Projet individuel

PJI (54) - Mais que font nos députés ? Une sociologie informatique du travail parlementaire

Auteur : Quentin Baert
Encadrant universitaire : Samuel Hym
Encadrant : Etienne Ollion

janvier à juin 2015

Sommaire

Remerciements

Merci à Samuel Hym et Etienne Ollion pour leur accessibilité, leur bienveillence, leur suivi et pour avoir répondu à chacune de mes questions.

Introduction

Le PJI (Projet individuel) se déroule dans le cadre de la première année de master informatique à l'Université de Lille 1. Le but est ici de développer un projet sur l'ensemble d'un semestre.

Ce rapport concerne le projet numéro 54, intitulé "Mais que font nos députés ? Une sociologie informatique du travail parlementaire" et encadré par messieurs Samuel Hym, enseignant chercheur au laboratoire CRIStAL de Lille, et Etienne Ollion, chercheur CNRS au laboratoire SAGE de Strasbourg.

Le but de ce projet est d'exploiter les comptes rendus intégraux de l'Assemblée nationale française afin de constituer des bases de données qui contiennent les informations des scrutins qui y sont votés : numéro, date, sujet du scrutin ainsi que nom, prénom, parti et vote des députés participants au scrutin.

Ce rapport présentera l'ensemble des travaux effectués sur le projet ainsi que leurs résultats. Pour commencer, le choix des outils utilisés sera exposé et justifié. Nous abroderons ensuite la manière dont les comptes rendus de l'Assemblée nationale ont été récupérés. Puis, nous verrons comment ces comptes rendus ont été filtrés et mis en forme afin de pouvoir en exploiter les données. Enfin, nous verrons comment les données ont été extraites et nettoyées afin de créer les bases.

1 Objectifs du projet

L'objectif de ce projet était de collecter l'ensemble des scrutins publics tenus à l'Assemblée nationale entre 1958 et 2002 et de les organiser dans une base de données. Les données initiales devaient être extraites de documents PDF, nettoyées et présentées sous forme d'une base de données facilement interrogeable (sous forme d'un fichier CSV par exemple).

Les différentes étapes étaient donc :

1. récupérer tous les PDF des comptes rendus intégraux sur le site de l'Assemblée nationale
2. convertir ces PDF en fichiers textes afin de pouvoir facilement travailler sur le texte des comptes rendus
3. tous les fichiers ne contenaient pas de scrutins, il fallait donc filtrer les fichiers pour isoler ceux qui nous intéréssaient
4. créer des fichiers CSV à partir des fichiers qui contenaient en effet un ou plusieurs scrutins

Toutes ces étapes ont été traitées. Leur mise en place et leurs résultats seront présentés dans la suite de ce rapport.

2 Choix des outils

2.1 Le langage Scala

2.2 SBT (Scala Build Tool)

3 Récupération des comptes rendus intégraux de l'Assemblée nationale

3.1 Organisation des PDF sur le site de l'Assemblée nationale

3.2 Récupération des PDF

4 Filtrage des données

4.1 Convertion des PDF en fichiers textes

4.2 Analyse des fichiers textes pour isoler ceux contenant des scrutins

5 Mise en forme des informations récupérées

5.1 Représentation objet d'un scrutin

5.2 Extraction des données des scrutins

5.3 Nettoyage des données

6 Construction des bases de données

Conclusion

Bibliographie