y a plusieurs choix, mais nous voulons les étudier de plus près pour choisir le meilleur pour notre cas).

Nous avons également expérimenté avec quelques librairies de dessin de graphiques en JavaScript, mais notre choix final reste encore à faire. Il s'agit d'un choix important, car on veut être capable de produire des graphiques interactifs et visuellement plaisants.

Conclusion

Tant que nous n'avons pas produit les graphiques et corrélé les données du gouvernement avec les autres données externes, nous n'avons aucune idée des résultats produits. Il est tout à fait possible qu'aucune corrélation solide ne puisse se faire, mais nous espérons que oui. Cela dépend fortement de la fiabilité des données du gouvernement, mais la prémisse de base de ce projet est que ces données sont solides et véritables, nous avons toutes les raisons de croire que c'est le cas.

Ce projet s'est révélé être plus complexe à mettre en pratique que nous l'avions cru au départ, mais nous croyons que ça en vaudra la peine puisque ça permettra de donner un sens aux données brutes qui sont autrement sans signification considérable.

Remerciements

L'équipe tient à remercier Sylvie Ratté pour ses commentaires judicieux et ses conseils nous guidant dans la bonne direction pour la suite du projet. Nous tenons également à remercier Erick Velázquez Godínez pour son aide concernant Python ainsi que ses commentaires en général.

Références

Pour en savoir plus sur le Partenariat pour des gouvernements ouverts :
http://www.opengovpartnership.org/about

Lien vers la source de données principale utilisée
http://ouvert.canada.ca/data/fr/dataset/9d38820b-41b1-442e-822f-f4c6d94f9ad6