Ecole **N**ationale **S**upérieure

d'**I**nformatique et d'**A**nalyse des **S**ystèmes

Rapport du projet Intelligence Artificielle

Option

Sécurité des Systèmes d'Informations

Sujet

Réalisation d'un moteur de recherche

Soutenu par :

Mr.Rachid OUBAOUG

Mr.Ilyas NAJI

Année Universitaire 2016-2017

Table des matières

Introduction générale

Le développement de l'informatique a conduit à la création de nombreux supports pour l'information. On admettra qu'il est impossible pour un individu d'organiser l'information et donc de chercher manuellement une information particulière lorsque la quantité d'information disponible dépasse un certain seuil. Pour répondre à ce problème, des outils informatiques qui automatisent cette tâche, appelés moteurs de recherche, ont été développés. Leur utilisation s'illustre par exemple sur internet, où des pages web sont créées continuellement et forment une masse d'information. Les individus qui cherchent une information spécifique sur internet sans connaître son emplacement font appel à un moteur de recherche tel que Google.

Internet est donné ici à titre d'exemple mais il existe de nombreux autres supports d'information, textuels ou multimédia, qui font l'objet de développement de moteurs de recherche qui leur sont propres.

Chapitre 1

Contexte et objectifs du projet

1. Contexte du projet

1.1. Introduction:

Un moteur de recherche est une application web permettant, de trouver des ressources à partir d'une requête sous forme de mots. Les ressources peuvent être des pages web, des articles de forums Usenet, des images, des vidéos, des fichiers, etc.. Certains sites web offrent un moteur de recherche comme principale fonctionnalité ; on appelle alors moteur de recherche le site lui-même.

Ce sont des instruments de recherche sur le web sans intervention humaine, ce qui les distingue des annuaires. Ils sont basés sur des « robots », encore appelés bots, spiders, crawlers ou agents qui parcourent les sites à intervalles réguliers et de façon automatique pour découvrir de nouvelles adresses (URL). Ils suivent les liens hypertextes qui relient les pages les unes aux autres, les uns après les autres. Chaque page identifiée est alors indexée dans une base de données, accessible ensuite par les internautes à partir de mots-clés.

C'est par abus de langage qu'on appelle également moteurs de recherche des sites web proposant des annuaires de sites web : dans ce cas, ce sont des instruments de recherche élaborés par des personnes qui répertorient et classifient des sites web jugés dignes d'intérêt, et non des robots d'indexation — on peut citer par exemple DMOZ et anciennement Yahoo!.

Les moteurs de recherche ne s'appliquent pas qu'à Internet : certains moteurs sont des logiciels installés sur un ordinateur personnel. Ce sont des moteurs dits desktop qui combinent la recherche parmi les fichiers stockés sur le PC et la recherche parmi les sites Web — on peut citer par exemple Exalead Desktop, Google Desktop et Copernic Desktop Search, Windex Server, etc.

On trouve également des métamoteurs, c'est-à-dire des sites web où une même recherche est lancée simultanément sur plusieurs moteurs de recherche, les résultats étant ensuite fusionnés pour être présentés à l'internaute. On peut citer dans cette catégorie Ixquick, Mamma, Kartoo, Framabee ou Lilo.

1.2. Historique :

Les moteurs de recherche sont inspirés des outils de recherche documentaire (à base de fichiers inversés, alias fichiers d'index) utilisés sur les mainframes depuis les années 1970, comme le logiciel STAIRS sur IBM. Le mode de remplissage de leurs bases de données est cependant différent, car orienté réseau. Par ailleurs la distinction entre données formatées ("champs") et texte libre n'y existe plus, bien que commençant depuis 2010 à se réintroduire par le biais du web sémantique.

Des moteurs historiques ont été Lycos (1994), Altavista (1995, premier moteur 64 bits) et Backrub (1997), ancêtre de Google.

