L'objectif de ce projet est d'extraire les tableaux Wikipedia et de les traduire dans un format plus simple et adapté à l'analyse statistique. De manière générale, l'extraction de données est une activité difficile et pourtant cruciale : il faut qu'elle soit fiable et cohérente. Le choix d'une extraction en CSV a été effectué car ce format est très simple et est supporté par de nombreux outils : R, excel, SAS, etc... La principale difficulté du projet sera de développer une procédure robuste et la plus générale possible. L'objectif est donc d'extraire le plus possible de tableaux Wikipedia (et évidemment l'extraction devra produire des fichiers CSV bien formés et corrects). Cette fiche résume mon projet.

Table des matières

1 Comment j'ai procédé :

Afin de répondre aux exigences je suis parti de l'url suivant : https://en.wikipedia.org/wiki/Comparison_of_Canon_EOS_digital_cameras, le but étant de réaliser un algorithme robuste d'extraction des tableaux de cette page afin de pouvoir ensuite être appliqué à n'importe quel tableau et n'importe quelle page wikipedia. J'ai donc procédé de la manière suivante :

- Création des différentes classes : Table, Body, Header, Ligne, Cellule, etc... En effet, une page est composée de tableau, lui même composé d'un body et d'un header, eux mêmes composés de lignes qui sont composées de cellules. J'ai donc directement codé en plusieurs classe pour ne pas avoir à fragmenter mon code par la suite. Le diagramme de classe est présent sur la dernière page.
- Utilisation de la librairie Jsoup pour l'extraction des objets à partir du code HTML.
- Après avoir été capable d'extraire les tableaux de l'url ci dessus, il a fallu extraire seulement les tableaux qui étaient potentiellement exploitables d'un point de vue statistique. Le programme extrait donc uniquement les tableaux de type : *wikitable sortable*.
- J'ai ensuite créé différents validateurs : page valide, tableau valide, etc, en essayant de respecter les différents design patterns vus en cours.
- Enfin, dans une dernière phase, j'ai testé mes différentes fonctions.
- J'ai ensuite généralisé aux autres pages données en exemple dans le fichier wikitext.txt

2 Quelques statistiques

2.1 Nombre de tableaux extraits

L'algorithme que j'ai créé permet d'extraire les tableaux déclarés comme *wikitable sortable* des pages wikipedia données dans le fichier : *wikitext.txt*. Avec le programme proposé on extrait 693 tableaux.
Pour ce faire, il met en moyenne 1 min 15, ce chiffre est à relativiser puisqu'il dépend à la fois de la connection internet et de l'ordinateur sur lequel le programme est lancé.

2.2 Statistiques sur les tableaux extraits

	min	moyenne	max	variance
Ligne	2	21	594	1844
Colonne	2	8	30	20

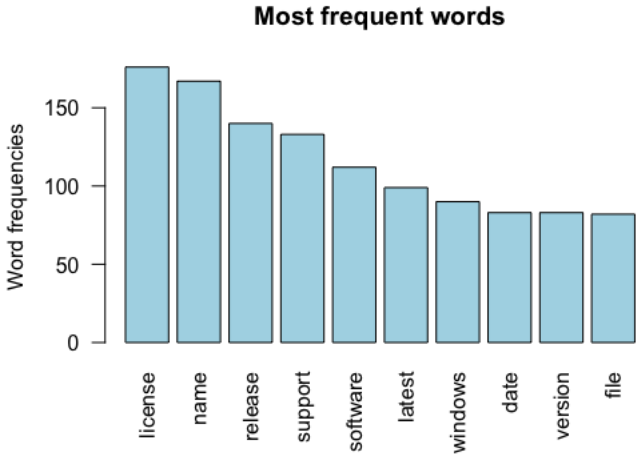
Voici les statistiques concernant les tableaux que j'ai pu extraire, le tableau ayant le plus de ligne est le premier tableau que l'on trouve sur cette page : https://en.wikipedia.org/wiki/Comparison_of_Standard_Malay_and_Indonesian ; celui avec le plus de colonnes est le premier sur cette page : https://en.wikipedia.org/wiki/Comparison_of_accounting_software

2.3 Analyse des noms de colonnes

Afin d'obtenir des statistiques exploitables sur les noms de colonnes j'ai procédé de la manière suivante :

- J'ai récupéré à partir de mon programme l'ensemble des noms de colonnes dans une liste.
- J'ai ensuite exporté cette liste au format CSV afin de l'ouvrir dans le logiciel R.
- J'ai ensuite pratiqué de l'analyse textuelle.

Le nom de colonne le plus utilisé est donc license avec plus de 170 occurrences suivi de name avec aussi presque 160 occurences. Afin de rendre cela plus visuel, j'ai réalisé le nuage de mots ci dessus, les mots les plus utilisés apparraissent en noir et en gros, les moins utilisés en petit et en vert.



3 Synthèse

Ce programme permet donc d'extraire tous les tableaux indiqués comme *wikitable sortable*. Cependant, tous les tableaux ne sont pas exploitables d'un point de vue statistique : on citera notamment les tableaux avec du texte ou les tableaux de conversion. Le critère "sortable" ne veut donc pas forcément dire exploitable.

