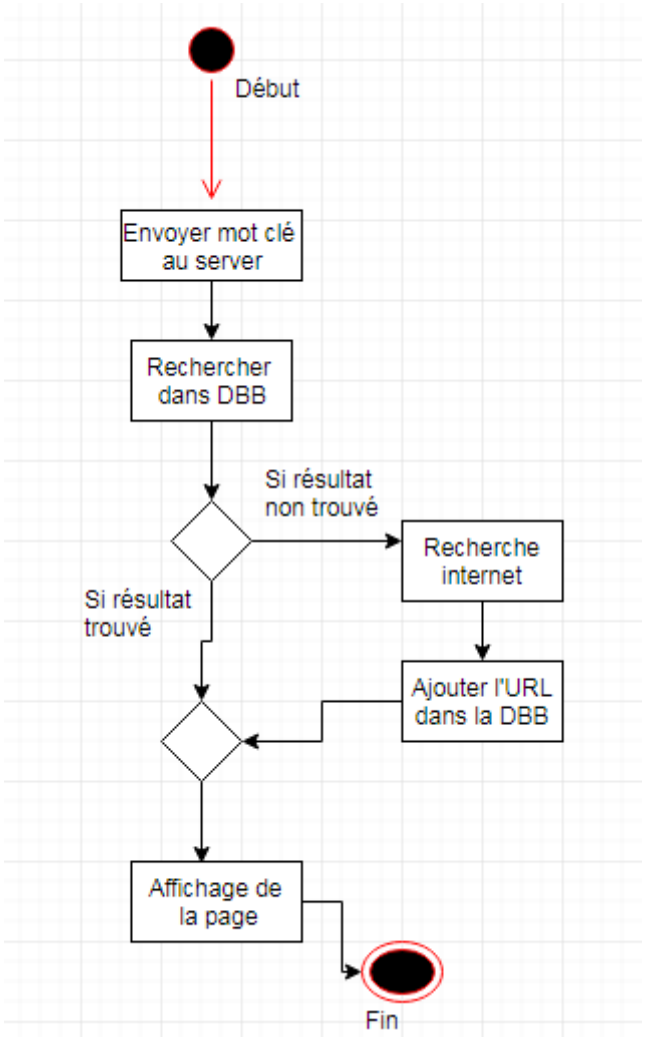
Feuille de route

Le projet consiste de réaliser un Crawler web, c'est-à-dire que c'est un logiciel qui recevra des données sous forme de mot clé comme un serveur à partir duquel, le programme devra effectuer des recherches web et trouver aussi des synonymes en fonctions de ces mots. Le crawler possède aussi une base de donnée vierge, qui se remplira par des adresses URL et les synonymes trouvés, au fur et à mesure de ses résultats sur le net. Le programme doit aussi afficher le résultat de ses recherches.

Voici un diagramme qui résumera les fonctionnement su projet :