1.3. Fonctionnement :

Le fonctionnement d'un moteur de recherche comme tout instrument de recherche se décompose en trois processus principaux :

- **L'exploration ou crawl :** le web est systématiquement exploré par un robot d'indexation suivant récursivement tous les hyperliens qu'il trouve et récupérant les ressources jugées intéressantes. L'exploration est lancée depuis une ressource pivot, comme une page d'annuaire web. Un moteur de recherche est d'abord un outil d'indexation, c'est-à-dire qu'il dispose d'une technologie de collecte de documents à distance sur les sites Web, via un outil que l'on appelle robot ou bot. Un robot d'indexation dispose de sa propre signature (comme chaque navigateur web). Googlebot est le user agent (signature) du crawler de Google
- **L'indexation** des ressources récupérées consiste à extraire les mots considérés comme significatifs du corpus à explorer. Les mots extraits sont enregistrés dans une base de données organisée comme un gigantesque dictionnaire inverse ou, plus exactement, comme l'index terminologique d'un ouvrage, qui permet de retrouver rapidement dans quel chapitre de l'ouvrage se situe un terme significatif donné. Les termes non significatifs s'appellent des mots vides. Les termes significatifs sont associés à un poids. Celui-ci reflète à la fois la probabilité d'apparition du mot dans un document et le « pouvoir discriminant de ce mot » dans une langue, conformément au principe de la formule TF-IDF.

- **La recherche** correspond à la partie requêtes du moteur, qui restitue les résultats. Un algorithme est appliqué pour identifier dans le corpus documentaire (en utilisant l'index), les documents qui correspondent le mieux aux mots contenus dans la requête, afin de présenter les résultats des recherches par ordre de pertinence supposée. Les algorithmes de recherche font l'objet de très nombreuses investigations scientifiques. Les moteurs de recherche les plus simples se contentent de requêtes booléennes pour comparer les mots d'une requête avec ceux des documents. Mais cette méthode atteint vite ses limites sur des corpus volumineux. Les moteurs plus évolués sont basés sur le paradigme du modèle vectoriel : ils utilisent la formule TF-IDF pour mettre en relation le poids des mots dans une requête avec ceux contenus dans les documents. Cette formule est utilisée pour construire des vecteurs de mots, comparés dans un espace vectoriel, par une similarité cosinus. Pour améliorer encore les performances d'un moteur, il existe de nombreuses techniques, la plus connue étant celle du PageRank de Google qui permet de pondérer une mesure de cosinus en utilisant un indice de notoriété de pages. Les recherches les plus récentes utilisent la méthode dites d'analyse sémantique latente qui tente d'introduire l'idée de cooccurrences dans la recherche de résultats (le terme « voiture » est automatiquement associé à ses mots proches tels que « garage » ou un nom de marque dans le critère de recherche).

De même, un article sur la récolte du blé en France sera jugé pertinent comme candidat à la réponse sur une question concernant la culture des céréales en Europe.

Des modules complémentaires sont souvent utilisés en association avec les trois briques de bases du moteur de recherche. Les plus connus sont les suivants :

1. **Le correcteur orthographique :** il permet de corriger les erreurs introduites dans les mots de la requête, et s'assurer que la pertinence d'un mot sera bien prise en compte sous sa forme canonique.
2. **Le lemmatiseur :** il permet de réduire les mots recherchés à leur lemme et ainsi d'étendre leur portée de recherche.
3. **L'anti-dictionnaire :** utilisé pour supprimer à la fois dans l'index et dans les requêtes tous les mots « vides » (tels que « de », « le », « la ») qui sont non discriminants et perturbent le score de recherche en introduisant du bruit.

1.4. Avantages :

- La possibilité de faire des recherches dans une grande masse d'informations.
- Le fait d'obtenir rapidement des informations précises sur des sujets divers et variés.
- Le fait de pouvoir trouver des documents spécifiques.
- Le fait de repérer des sites récents ou ayant été écartés des annuaires.
- Le fait de pouvoir réaliser des recherches complexes en utilisant la logique booléenne.

1.5. Inconvénients :

- Absence de contrôle des informations (présence d'URL périmées dans les résultats).
- Des interrogations qui semblent complexes car les interfaces changent d'un moteur de recherche à l'autre.
- Des résultats parfois surprenants car l'indexation est automatique.
- Des résultats qui sont classés selon un ordre qui donne la priorité à la popularité des informations

Chapitre 2

Analyse & Conception

2.1. La collection d'information ou Crawling :

Comme la taille entière du Web est trop large et ne cesse pas d'augmenter, même un grand moteur de recherche ne peut couvrir qu'une petite partie du contenu de Web. Selon une étude de Lawrence et Giles (Lawrence and Giles, 2000), aucun moteur de recherche n'indexe plus de 16% du Web. Pour la raison de l'explosion de la taille du Web, les moteurs de recherche deviennent de plus en plus importants comme un moyen primaire de localiser l'information sur Web. Les moteurs de recherche se fondent sur les collections massives de pages Web qui sont acquises à l'aide des crawlers du Web. Le crawler parcourt le Web en largeur (tous les liens de même niveau hiérarchique) en suivant les liens hypertextes contenues dans chaque page visitée jusqu'à une profondeur donnée.

