Qui a tué qui ? Où ? Quand ?

Présentation Projet TAL
Groupe F

Sommaire

Constitution du corpus

Mots clés choisis :

- Fictional murderers
- People executed for murder
- Male serial killers

Plusieurs essais infructueux lors du passage en txt

Sélection des pages wikipedia dans ces catégories

Tokenization

Stanford Core NLP

Exécution de la commande vue lors du td1

XML de 19,5Mo

Scripts de traitements xml

Parcours des tags

Identification des Entités Nommées

Stanford Named Entity Recognizer

Fichier tsv

EN[en numéro 1]= (“Jean Paul”, “PERSON”, “est parti manger des crêpes en”)
EN[en numéro 2]= (“Bretagne”, “LOCATION”, “avec son ami”)
EN[en numéro 3]= (“Pierre”, “PERSON”, “qu’il connaît depuis ”)
EN[en numéro 4]= (“1996”, “DATE”, “.”)

Synonymes de “Killed”

Récupération des verbes dans le xml tokenisé et taggué dans les “POS” sous différentes formes (VB, VBN, VBD, VBG & VBZ)

Lemmatisation de ces verbes puis stockage

Liste des synsets de “killed” (wordnet)

Synonymes de ces synsets

Comparaison des synonymes et des verbes du corpus

Si match entre eux, ajout du verbe du corpus dans une liste et de son index

Identification

Récupérer emplacement verbe dans une phrase

Regarder si on trouve un nom en F faisant partie de la liste d'EN dans la phrase

Si oui, regarder si on retrouve une date liée dans la phrase/le tableau d’EN

Même chose avec la victime

Si le programme ne trouve pas de victime ou de date, on regarde la phrase d'après

Vérifier l’ordre d'apparition de ces EN

Conclusion

Corpus décevant

Bon début

Jusqu’à l’identification

Mauvaise gestion de notre temps