# PJI - DownloadHTML

## Auteur


- Ouamar SAIS


## Description

Le projet consiste à extraire du html des comptes rendus intégraux de l'Assemblée Nationale française des données concernant les prises de paroles, des intervenant, et aussi des humeurs (réactions aux interventions) qui y sont effectués. Les informations récupérées seront ensuite nettoyées, et organisées dans une ou plusieurs bases de données.

## Récupération des HTMLs

Les sources des classes permettant de télécharger les pages Html des comptes rendus intégraux de l'Assemblée Nationale se trouvent dans le package `download`.

### Procédé utilisé

Les pages d'indexes de toutes les legislatures depuis la 12eme (http://www.assemblee-nationale.fr/X/debats/index.asp, avec X compris entre 12 et 14) sont filtrées afin d'en tirer les URLs des pages de sessions. Ces pages de sessions sont filtrées à leur tour afin d'obtenir les URLs des pages html.

Une fois la liste de toutes les URLs établie, le contenu html correspondant à chacune d'entre elles sont téléchargés et organisés localement selon leur legislature, années et session.

## Netoyage du html

Une fois les pages html récupérée il faut d'abord les nettoyer, pour cela j'ai utilisé la librairie java htmlCleaner qui permet de netoyer le code et de l'indenter correctement pour qu'il soit plus facile de se repérer.

Une fois celà effectué, il est possible de commencer à recuperer les interventions, intervenant et faire des calculs sur ces informations.

## Analyse des fichiers html

Pour l'analyse des fichier html néttoyés, j'ai utilisé la librairie Jsoup qui permet de naviguer dans les différentes div et attributs d'un fichier html.

Les informations récupérés sont: - Liste des interventions associées à leur intervenant, classé par session. - Nombre de mots prononcés par intervention et par session

## Structure

Les sources sont organisées selon le schéma d'un projet SBT (Scala Build Tool), c'est à dire :

```
|_scala
|_ build.sbt
|_ src
    |_ main
        |_ scala
            |_ download
                |_ PDFDownloader.scala
                |_ URLManager.scala
|_PJI_project
  |_src
     |_analyse
        |_AnalyseTalks.java
        |_Main.java
     |_cleaning
        |_Cleaner.java
```

# Utilisation

L'execution des différentes fonctionnalités necessite Scala Build Tool (http://www.scala-sbt.org/).

## Récupération des fichiers html

- Pour récupérer toute les pages html en une fois

```
$ sbt
> compile
> console
scala> HTMLDownloader.downloadAll
```

- Pour récupérer le n-ième paquet de 100 PDFs

```
$ sbt
> compile
> console
scala> HTMLDownloader downloadGroupNb X
```

## Lancement de l’analyse des fichier

Il faut lancer le main du package analyse dans le dossier PJI_Project