



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	2
Introduction	3
1 Objectifs du projet	4
2 Choix des outils	5
2.1 Le langage Scala	5
2.2 SBT (Scala Build Tool)	5
3 Récupération des comptes rendus intégraux	6
3.1 Organisation des PDF sur le site de l'Assemblée nationale . .	6
3.2 Récupération des PDF	6
4 Filtrage des données	7
4.1 Conversion des PDF en fichiers textes	7
4.2 Analyse des fichiers textes pour isoler les scrutins	7
5 Mise en forme des informations récupérées	8
5.1 Représentation objet d'un scrutin	8
5.2 Extraction des données des scrutins	8
5.3 Nettoyage des données	8
6 Construction des bases de données	9
Conclusion	10
Bibliographie	11

Remerciements

Merci à Samuel Hym et Etienne Ollion pour leur accessibilité, leur bienveillance, 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 CRISAL 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 aborderons 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

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 Conversion 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