CES Data Scientist, Télécom ParisTech

HBase

Pierre Senellart (pierre.senellart@telecom-paristech.fr)

13 juin 2016

Le but de ce TP est de créer à partir des articles de Simple English Wikipedia (une version dans un anglais simplifié de Wikipedia, qu'il faudra crawler, depuis un site local) un index inversé, lui aussi stocké dans HBase. Dans ce TP, nous construirons cet index de manière centralisée. Nous verrons dans une session ultérieure comment paralléliser la construction de l'index.

Pour évaluation de la session, vous devez envoyer à Pierre Senellart <pierre.senellart@telecom-paristech.fr> vos programmes Python au plus tard le 3 juillet (pénalités en cas de rendu après cette date, et aucun rendu accepté à compter du 10 juillet).

Nous utiliserons le langage Python pour ce TP. Vous pouvez utiliser l'éditeur de votre choix, Scrapy étant par exemple préinstallé.

Nous utiliserons le module python HappyBase pour accéder à HBase en lecture et écriture. La documentation de ce module est disponible à l'URL `http://happybase.readthedocs.org/en/latest/`.

1 Prise en main de la machine virtuelle

Le site pédagogique comporte un lien vers une machine virtuelle VirtualBox comprenant une installation de Hadoop, HBase et Spark (pour le TP suivant) prête à l'emploi. En raison de limitations liées à l'environnement du TP, ces installations fonctionnent en mode « pseudo-distribué » : les données et calculs sont sur une seule et même machine, mais l'accès aux données et l'exécution du calcul se fait de la même manière qu'en mode distribué, avec des accès réseaux simulés par une connexion à `localhost`.

Importer la machine virtuelle dans VirtualBox. Avant de la lancer, vérifier la quantité de RAM allouée à cette machine virtuelle, et positionnez-la à une valeur adéquate en fonction de la quantité de RAM disponible sur votre ordinateur. Démarrer la machine virtuelle (en cas de crash au démarrage, mettre à jour votre version de Virtualbox en suivant les instructions de la page `https://www.virtualbox.org/wiki/Downloads`). Le système est un Debian Linux, avec un utilisateur préconfiguré, `user`, dont le mot de passe est `cesds2015`. Cet utilisateur est *sudoer* mais il ne devrait pas être nécessaire d'utiliser les droits d'administrateur au cours de cet atelier.

1.1 Configuration du clavier

La machine virtuelle est configurée pour un clavier français standard. Si vous avez un autre type de clavier, par exemple un clavier Apple américain, vous pouvez le configurer avec la commande (dans un terminal) :

```
setxkbmap -layout us -variant mac
```

(adapter en fonction du type de clavier). Pour rendre ce changement permanent, vous pouvez ajouter cette ligne à la fin du fichier `.config/lxsession/LXDE/autostart` à partir de votre répertoire racine.

1.2 Spyder

La version installée de Spyder a un bug, qui l'empêche de se connecter à iPython. Il est possible de contourner ce bug en utilisant le menu « Consoles » / « Connecter à une console existante » et en choisissant un noyau existant, démarré par Spyder au démarrage. Cependant, il est recommandé (et parfois nécessaire, par exemple pour scrapy) pour ce TP de lancer directement les programmes dans un terminal.

1.3 Environnement Hadoop

Une fois la machine démarrée, utiliser le script `start_hadoop.sh` dans un terminal (raccourci sur le bureau) pour lancer le serveur Hadoop au nom de l'utilisateur `user`. Vous pouvez consulter l'URL suivantes pour des informations sur la configuration HDFS (système de fichier distribué) `http://localhost:50070/`.

Vous pouvez tester l'installation HDFS en tapant :

```
hadoop fs -ls /
```

dans un terminal. Vous devriez voir le contenu du répertoire racine du système de fichiers HDFS. Vous pouvez utiliser `hadoop fs` pour manipuler en ligne de commande les fichiers sous HDFS. Voir la documentation avec `hadoop fs -help`.

2 Crawler un site Web et le stocker dans HBase

Dans la machine virtuelle, vous trouverez à l'URL `http://localhost/` une copie locale du site de Simple English Wikipedia. Le but de cette partie est de récupérer l'ensemble du contenu textuel de chaque article Wikipedia contenu sur ce site, et de les stocker au sein d'une table dans HBase.