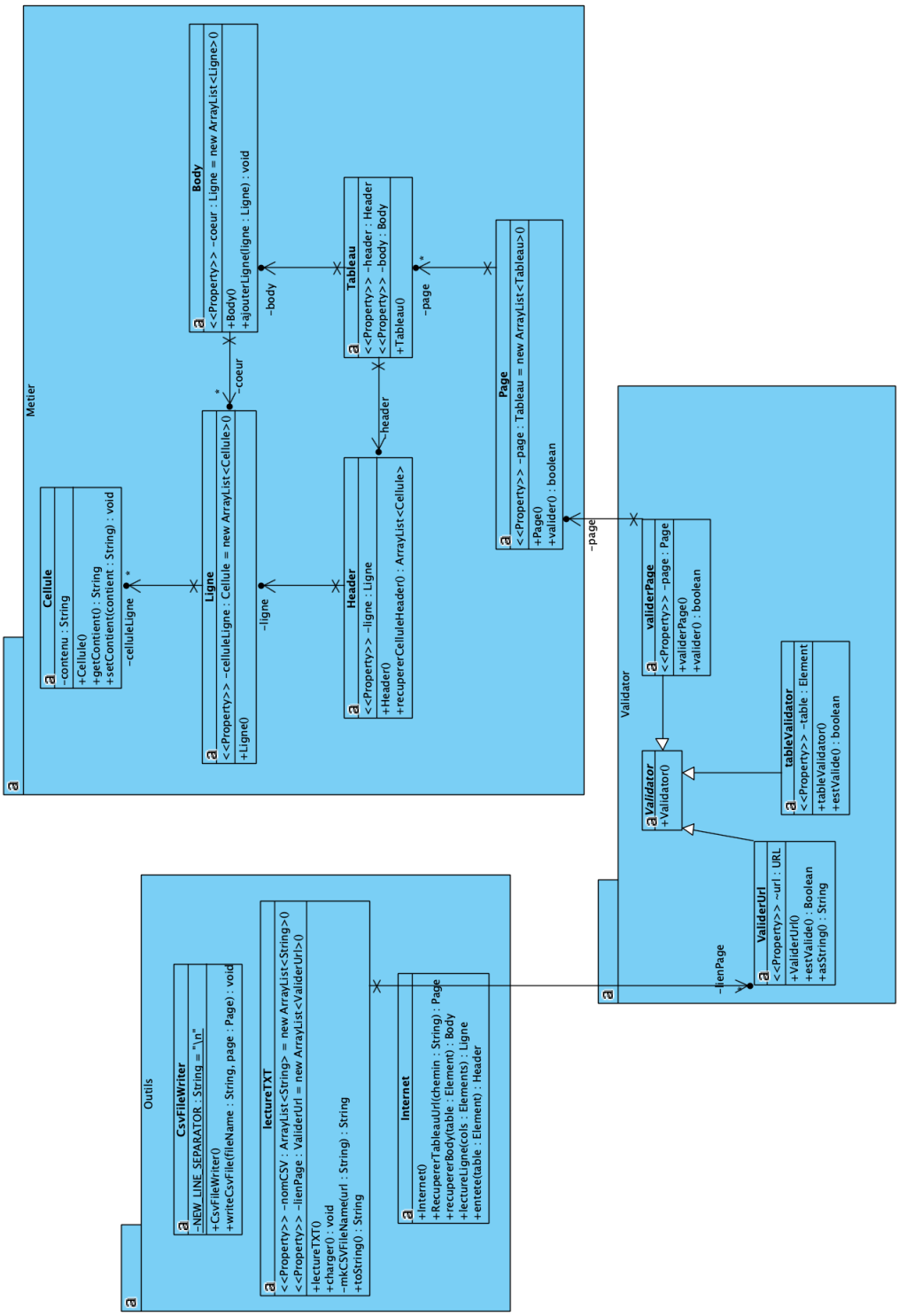
Un autre problème à signaler est celui lié aux problèmes d'encodage notamment avec le tableau chinese romanization.
L'export csv utilisé semble être de bonne qualité :

- Un seul tableau pose problème sur le nombre de ligne exporté : $Comparison\_of\_neurofeedback\_software-1$
- L'algorithme proposé ne prend pas en compte les tableaux avec des header composés de plusieurs lignes (cf exemple)

Il permet néanmoins d'extraire les tableaux qui sont triables à la fois en ligne et en colonne

L'algorithme proposé semble avoir une bonne robustesse. La majorité des tableaux sont bien extraits.
Il faudra néanmoins veiller au changement de la structure de l'information : même si des critères adéquats pour l'extraction des informations pertinentes sont établis, dans le cadre d'une analyse qui s'étale dans le temps, il y a la possibilité que wikipedia change de structure et par conséquent les critères utilisés auparavant ne seront plus valides.

Enfin j'ai eu beaucoup de difficulté à mettre en place les différents design patterns vus en cours.



Graphique 3 – Diagramme de classes