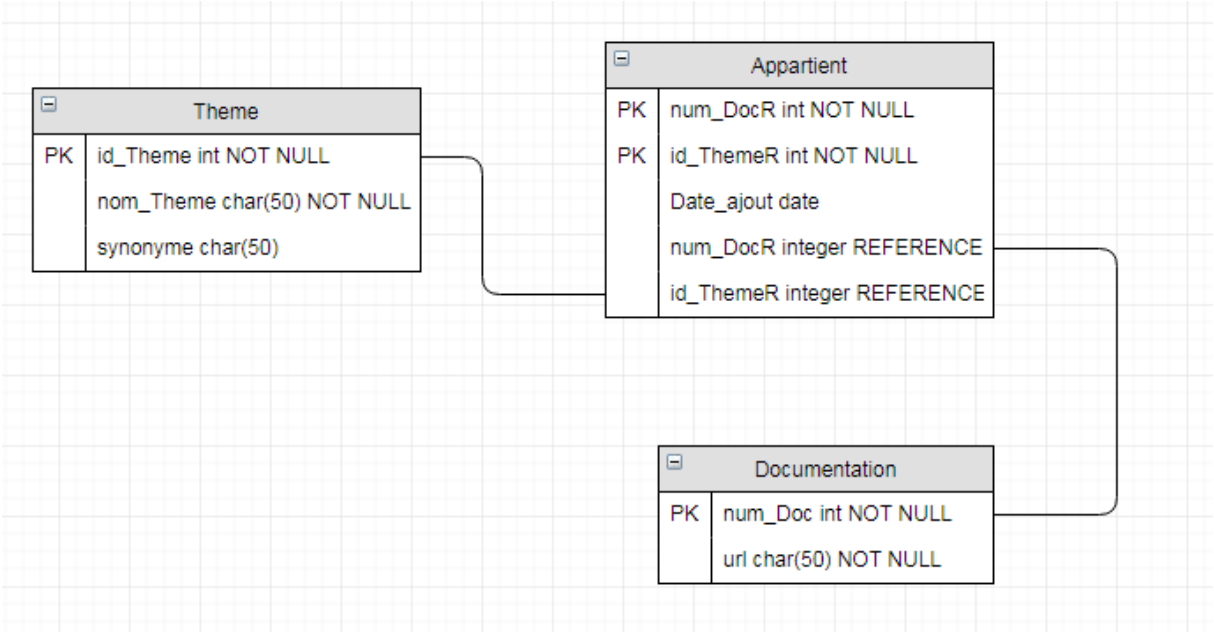


Le projet vérifiera d'abord sa base donnée s'il y a déjà des documentations sur que mot clé que le programme aura reçue. S'il y a aucun résultat dans la base de donnée, le programme effectuera des recherches sur le net afin de trouver un synonyme du mot et de trouver des documentations pertinentes sur le mot clé reçu. C'est ensuite après que les résultats de ces recherches seront stockés dans la base de donnée.

Donc le principal objectif du programme sera d'indexé une source de documentation et de synonymes via ses requêtes web.

Les fonctionnalités du crawler sont :
- communiquer (recevoir des informations et d'envoyer) ;
- réaliser des requêtes web avec plusieurs moteurs de recherche (google, yahoo, bing);
- enregistrer l'url dans la base de donnée ;
- enregistrer un synonyme dans la base de donnée ;
- rechercher dans la base de donnée ;
- être capable de fonctionner sous Linux et Windows

Le programme possède une base de donnée qui ne contiendra trois tables, qui suffisant pour le projet. Le but principal de la table sera juste de stocké des adresses URL ainsi que synonymes associé au mot clé.



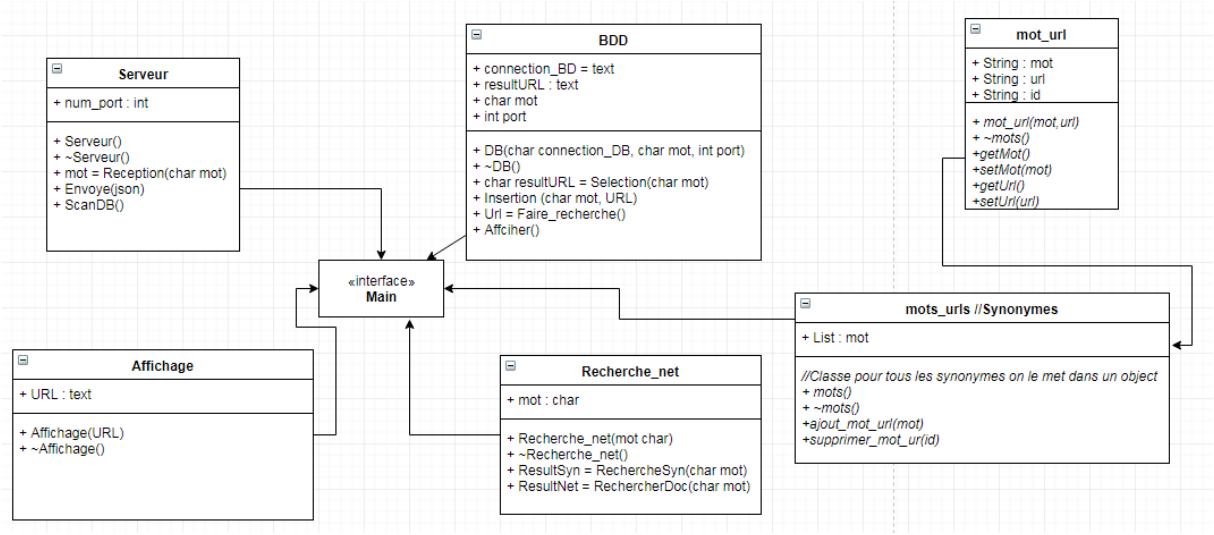
Le champs nom_Theme de la table Theme est le mot clé que le programme aura reçu. La table Documentation contient une liste de d'adresse URL, cependant le mot clé (nom_Theme) peut avoir plusieurs URL, afin pouvoir les relier, la table Appartient est créée pour réaliser dans ce but.

