	MatrixSynthesizeWikipedia	Ref : PDL-Project3-Group4 Issue : 1 Révision : 0 Date : 20/12/2017
-----------------------------------------------------------------------------------	----------------------------------	-----------------------------------------------------------------------------

I - Présentation du projet

Le but de ce projet est de générer des matrices de comparaison au format “.csv” à partir des données de Wikipédia ou Wikidata (au choix). Aucune données ne devra être saisies manuellement, l'utilisateur devra simplement renseigner les champs des valeurs à récolter. Ces matrices seront ensuite importées dans “OpenCompare” qui se chargera de la mise en forme de celles-ci. Après exécution de votre programme Java, vous devriez donc obtenir un fichier “.csv”


2 démarche du projet :

2-1 Présentation de wikidata:

Wikidata est une base de données participative basée sur des technologies web. Elle a été lancée en 2012 par Wikimedia. Elle permet de centraliser des données présentes dans différents projets tels que Wikipédia. Contrairement aux encyclopédies classiques, Wikidata prend la forme de données brutes, à savoir des couples clés/valeurs sans rédaction autour des informations. Par exemple dans le cas d'un pays, on trouvera les clés “Langue”, “Taille”, “Population” associées aux valeurs “Français”, “5000 Km²”, “30 000 000”.

2-2 Récupération des données de Wikidata:

il y a plusieurs façon d'accéder aux données de Wikidata. Par exemple, il est possible d'obtenir les données pour un élément ou bien d'obtenir l'intégralité des données grâce aux fichiers de dumps, ou encore, des exportations périodiques des données de la base dans un ou des fichiers téléchargeables. Chaque élément présenté par Wikidata possède une clé unique, un id permettant de retrouver facilement ces données. Cette clé prend la forme suivante : la lettre “Q” suivie d'un numéro (identifiant unique). Pour illustrer notre explication, avec la page Wikidata de la France, on obtient la clé “Q142”. Celle-ci est récupérable à plusieurs endroits : sur la page elle-même ou à côté du titre (voir ci-dessous), et donc accessible via son url et son code source.

	MatrixSynthesizeWikipedia	Ref : PDL-Project3-Group4 Issue : 1 Révision : 0 Date : 20/12/2017
-----------------------------------------------------------------------------------	----------------------------------	-----------------------------------------------------------------------------



Clé de l'élément d'une page Wikidata.

Ensuite, une fois la clé obtenue, il est possible de récupérer toutes les données de la page par l'intermédiaire de son json. Celui-ci est accessible par une url, suivant le modèle ci-après : <http://www.wikidata.org/wiki/Special:EntityData/Q142.json> (exemple des données JSON pour la France).

2-3 API wikidata:

-Wikidata propose une API "officielle" qui est disponible sur Github via le lien suivant : <https://github.com/Wikidata/Wikidata-Toolkit>.


Cette API est en fait une bibliothèque Java permettant d'accéder aux APIs Wikidata. Il s'agit d'une API conséquente et complexe, dont la plupart des fonctionnalités ne sont pas en adéquation avec nos problématiques.

2-3 Bibliothèques utilisées:

-Plusieurs librairie externe pour Java sont disponibles pour manipuler les fichiers CSV. Nous avons utilisé Commons CSV d'apache. Cette librairie fournit beaucoup de fonctionnalités. Comme le traitement de fichiers volumineux.(voir annexe 1)

-Nous avons aussi utilisé deux bibliothèques JSON(JSON+JSONSIMPLE):

-JSON (JavaScript Object Notation) est un format léger d'échange de données. Il est facile pour les humains de lire et d'écrire. Il est facile pour les machines d'analyser et de générer. Il est basé sur un sous-ensemble du langage de programmation JavaScript, Standard ECMA-262. JSON est un format de

	MatrixSynthesizeWikipedia	Ref : PDL-Project3-Group4 Issue : 1 Révision : 0 Date : 20/12/2017
-----------------------------------------------------------------------------------	----------------------------------	-----------------------------------------------------------------------------

texte complètement indépendant du langage mais utilisant des conventions familières aux programmeurs de la famille des C, y compris C, C++, C#, Java, JavaScript, Perl, Python, et bien d'autres. Ces propriétés font de JSON un langage d'échange de données idéal.

2-4 problèmes rencontrés durant le projet:

-le plus complexe dans l'utilisation de Wikidata est la récupération de la clé de la page associée à la recherche saisie par l'utilisateur. De plus, il faudra mettre en place un système permettant de choisir le sens le plus pertinent d'un mot lorsque celui-ci possède des homonymes.

-le problème a été réglé en proposant à l'utilisateur de choisir la ligne (donnée) qu'il souhaite.

2-5 Conduite de projet:

La conduite du projet a été confiée à Jeremy Cavron.

on peut dire quand à réaliser notre projet suivant plusieurs étapes:

-L'étape d'observation correspond à les rencontres quand à fait durant le projet.

-L'étape de cohésion ou on a constitué une équipe soudée.

-L'étape de différenciation qui nous a permis de tirer parti des différences entre les membres de l'équipe.

-L'étape d'organisation ou on a utilisé les techniques traditionnelles de la gestion de projet.

-L'étape de production ou on a mis en disposition la stratégie qu'on va suivre pour réaliser le projet.