UNIVERSITÉ DE LORRAINE | UFR MATHÉMATIQUES ET INFORMATIQUE

L3 MIASHS 2017
Ingénérie Linguistique
Projet 1

Qui a tué qui ? et quand ? et où ?

Le but du projet est de programmer un enquêteur qui doit traverser WikiPédia pour déterminer qui a tué qui ? et quand ? et où ?

Le travail sera réalisé en groupe de 3-4 étudiants. Une partie importante du projet sera à réaliser pendant les TD, mais ça ne suffira pas. Vous devrez travailler en dehors des cours et des TD.

Objectif

Votre enquêteur devra présenter une liste de 50 assassins, ainsi que les informations sur le meurtre : qui a été tué, quand et où. Les informations doivent être automatiquement extraites de Wikipédia et les noms des meurtriers doivent commencer par la même lettre.

Évaluation

Vous serez essentiellement évalués sur votre capacité à atteindre l'objectif et sur votre aptitude à travailler efficacement en équipe. évidemment, la composante résolution de problèmes et TAL du projet est la plus importante.

Attention : ce projet est l'unique évaluation de ce cours. Vous obtiendrez des notes différenciées en fonction de votre investissement dans le projet. La note finale sera composée de :

— une note d'investissement dans les TD
— l'évaluation d'un rapport décrivant le travail réalisé
— une présentation orale du travail (avec des transparents)
— la qualité globale de la tenu du travail

Vous devrez avoir constitué votre corpus, segmenté les données, appliqué un POS-tagger, reconnu les entité nommés et identifié les indices temporels.

Rédaction

Parallèlement au programme, nous vous demandons de rédiger un rapport de 6 à 10 pages. Ce travail de rédaction est essentiel pour votre évaluation. Dans ce compte- rendu doit figurer un rappel des différentes étapes du projet, avec les choix que vous avez fait, avec éventuellement quelques éléments historiques et des illustrations.

Vous détaillerez les techniques génériques de TAL utilisées que vous avez employées pour programmer votre enquêteur artificiel. Vous expliciterez les choix d'implémentation qui vous ont posé des difficultés et/ou qui vous ont semblé importants.

La présentation du rapport devra être soignée. L'utilisation de LATEX est recommandée. Reportez-vous au calendrier pour connaître les dates de rendus.

Organisation du travail : GIT

Vous devez travaillez sur Git. Vous devrez mettre en place le répertoire de travail. Le rapport final devra contenir une image de l'activité du répertoire afin de montrer la répartition du travail et la régularité de ce dernier. L'absence de cette illustration sera lourdement sanctionnée.

Choix d'une lettre

Vous ne devrez pas trouver tous les meurtriers présents dans wikipédia. Nous vous demandons d'identifer 100 meurtriers, dont le nom de famille commence par la même lettre. Ainsi chaque groupe obtiendra des résultats différents, et donc rencontrera des problèmes différents. La première étape est de donc de choisir votre lettre (vérifier que vous êtes le seul groupe avec cette lettre).

Constitution du corpus

Les applications du TAL sont généralement bâties à partir d'un corpus. C'est le cas ici. Il s'agit pour vous d'extraire une partie de Wikipédia, puis de travailler à identifier automatiquement des informations dans le texte.

La première étape est de choisir des thèmes. à partir de ces thèmes, vous extrairez automatiquement les pages wikipédia concernées par ce thème. C'est à vous de choisir vos thèmes. Donc soyez intelligent pour limiter la quantité de données à traiter, tout en ayant des données pertinentes.

1. Allez sur la page `https://fr.wikipedia.org/wiki/Sp\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{e\global\mathchardef\accent@spacefactor\spacefactor}\accent19e\egroup\spacefactor\accent@spacefactorcial:Exporter`
2. Utilisez cette seconde page pour télécharger un fichier .xml contenant des pages wikipedia sur un thème donné.
3. Allez sur la page `http://talc2.loria.fr/import-text/`
4. Chargez le fichier .xml sur cette page afin d'en extraire le texte, que vous enregistrerez dans un fichier .txt.
5. Utilisez le programme de segmentation sur ce fichier.

Traitements Automatiques

Pour parvenir au bout de votre projet, il vous faudra plusieurs outils du TAL. La majorité des outils sont développés pour l'anglais, aussi, vous travaillerez uniquement sur l'anglais.

Vous trouverez deux séries d'outils utilisable NLTK (en python) et les outils stanfordNLP (majoritairement en java). Vous travaillerez avec les outils stanfordNLP

http://nlp.stanford.edu/software/

Nous reviendrons dans une prochaine feuille sur la description d'autres traitements.