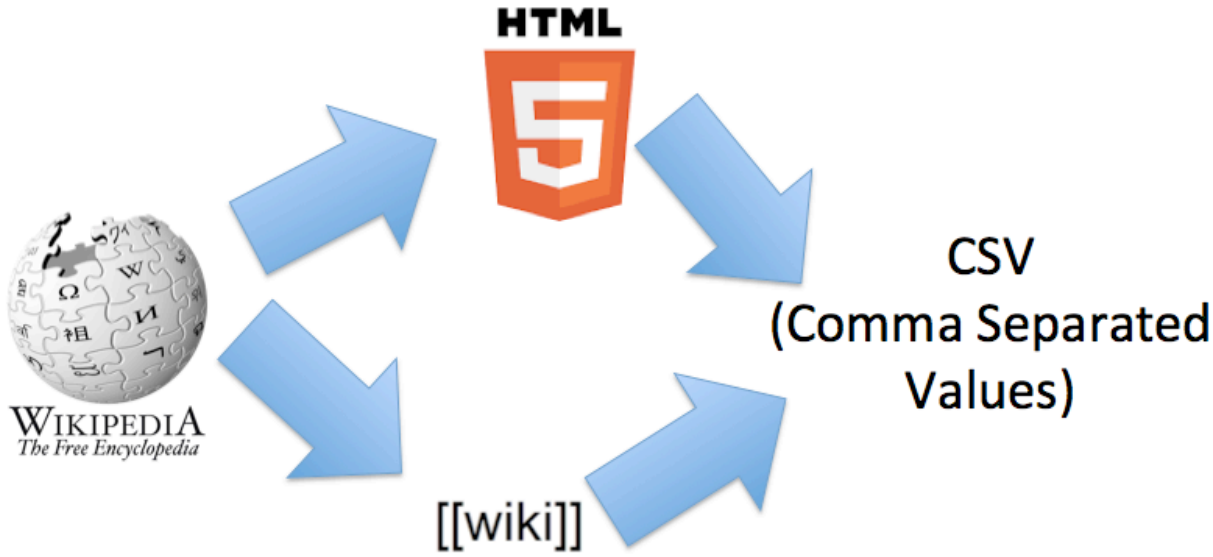
MIAGE 1
Projet Développement Logiciel –
Responsable : Mathieu Acher
Contact : mathieu.acher@irisa.fr
2018/2019

Gestion de projet – Projet Développement Logiciel (PDL)

Projet: « Wikipedia Matrix»

Ressources :https://github.com/acherm/PDL1819

Wikipedia est une fantastique source de données, principalement composée d'articles écrits en langage naturel (e.g., français, anglais).



L'objectif de ce projet de PDL est d'extraire des tableaux au format CSV à partir de pages Wikipedia.

Pourquoi extraire des tableaux dans Wikipedia?

La motivation générale est que les tableaux Wikipedia sont difficiles à exploiter par des outils statistiques, de visualisation ou n'importe quel outil capable d'exploiter les tableaux (e.g., Excel, OpenOffice, RStudio, Jupyter). Ces tableaux sont en effet écrits dans une syntaxe (Wikitext) difficile à analyser et non nécessairement conçue pour la spécification de tableaux. De plus, il y a une forte hétérogénéité dans la manière d'écrire des tableaux, ce qui complique encore plus le traitement des données tabulaires de Wikipedia. Le même constat peut être fait pour le format HTML qui peut être utilisé pour présenter un tableau dans un navigateur Web: il n'est pas facilement exploitable par des outils statistiques ou des tableurs. L'objectif est donc d'extraire les tableaux Wikipedia et de les traduire dans un format plus simple et adapté.

Le choix de CSV https://fr.wikipedia.org/wiki/Comma-separated_values a ainsi été effectué car il a le mérite d'être très simple et il est supporté par de nombreux outils.

La principale difficulté du projet sera de développer une procédure ***robuste et la plus générale possible.*** Pour évaluer les qualités de votre solution, un concours sera organisé : l'objectif sera d'extraire le plus possible de tableaux Wikipedia (et évidemment l'extraction devra produire des fichiers CSV bien formés et corrects).

