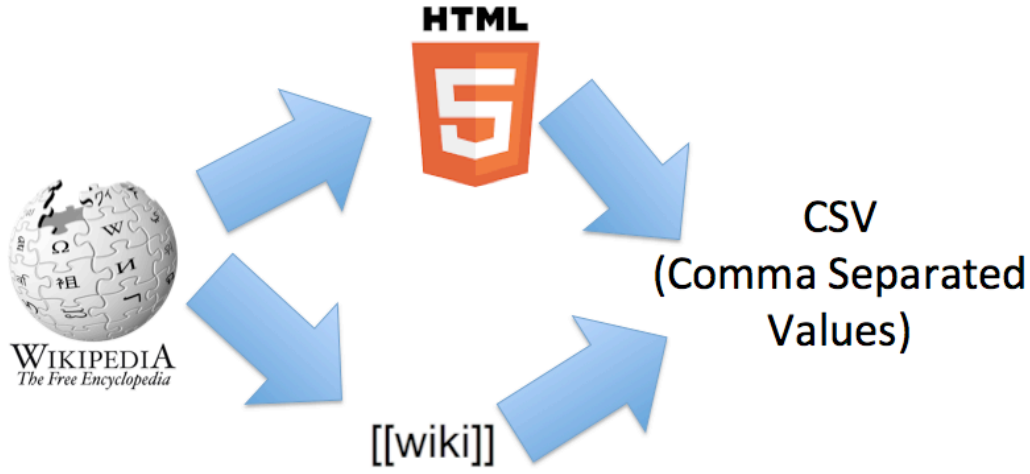
MIAGE 1
Projet Développement Logiciel –
Responsable : Mathieu Acher
Contact : mathieu.acher@irisa.fr
2019/2020

Gestion de projet – Projet Développement Logiciel (PDL)

Projet: « Wikipedia Matrix : The Truth»

Ressources : https://github.com/acherm/PDL1920
http://blog.mathieuacher.com/WikipediaMatrixChallenge/

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. 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.

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.

Une des grandes leçons de l'année dernière (cf http://blog.mathieuacher.com/WikipediaMatrixChallenge/) a été la difficulté à valider et automatiquement tester les extracteurs. Comment s'assurer que la procédure d'extraction fonctionne correctement pour n'importe quelle page Wikipedia ?

C'est le grand défi de cette année : vous allez devoir implémenter une solution pour trouver et spécifier une vérité terrain (« ground truth ») et ainsi pouvoir évaluer différents extracteurs en les confrontant à la vérité terrain.

Plutôt que d'implémenter des extracteurs HTML ou Wikitext à partir de zéro, vous allez considérer les projets de l'année dernière.

Le premier objectif est de démontrer que les extracteurs développés jusque là ont des manques, des faiblesses, des « bugs ». Vous allez écrire des tests automatiques (des programmes donc) qui mettront en exergue ces dysfonctionnements.

Le deuxième objectif est ensuite d'améliorer les extracteurs existants. Vous décrirez vos améliorations et vous évaluerez le gain obtenu, en démontrant par exemple que vos modifications passent la suite de test (alors que les précédentes implémentations non).

Le troisième objectif est de concevoir un ensemble d'outils pour pouvoir analyser les résultats des extracteurs et ainsi spécifier un ensemble de résultats attendus (qui sera utilisé ensuite lors de la phase de test automatique). Concrètement, étant donné une page Wikipedia, on aimerait pouvoir spécifier le résultat attendu (e.g., au format CSV). Or, cette tâche est fastidieuse : écrire à la main des centaines d'entrée n'est pas viable. Aussi, vous développerez un outil qui permet de visualiser une matrice (résultant d'un extracteur automatique), éventuellement de corriger la matrice, et ensuite l'exporter au format CSV. Le bénéfice en retour sera que la tâche du développeur est simplifiée : cela consistera simplement à visualiser/évaluer des matrices avec un outil dédié.

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

La première chose est de prendre en main le projet de l'an passé qui vous est assigné. Il faut le faire fonctionner le plus rapidement possible, le comprendre, et commencer son « audit ».

En parallèle, une lecture intensive de http://blog.mathieuacher.com/WikipediaMatrixChallenge/ est nécessaire pour bien comprendre la difficulté et les challenges de l'extraction.

Une fois que le sujet et le projet sont bien maîtrisés, il faudra passer à l'écriture de tests automatiques pour démontrer les manques des extracteurs. Il vous faut construire une suite de tests la plus complète possible.

Très rapidement, vous vous rendrez compte qu'il est idéalement nécessaire de collecter une vérité terrain avec le résultat attendu pour chaque page. Comme cette tâche est fastidieuse et non triviale, il vous faudra développeur une suite d'outils.

Vous utiliserez alors vos propres outils pour fabriquer une suite de tests et une vérité terrain, permettant ainsi d'améliorer vos extracteurs au fil de l'eau, tout en démontrant la fiabilité de votre solution.

Comment commencer ?

Outre la mise en place d'un repository Github, étudier Wikipedia et notamment les APIs, parsers, etc. qui sont proposés. Lire
http://blog.mathieuacher.com/WikipediaMatrixChallenge/

Tester et évaluer la pertinence du projet dont vous êtes en charge. Itérer, tester.

En parallèle de vos expérimentations, écrire le document de spécification.

Résultat final

Il y aura trois résultats concrets:

- Des extracteurs de bien meilleure qualité (avec code source, documentation, suite de tests, intégration continue, etc.)
- Une suite d'outils pour pouvoir spécifier plus facilement une vérité terrain et ainsi aider à l'évaluation des extracteurs
- Un jeu de données réutilisable par n'importe qui voulant tester un extracteur de tableaux