Compte Rendu
"Qui a tué qui? Où ? Quand"

Projet de Projet Ingénieurie Linguistique, Groupe F

19/04/2017

Composition de l'équipe

Equipe F :
— BIVER Alexis
— BRICHE Arnaud
— COLIN Emile
— CULARD Mathieu
— RICHIER Valère

Gestion du projet

— Mise en place d'un espace de travail collaboratif :
  — Trello
  — GitHub
  — Définition des tâches et répartition
  — Communication via groupe messenger courante
— Documentation sur l'utilisation de CoreNLP
— Acquisitions de compétences en LaTex via OpenClassroom

Liste des Tâches

Tâches :

corpus tokenxml td1

entités nommées

td1 aussi, identification des entités nommées personnes et pas locations par exemple

réductions des phrases (ATTENTION, PAS DES PHRASES C TRO PETIT) avec les personnes en F algo dictionnaires de synonymes : wordnet sur esl verbes qui nous interessent réduction des phrases avec ces verbes CoreNLP encore identitfication (PERSONNE, VERBE SYNONYME DE TUER, PERSONNE)

Constitution du corpus

Mots clés choisis :
— Fictional murderers

— People executed for murder
— Male serial killers

Après avoir parcourru les pages spéciales de wikipédia regroupant différents types de meurtriers par catégories à partir de la page https ://en.wikipedia.org/wiki/Categ derers, nous avons choisis nos mots clés de catégories parmis ceux qui contenaient une grand nombre de pages de meutriers dont le nom commence par F (lettre qui ne nous a pas semblée être parmis les plus faciles car peu courantes dans les noms de meurtriers).
D'après notre recensement manuel et approximatif sur Wikipédia dans le classement par nom, la catégorie Fictional murderers contient 24 tueurs en F, la catégorie People executed for murder en contient 21 et la catégorie Male serial killers en contient 17.
Il nous paraîssait tout d'abord important d'avoir plus de pages et de tueurs possibles afin d'être sûrs de trouver suffisement de données sur chacun pour répondre à la consigne, pour les serial killers il pouvait êrte difficile d'identifier les victimes et pour les tueurs de fictions il pouvait être difficile de trouver un lieu et un moment.
Ainsi avec ces trois catégories nous avons (naïvement) exporté via wikipedia un XML de 25,5 Mo que nous avons essayé en vain de traiter avec le site du loria fournis dans la description du projet : http ://talc2.loria.fr/import-text/ sans succès, puis avec un deuxième fichier légèrement plus léger duquel nous avons retiré les pages ne menant qu'à d'autres catégories.
Devant ces échecs ous avons alors fait le choix d'extraire uniquement les pages qui nous intéressent en les séléctionnant via les pages de catégories cités plus haut uniquement les pages de tueurs dont le nom commence par F dont le txt contenant le nom de ces pages entrés dans wikipedia est en annexe.