Exercices : travailler avec les fichiers textes sous Unix
Application au traitement du wikicode

Franck Sajous (CLLE-ERSS, CNRS & Université de Toulouse 2)

Les données à télécharger sont accessibles à l'adresse : `http://fsajous.free.fr/`

1 Observation de fichiers volumineux

1. Récupérez et décompressez un extrait du dump du Wiktionnaire : (`bzip2 -d` *nomFichier.bz2*)
2. Observez le contenu du fichier. Il est trop volumineux pour l'ouvrir dans un éditeur de texte. Ouvrez un terminal et maximisez votre fenêtre. Puis faites afficher et défiler le fichier avec la commande `more` ou `less`. Une autre possibilité est d'extraire les N premières (ou N dernières) lignes du fichier dans un autre (moins volumineux, donc), pour l'observer dans un éditeur de texte.
3. Repérez les balises qui délimitent les articles, les titres, etc.

2 Manipulations

1. Extrayez la nomenclature (ici, le titre des articles) correspondant à l'extrait du dump.
2. Extrayez la nomenclature en supprimant les pages "méta" (ces pages ont un préfixe qui contient le caractère `:` dans leur titre). Donnez la taille de cette nomenclature.

3 Dump et pages du Wiktionnaire

1. Extrayez dans un fichier le contenu de tel ou tel article (titre que vous avez repéré dans la nomenclature). Méthode : pour extraire la portion du fichier qui correspond exactement à l'article, il vous faudrait recourir à un programme Perl, ou une commande Unix telle que `awk`, que nous n'avons pas étudiée. En revanche, il vous est possible de repérer dans votre fichier la ligne correspondant au titre de l'article que vous cherchez et sélectionner N lignes avant et M lignes après (→ commande `grep`). Sélectionnez, par exemple, 1000 lignes après le titre et écrivez la sortie dans un fichier. Vous pouvez alors ouvrir ce fichier dans un éditeur de texte.
2. Une fois que vous avez le code de votre article sous les yeux, ouvrez un navigateur et observez la page correspondant à cette entrée dans le Wiktionnaire
3. Dans le Wiktionnaire, cliquez sur l'onglet *Modifier le wikicode* et comparez avec le code que vous avez extrait

4 Analyse et conversion du wikicode

4.1 Sections, sous-sections et patrons

Les sections de langue sont repérées par des lignes telles que : `== {{langue|fr}} ==` ou `== {{langue|en}} ==` Les sections de second niveau sont repérées par un triple signe égal `===` en début et fin de ligne.

Un patron est une portion de wikicode qui se trouve entre `{{` et `}}`. On en trouve dans les titres de section, comme dans `=== {{S|nom|fr|num=1}} ===` où le patron `S` prend plusieurs arguments et introduit ici un titre de section (ici, "nom"), ou comme le patron de traduction `trad+` : `{{trad+|en|moss}}` qui prend en paramètre une langue (en pour English) et un mot *"moss"*.

1. Extrayez les type de section (premier paramètre du patron `{{S|...}}`) et triez-les par fréquence décroissante.
2. Combien votre extrait de dump compte-t-il de noms ? De verbes ? D'infinitifs ? De formes verbales fléchies ?
3. Déterminez quels sont les patrons les plus fréquents (en dehors de ceux indiquant une section de langue et un titre de section/sous-section). Vous pouvez procéder par étapes :
   (a) sélectionnez les lignes qui contiennent des patrons (`{{ ... }}`) ;
   (b) supprimez tout ce qui se trouve avant le patron, puis supprimez tout ce qui se trouve après ;
   (c) supprimez les éventuels paramètres du patron ;
   (d) triez les patrons, comptez leur nombre d'occurrences et triez-les par ordre décroissant.

4.2 Langue des étymons

1. Comment sont signalées les étymologies ?
2. Extrayez les étymologies dans un fichier.
3. Quel patron indique la langue des étymons ?
4. Triez la langue des étymons par fréquence décroissante (on parle ici du code composé la plupart du temps de deux ou trois caractères).

4.3 Extraction et conversion des gloses

Cette partie propose de convertir les gloses du format wiki au format texte.

1. Comment sont signalées les gloses ? Combien votre extrait compte-t-il de sens (un sens = une glose) ?
2. Extraire l'ensemble des gloses de votre extrait dans un fichier.
3. Comment sont encodés les hyperliens ? Produisez un fichier dans lequel le formatage des hyperliens est supprimé (seul le texte des liens doit subsister). Supprimez les liens vers des fichiers ou des images.
4. Comment est encodé le gras ? L'italique ? Le cumul des deux mises en forme ? Produisez un fichier dans lequel ces mises en forme sont supprimées.
5. Après ces traitements, quels sont les 10 patrons les plus fréquents qui restent dans vos gloses ? Tentez de les traiter pour produire une version texte des gloses la plus proche de celle que l'on peut lire dans les pages du Wiktionnaire.

Si vous êtes à l'aise avec cette partie, vous pouvez aussi encoder le formatage, plutôt que de le supprimer, avec des balises XML de votre choix.

5 GLÀFF

Téléchargez et décompressez le lexique GLÀFF (`tar xjf  GLAFF-1.2.1.tar.bz`).

1. Combien de formes comporte ce lexique ? Combien de lemmes ?
2. Triez les étiquettes morphosyntaxiques par fréquence décroissante.
3. Donnez les formes fléchies (uniquement) du verbe *googler*.
4. Donnez les formes fléchies, avec leurs étiquettes morphosyntaxiques, du verbe *hacker*.
5. Donnez pour chaque verbe le nombre de formes fléchies contenues dans le lexique.
6. Dressez une liste de verbes surabondants et une liste de verbes défectifs (dans ce lexique).

6 Comparaison de les nomenclatures de GLÀFF et de Morphalou

1. Téléchargez et décompressez Morphalou 2.0.
2. Extrayez ses formes.
3. Construisez :
   (a) la liste des formes qui sont dans Morphalou et pas dans GLÀFF
   (b) la liste des formes qui sont dans GLÀFF et pas dans Morphalou

7 Conversion de GLÀFF en format CSV

Le format CSV (comma separated values) est une format d'échange entre tableurs. Chaque champ est séparé par une virgule. Produisez une version de GLÀFF qui contienne : la forme, son étiquette morphosyntaxique, son lemme son éventuelle transcription en API (*i.e.* éliminez les transcriptions SAMPA et les informations fréquentielles).

Convertissez cette dernière version en format CSV, *i.e.* faites en sorte que les champs soient séparés par des virgules au lieu des caractères `|`. Suffixez le nom du fichier résultat par l'extension `.csv`

Vous pouvez désormais ouvrir ce fichier (ou ses N premières lignes) avec OpenOffice/LibreOffice.