Sujet PLDAC
Analyse de sous-titres

Vincent Guigue - Nicolas Baskiotis - prenom.nom@lip6.fr

Sujet logiciel et bibliographique

Sujet ouvert à un monôme ou un binôme

On dispose d'une base de données de sous-titres de séries télévisuelles (plus de 3000) ainsi que des informations sur les séries elles-mêmes : avis, résumé, année de production, ... L'enjeu du projet est d'explorer ce que le traitement de ces données à partir des techniques usuelles d'apprentissage automatique peut apporter.

On s'intéressera d'abord à une analyse des sous-titres, de manière à comparer le champ lexical, dresser une cartographie de l'offre, extraire les mots-clés associés à chaque série. L'analyse de ces sous-titres, à l'aide des techniques croisées en MAPSI, ARF ou TAL, permet de créer une topologie des genres de séries. D'un autre coté, internet est une source infinie d'avis en tous genre. L'idée du PLDAC est de combiner les données que nous avons à disposition avec des avis issus du web pour construire un système de recommandation efficace.

Les principales étapes sont donc les suivantes :
1) Prise en main des données de sous-titres et application des algorithmes de base pour analyser cette base et comprendre le positionnement des séries.
2) Récupération d'avis sur le web et application d'algorithmes de filtrage collaboratif.
3) Comparaison et fusion des topologies issues des deux univers.

Autour de ce cadre, de nombreuses variantes sont possibles et seront discutées avec les candidats au début du projet puis au cours de son évolution. Les développements seront principalement effectués en python. D'autres langages sont négociables en concertation avec l'étudiant.