Projet de groupe

# Remarques préliminaires :

Le but est d’obtenir un logiciel utile réellement.

Prendre décisions pour l’utilisateur.

On considère chaque phrase comme un document indépendant.

Remarque : apprendre un mot = apprendre le sens d’un mot, faire poignet = lequel ?

Si on ajoute le Français on peut en déduire le sens.

On considère qu’on connait le mot plus ou moins bien quel que soit le sens

Méthode agile : observation du projet au fur et à mesure et ajout de fonctionnalité, avancement effectif et révision. Sprint à des moments, objectifs de la période suivante. Tous les matins, réunion de projet :

-avec problèmes, chercher idées ensemble, si blocage,

-trouver méthodes ensemble.

# Outils :

<http://tatoeba.org/>

<http://Linguee.fr/>

# Partie indexation

Mise en place des structures à partir de la base de données de tatoeba. La base de donnée est statique une fois crée. Structurer de façon solide.

Prétraitements :

-Préanalyser normaliser, supprimer majuscules, groupes de mot. Trouver « open » si on cherche « opened »

-Partie non traitée à afficher.

-Index de position (« to be or not to be »),

Tableau de suffixe permet d’indexer sans faire péter la mémoire et le temps de recherché.

# Critères de recherche de l’utilisateur

La partie dépendante de l’utilisateur est variable, avec un graph d’apprentissage.

Utilisateur peut préciser fenêtre en français et trouver vu qu’il apprend l’anglais, trouver les textes avec window.

Requête de base « window » ou « open window ».

Partie 0 : chercher mot clef directement, faire des tests de pondération pour obtenir des résultats le plus pertinent possible. PageRank de Google, mais aussi densité dans document. Jouer avec paramètre au départ puis figer ou laisser curseurs utilisateurs.

Partie 1 : Graph de l’apprenant, similarité entre mots, recherche d’exemple utiles. Partie graphique, avec mots qu’on connait, cliquer sur nœuds et obtenir textes. Trous dans le lexique. Catégories : « aérospatiale » « agriculture ».

# A faire

-Cahier des charges de l’API

-Document du code

# Divers

Conventions communes de nommage.