WPTokens

Léo Joubert

WPTokens

Une base pour l'analyse de la régulation de Wikipédia

Léo Joubert – DySoLab – Université de Rouen
leo.joubert@univ-rouen.fr

Contexte

Wikipédia une des encyclopédies de référence dans les usages des internautes. Contrairement à une encyclopédie classique, où l'écriture des articles est supervisée par un comité entouré d'experts, l'écriture de Wikipédia repose sur la mobilisation d'une communauté de contributeurs. Cette communauté utilise un « wiki », c'est-à-dire un système d'écriture collaborative où chacune des contributions s'ajoute les unes aux autres pour former des textes.

Analyser sociologiquement la régulation de Wikipédia demande donc d'abord de déplier ce processus, en se posant des questions permettant de décrire le flux de coécriture : « qui ajoute quoi, quand et où ? » et « qui supprime quoi, quand et où ? ».

C'est pour répondre à ces deux questions que nous avons construit la base *WPtokens*. Nous procèderons pour cela, après avoir nettoyé le texte (cf. « *Pre-processing* du texte ») et filtré les pages (cf. « champ de la base de données »), a une comparaison des versions consécutives des pages (*diff*) en stockant au fil du temps les mots ajoutés et supprimés par chaque version. **L'individu statistique de la base est donc un mot, qui est attaché à deux versions de page : celle où le mot a été ajouté et celle où il a éventuellement été supprimé.**

Champ de la base de données

Dans cette première version de la base, nous nous sommes focalisés sur la version francophone de Wikipédia. Le but étant de permettre une analyse sociologique de l'écriture des articles de l'encyclopédie wikipédienne, l'analyse des différentes versions sur texte se limite aux articles en négligeant l'ensemble des pages de discussion et de délibération. **Seuls les articles sont versionnés.**

Pour autant, et dans la mesure où les pages de discussions sont utiles pour comprendre la dynamique sociale d'écriture d'un article, nous avons également inclus dans la base de données des éléments de cadrage à propos des pages de discussions.

Un filtre est ensuite appliqué parmi les versions d'une page, où nous retirons les versions révoquées (*reverts*), c'est-à-dire les versions qui ont été intégralement effacées d'un clic par un autre contributeur. Cette décision a été prise, car la révocation, si elle est incontestablement un effacement des contributions réalisées par les autres, ne relève pas *a priori* du même processus sociologique qu'un effacement ou ajout après lecture. Pour être certains de pouvoir mettre à l'épreuve cette hypothèse dans le futur, nous avons néanmoins conservé les données de cadrage sur les révocations (cf. « structure de la base de données).

Pre-processing du texte

Dans cette première version de la base, le nettoyage le plus simple est réalisé pour ne conserver que les mots pleins composant les textes wikipédiens. À partir du texte des articles wikipédiens, voilà donc ce qui est supprimé :

- Les marques de « wikicode » servant à délimiter les titres, les liens hypertextes, les modèles, les illustrations, etc.
- Les mots vides du français comme « le », « la », « du », etc.
- Les marques de ponctuations.

Figure 1 – extrait de texte provenant d'un article

Le **nationalisme** est un principe politique apparu à la fin du XVIII[e] siècle, tendant à légitimer l'existence d'un État-nation pour chaque peuple, défini par des caractéristiques propres et communes à ses membres, comme une langue, des traditions historiques et culturelles ou des valeurs politiques[1].

Le texte de la figure 1 est codé de la manière suivante en wikicode :

```
Le '''nationalisme''' est un principe politique apparu à la fin du {{s-|XVIII}}, tendant à légitimer l'existence d'un [[État-nation]] pour chaque [[peuple]], défini par des caractéristiques propres et communes à ses membres, comme une [[langue]], des traditions historiques et culturelles ou des [[Valeur (sociologie)|valeurs politiques]]<ref>{{harvsp|Gellner|1989 (1983 pour l'édition anglaise)|loc=chapitre 1}}. Le terme de ''principe politique'' est repris dans {{harvsp|Hobsbawm|1992|loc=chapitre I}} et revendiqué comme venant du texte de Ernest Gellner.</ref>.
```

Après notre traitement, voilà ce qu'il advient de ce même texte :

```
nationalisme principe politique apparu fin XVIII tendant légitimer existence État-nation chaque peuple défini caractéristiques propres communes membres langue traditions historiques culturelles valeurs politiques terme principe politique repris dans revendiqué venant texte Ernest Gellner
```

Structure de la base de données

Le résultat de l'opération des *diffs* est stocké sous la forme d'une base de données relationnelles comprenant 6 tables : *page, reverts, revisions, user, map et analytics*. La description de chacune des tables est disponible sur le dépôt GitLab figurant dans les ressources documentaires.

Tableau 1 – cardinalité de la base de données

Nombre d'articles	2 347 315
Nombre total de page (incluant les articles et les pages de discussion non versionnées)	9 787 988
Nombre de mots issus du versionning de l'ensemble des articles	2 456 191 888

Comment obtenir la base de données ?

Du fait de sa lourdeur, le fichier occupant autour de 400Go, il est nécessaire de faire la demande par mail à leo.joubert@univ-rouen.fr pour que puisse être mis en place un lien de téléchargement spécifique pour une durée limitée. À moyen terme, une solution de stockage pérenne sera mise en place pour la diffusion du jeu de données.

Une fois la base de données mise à disposition, il vous sera demandé de systématiquement citer ce document en complément de toute réutilisation.

Limitations du jeu de données

La construction actuelle de WPtokens comprend deux types de limites. Les premières sont liées à la mobilisation des *dumps* comme matériau pour la construction de la base. Ainsi les pages supprimées avant la création du *dump* ne figurent pas dans la base de données. Dans la mesure où le *dump* francophone a été le seul à être parsé ici, seules les pages francophones sont incluses. Cette limite est cependant facile à lever dans la mesure où le code est construit pour fonctionner sur l'ensemble des langues pour peu que le *dump* soit disponible.

Dans les prochaines versions, le nettoyage devra être raffiné de façon à mieux tenir compte des spécificités des textes wikipédiens, probablement en tenant compte dans la base de données des modifications liées au formatage.

Si le stockage en base relationnelle sous SQLite s'est dans un premier temps imposé à nous pour des raisons de simplicité d'usage, il est clair que cette technologie n'est pas adaptée WPtokens pour des raisons relatives à la volumétrie des données et à l'usage qui en est fait. Dans le futur, nous nous tournerons vers des techniques plus adaptées comme Parquet.

Ressources complémentaires

Adresse de téléchargement des dumps
https://dumps.wikimedia.org/frwiki/

Dépôt contenant les scripts de *parsing* des *dumps*
https://gitlab.huma-num.fr/ljoubert/wptokens