Semaine 1.1

Prof absent : Gerchy

Groupe 1 : Supervision

Faire la demande aux profs pour les journaux payants (Dépêche)

Groupe 2 : BD index

API créée avec Laravel fonctionnelle mais pas encore (bientôt) sur le serveur

MCD validée : modifications à faire et des données à rajouter

Les groupes : n'hésitez pas à donner des informations pour perfectionner le MCD

Demain : Procédures, déclencheurs, API configuration serveur, vue matérialisée TF-IDF-DF, liens entre les tables et les procédures + faire un point avec Valentin.

Groupe 3

Tant que le groupe 2 n'a pas fait la BD, ils ne peuvent pas travailler (stand-by pour les requêtes)

Essayer de faire l'API – problème entre le groupe 8 et le groupe 3 (pas fonctionner – Serrurier vient demain) + procédures pour les demande du Groupe 8

Résoudre le problème de l'api + faire le point avec les besoins de Valentin

Groupe 4

Transformation des données dans le bon format json et uniformisé (utf-8)

Les titres des fichiers seront notés avec des tirets (-)

Fini Le Gorafi : récupération des articles, traitement (50 aine articles par catégorie)

La Dépêche (ils ont extraits, reste le RSS et crawler), FuturaSciences : en cours

Continuer les journaux, prendre Femina

Groupe 5

Pré-traitement des données : bien avancé (nltk limité pour le français donc recherche de nouveaux packages)

Fonction d'import selon l'arborescence des différents fichiers

Sorti du premier TF/IDF, tokenisation et lemmatisation pas parfaite mais premier jet pas mal *(dans la matinée transmettre à Clara et Geoffrey)*

Stopwords à garder : utiles pour la sémantique

Post-tagging + affinement des programmes + entités nommées + se mettre d'accord pour les TF/IDF avec la base de données

Groupe 6

Synonymes des mots avec nltk, *recherche de méthodes un peu mieux*

Traitement des entités nommés : ne pas sortir de synonymes

Polarité – utiliser l'écart à la moyenne : implémentation, sentiments des mots (puis par article), *tester de méthodes pour avoir la plus juste (utiliser Wikipedia pour des entités nommés)*

Travail sur les entités nommées : scraper les titres et les liens (super long)

Toutes les pages Wikipedia des entités nommées doivent pouvoir être accessibles à partir de la barre de recherche.

Groupe 7

Recodage des catégories

Prend les données de Maxime et on commence le boulot

Groupe 8

Chouquet valide les tests stat, Serrurier séries chro : les profs ne sont pas d'accords

Mise en forme que la BD doit donner à Valentin (Mehdi doit les mettre en forme)

Groupe 9

Maquette du site web, fonctionnalités (quasi-finie)

Définition des requêtes (appel de l'api)

Structuration du site

Groupe 10

Travail avec le groupe 9 (pour le visuel)

Charte graphique

Template des rapports