

Octobre/Novembre 2020
Projet « RSS-Intelligence »
Sous projet « INDEXER/SEARCHER »

Livrables

Vous devrez fournir :

- 1) La mise à jour de la documentation globale de votre projet
- 2) Votre code source, avec la documentation et le diagramme des classes
- 3) Un fichier README qui décrit ce qu'est censé faire votre programme et comment l'exécuter.
- 4) Un exemple documenté d'exécutions possible de votre exécutable et des données de test.

Module “Indexer/Searcher”

Éléments de cahier des charges

On se limite au stockage local des dépêches RSS, à l'indexation **incrémentale** des éléments RSS collectés et à leur recherche en s'appuyant sur une API python3.

La réalisation de ce module sera effectué grâce à l'API Python du moteur de recherche Elasticsearch <https://elasticsearch-py.readthedocs.io/en/master/>

Un serveur Elasticsearch devra être téléchargé et installé à partir de <https://www.elastic.co/fr/>

Une introduction sommaire sur cette API, vu en cours, est accessible sur l'ENT.

L'interopérabilité avec les modules *Collector* et *Classififier* (*Classififier* est le sous-projet suivant) devra être assurée.

Tous les champs de la structure Item_RSS devront être indexés et ainsi être interrogeable, y compris les champs Catégorie_du_flux, Catégorie_prédite, et date_de_collecte.

Une réflexion sera conduite spécifiquement sur la question de la mise en correspondance Requête/Item_RSS. Un modèle dédié de mise en correspondance pourra être développé et mis en oeuvre pour tenir compte de choix d'importance spécifiés par l'utilisateur (l'utilisateur pourrait spécifier des poids d'importance pour chacun ou parti des champs contenus dans les Item_RSS).

Un exemple d'utilisation de l'API Elasticsearch python est disponible sur l'ENT dans le sous répertoire INDEXER/SEARCHER.