Le Crawler sera réalisé avec python3.5, puisque c'est un langage qui est très polyvalent, puisque c'est un langage orienté objet, donc très utile pour la modularité. De plus, ce langage est capable de liée à plusieurs SGBD, mais surtout, il est capable de faire des requêtes web beaucoup plus facilement que les autres langages et il est aussi capable de fonctionner aussi bien sur Linux que Windows. L'autre technologie utiliser sera le langage SQL, car elle est capable de réaliser des

base de donnée. Mais en vue de la petite taille de notre base de donnée, le choix du moteur Sqlite3 sera plus propice. Sqlite3 est très léger facilement transportable sur différents types d'OS.
Comme le programme doit communiquer avec les autre, il doit recevoir et envoyer des données sous le protocole TCP, cela nécessite une connexion. Il est important pour garantir une bonne réception des mots clés, sinon le programme ne fonctionnera pas.

Les étapes clé du projet sont :
- la partie conception : puisque les modules doivent communiquer sans problème avec elle ;
- la partie réception des donnée ;
- la partie enregistrement de l'URL est la plus importante du projet ;

Ci-dessous vous pouvez voir l'architecture du programme :



Il y a 6 classes :
- la classe serveur permet de principalement de recevoir le mot clé;
- la classe base de donnée permet d'insérer des donnée (les URLs), de les sélectionnées dans une base de donnée;
- la classe Recherche_net contient les fonction de connexion à internet, mais aussi des fonctions de recherche de documentation.
- la classe Sysnonyme contient les fonctions de recherches des synonymes, cependant cette doit utiliser une autre classe est mot_url, afin de créer une liste de variable ;

L'interface main permet de gérer tous les classes et conditionne le fonctionnement du programme.

Par la nature du projet, il n'y aura pas d'interface utilisateur.

Voici le planning du projet :

Num Tâche	Tâches	Début	Fin	Début reel	Fin réel	Jour retardé	Jour avance
1	Dernière spécification	24/09/2018	29/09/2018	24/09/2018	29/09/2018	Aucun	Aucun
2	Réalisation base de donnée	01/10/1993	05/10/2018	29/04/2018	29/04/2018	Aucun	2 Jours
3	Test base de donnée	05/10/2018					
4	Module python3 et DBB	09/10/2018	12/10/2018				
5	Création serveur et client	30/09/2018	06/10/2018	29/09/2018			
6	Test DBB et serveur	15/10/2018	20/10/2018				
7	Partie web : accès internet	05/11/2018	07/11/2018				
8	Partie web : mecanisme URL	07/11/2018	19/11/2018				
9	Remplissage DBB	21/11/2018	30/11/2018				
10	Tout unifier	03/12/2018	10/12/2018				
11	Test	12/12/2018	17/12/2018				
12	Soutenance	19/12/2018					

Nous pouvons voir qu'il y a déjà 2 parties qui sont en avance :
- La partie réalisation de la base de donnée ;
- La partie création du serveur ;