Quelques défis

Wikipedia contient énormément de données (du texte, des figures, des sections, etc.) qui ne sont pas des données tabulaires. Il est même possible qu'une page Wikipedia ne contienne tout simplement pas de données tabulaires. Un défi sera donc d'occulter ce genre d'informations et ne considérer que les données pertinentes pour l'extraction de tableau. Il faudra aussi être en mesure d'extraire plusieurs tableaux sur une même page Wikipedia.

Une autre difficulté est que certaines données tabulaires sont difficiles à convertir au format CSV (par exemple, les tableaux imbriqués). Une barrière supplémentaire concerne l'hétérogénéité des données tabulaires qui complique la tâche d'extraction.

La conception et l'implémentation d'un algorithme général et robuste nécessitent de nombreux essais et erreurs sur différents cas réels. Il faudra veiller, lors de l'évolution de votre solution, à ne pas casser ce qui marchait précédemment (test de non régression). Typiquement, vous devez vérifier que certaines extractions qui fonctionnaient sur certaines pages Wikipedia sont toujours à même d'y parvenir.

Quelques contraintes supplémentaires

Une page Wikipedia peut être analysée de deux manières différentes :

- En allant chercher le code Wikitext correspondant
- En exploitant le rendu HTML de la page Wikipedia

Aussi, il est demandé d'effectuer deux types d'extraction : via HTML et via Wikitext. Les deux approches devront être comparées et testés (e.g., donnent-elles le même CSV en sortie ?) et la meilleure extraction devra être justifiée.

Auto-évaluation de votre approche

En plus d'une solution technologique (en Java) pour extraire des données tabulaires à partir de Wikipedia, l'objectif de ce projet est aussi d'explorer la pertinence de l'idée : obtient-on des matrices de qualité ? est-ce utile de fabriquer des matrices de comparaison à partir de Wikipedia ?

Vous devez adresser ces deux questions en utilisant votre solution sur plusieurs exemples concrets. Vous devez ainsi démontrer les aspects positifs et la plus value de votre solution mais aussi les limites actuelles, qu'elles soient liées à la qualité des données de Wikipedia, à la difficulté d'extraire des données tabulaires, ou à la qualité de votre solution. L'évaluation de votre approche comptera pour 25% de la note : ce n'est donc pas qu'un travail d'implémentation.

Concours

Un concours sera organisé en cours de projet: l'objectif sera d'extraire le plus possible de tableaux Wikipedia. En particulier, nous fournirons un jeu données (un ensemble d'URLs Wikipedia). Evidemment l'extraction devra produire des fichiers CSV bien formés et corrects. Le groupe gagnant aura réussi à extraire le plus de CSVs.

Agenda

L'implémentation complète et de bout en bout n'est pas triviale, mais tout à fait faisable sous la condition de commencer par des tâches simples et de monter en complexité au fur et à mesure. Aussi il est fortement conseillé :

- De ne considérer que l'extraction avec Wikitext ou HTML (dans un premier temps)
- D'obtenir rapidement une solution « basique » qui étant donné une et une seule URL d'une page Wikipedia produit un CSV (eg

https://en.wikipedia.org/wiki/Comparison_of_Canon_EOS_digital_cameras)

- De tester automatiquement (avec JUnit) votre solution « basique »
- D'étendre votre solution pour supporter non plus une seule URL mais plusieurs. Vous pouvez alors songer à implémenter des techniques de test « génériques ». Il faut aussi songer à expérimenter dans le large, sur plusieurs cas possibles, votre solution.
- De supporter Wikitext et HTML
- De tester automatiquement et de valider de manière croisée les deux types d'extraction
- De s'attaquer au problème de l'extraction de multiples tableaux
- Tester, expérimenter, tester, tester

Une attention particulière devra également être portée à la conception de l'API Java pour que votre solution soit la plus réutilisable et paramétrisable possible.

Comment commencer ?

Outre la mise en place d'un repository Github, étudier Wikipedia et notamment les APIs, parsers, etc. qui sont proposés. Prototyper une première transformation « simple » (cf ci-dessus). Itérer.

Tester. Evaluer la pertinence de votre solution. Itérer, tester.

En parallèle de vos expérimentations, écrire le document de spécification. En particulier, il est nécessaire de sélectionner des librairies Java pour récupérer des données Wikipedia, « parser » le Wikitext, le HTML, produire le CSV : vous devez justifier vos choix dans le document de conception et cela nécessite d'expérimenter.