Nous utiliserons le module Python scrapy pour réaliser le crawler. Vous pouvez consulter la documentation de ce module `http://doc.scrapy.org/en/1.0/index.html` et vous pouvez vous inspirer de l'exemple de crawler de la page d'accueil de Scrapy `http://scrapy.org/`. Pour lancer un crawler Scrapy créé dans un fichier `moncrawler.py`, il faudra lancer la ligne de commande `scrapy runspider moncrawler.py`.

Votre programme doit avoir les fonctionnalités suivantes, pas nécessairement codées dans cet ordre :

— Vous devez créer une table HBase afin d'y stocker les articles.

— Ne pas oublier de vider cette table chaque fois que vous redémarrez le programme (le plus simple est de la supprimer et de la recréer avec les méthodes HappyBase adéquates).

— Vous démarrerez le crawl depuis l'URL `http://localhost/`.

— À part la page d'accueil, vous ne vous intéresserez qu'aux URL commençant par `http://localhost/articles/`, et vous ne vous intéresserez pas aux URL contenant `%7E` (l'encodage d'un deux-points, qui apparaît dans le titre des pages de catégories, d'images, de discussions, et pas de celles des articles).

— Depuis une page d'articles, vous pouvez extraire le texte pertinent à l'aide de l'expression XPath « `//div[@id='bodyContent']//*[self::p or self::ul]//text()` ».

— Vous pouvez récupérer tous les liens classiques d'une page Web à l'aide de l'expression XPath « `//a/@href` ».

— HappyBase stocke des chaînes binaires d'octets, tandis que Python s'attend à manipuler des chaînes de caractère. Pour passer de l'un à l'autre et réciproquement, vous pouvez utiliser `octets.decode('utf-8')` et `caracteres.encode('utf-8')`.

— Comme le crawl se fait en local, il est inutile d'imposer des contraintes de politesse de crawl (mais ce serait indispensable en situation réelle !).

3 Construire un index inversé dans HBase

Implémenter la construction d'un index inversé des articles de Wikipedia stockées dans HBase, qui doit avoir les fonctionnalités suivantes, pas nécessairement codées dans cet ordre :

— Vous devez créer une table HBase afin d'y stocker l'index.

— Ne pas oublier de vider cette table chaque fois que vous redémarrez le programme (le plus simple est de la supprimer et de la recréer avec les méthodes HappyBase adéquates).

— Chaque article de Simple English Wikipedia doit être traité tour à tour ; pour des raisons de performance, pendant le développement, limitez-vous aux 1000 premiers articles.

— Pour le découpage du terme en termes (*tokenization*), on pourra utiliser la commande suivante : `it = re.finditer(r"\w+",text,re.UNICODE)` qui met dans `it` un itérateur vers l'ensemble des suites de caractère alphanumériques d'une chaîne de caractères `text`.

— Les mots vides (stop words) ne sont pas ajoutés dans l'index ; se reporter à la liste de mots vides fournie sur le site pédagogique (et dans votre répertoire personnel dans la machine virtuelle)

— Les termes sont racinisés avec l'algorithme de racinisation de Porter (déjà implémenté dans le module Python `stemming`, pré-installé).

— Chaque terme doit être accompagné de son score tf–idf.

— Les mises à jour sur la base doivent se faire en *batch* pour des raisons de performance (voir la documentation de HappyBase).

— HappyBase stocke des chaînes binaires d'octets, tandis que Python s'attend à manipuler des chaînes de caractère. Pour passer de l'un à l'autre et réciproquement, vous pouvez utiliser `octets.decode('utf-8')` et `caracteres.encode('utf-8')`.

Une fois le développement terminé, tenter d'enlever la limitation aux 1000 premiers articles. Que constatez-vous ?

4 Interrogation de l'index inversé

Implémenter un programme Python qui demande à l'utilisateur (à l'aide de la fonction `raw_input`) une requête par mots-clefs, et qui exécute cette requête sur l'index inversé. On implémentera une sémantique conjonctive (tous les mots de la requête qui ne sont pas des mots vides doivent être présents), en utilisant l'addition pour combiner les scores de chacun des termes.

Afficher les 10 résultats les plus pertinents, par ordre décroissant de pertinence.

Tester votre programme avec la requête « Which associations are both Singapore and Brunei in ? » Si tout va bien, vous devriez récupérer, bien classés, les articles Simple English Wikipedia sur ASEAN et le Commonwealth (si ces deux articles ont été traités dans l'index).