Un crawler commence son exécution par une page « seed » dans notre travail, nous utilisons L'Open Directory Project (OPD) *(http://blog.dmoz.org/)*, aussi appelé DMOZ (Directory Mozilla) qui est un répertoire de sites web créé en 1998. Il est géré, développé et maintenu par des éditeurs bénévoles, chacun de ceux-ci étant responsable de la vérification des sites et de leur classement. DMOZ est un projet gratuit pour l'utilisateur et collaboratif. Les principales catégories de DMOZ international sont : Arts – Jeux - Enfants et adolescents – Références – Achats - Commerce et économie – Santé – Actualités – Régions – Société – Informatique - Vie quotidienne – Loisirs – Sciences – Sports.

Le Web est souvent considère comme des graphes dont les nœuds sont des pages et les arcs sont les relations entre les pages Web, c'est dans ce sens que nous allons stocker dans un dictionnaire le lien d'une page donnée come clef et les liens qu'elle contienne comme valeurs pour les visiter après.

2.2. L'indexation des données collectées :

A partir des données collectées par le robot explorateur (crawler), le module indexeur construit un index général de recherche des données.

Le contenu de chaque page web parcourue déjà par le crawler est analysé pour l'indexer. Différents champs sont alors pris en compte lors de l'indexation en fonction de l'importance relative de ce champ dans le document (utilisation dans le titre, mise en évidence, fréquence d'utilisation, ancres des liens menant vers la page, popularité …) :

- **Le titre des pages web :** 1er critère de pertinence pour la plupart des moteurs, il est proposé par le concepteur du site et situé entre balises <TITLE> et </TITLE>.

- **Les métadonnées :** les balises des métadonnées donnent des informations sur une page web (données sur les données) et ne sont pas visibles par l'utilisateur, mais sont dans le code source de la page. Il existe deux grands types de métadonnées :

  - *Balises META "Description" : <META NAME="description" CONTENT=".....">*

  Avec une longueur souvent limitée : 150-200 caractères, elles permettent de décrire le contenu d'une page sous forme de résumé.

  - *Balises META "Keyword" : <META NAME="keywords" CONTENT=".....">*

  Elles Permettent de caractériser le document par un ou plusieurs mots-clés.

- **Le corps du texte :**

  Aujourd'hui, la plupart des grands moteurs de recherche indexent le texte des pages web, de manière limitée (jusqu'à une certaine taille du texte).

- **Les URL :**

  URL considérée comme un champ de recherche interrogeable. Presque tous les moteurs aujourd'hui indexent l'URL des pages web

Dans notre travail, nous avons choisis d'indexer l'URL d'une page donnée avec son contenue de la balise *META "Keyword"* dans un dictionnaire et utiliser des fichiers textes comme support de stockage.

Champs	Moteurs indexant le champ	Moteurs n'indexant pas le champ
Titre des pages web	Tous	
Balises <META Description>	Alta Vista, HotBot, InfoSeek, Voilà	Lycos Northern Light, Google, AlltheWeb
Balises <META Keywords>	Alta Vista, HotBot, InfoSeek, Voilà	Lycos Northern Light, Google, AlltheWeb
Corps du texte	Tous (avec des variantes)	
URL	Presque tous	Lycos, AllTheWeb

Tableau récapitulatif des champs indexés sur quelques-uns des principaux moteurs

2.3. Le classement des résultats :

Maintenant qu'on a indexé les pages, il est indispensable de classer ces pages par pertinence selon la requête envoyé par l'utilisateur, c'est à dire mettre en premier les pages qui sont le plus susceptibles d'apporter la bonne réponse au mot demandé.

Il existe des tas de méthodes différentes, en voici celle utilise dans notre travail :

- **Le classement par liens :**

Plus il existe de liens vers une page, plus cette page doit faire référence dans un domaine. C'est ce que fait (entre autres) Google pour classer les pages. Cet algorithme est appelé PageRank :

> Le PageRank est l'algorithme d'analyse des liens concourant au système de classement des pages Web utilisé par le moteur de recherche Google. Il mesure quantitativement la popularité d'une page web.
>
> Le principe de base est d'attribuer à chaque page une valeur (ou score) proportionnelle au nombre de fois que passerait par cette page un utilisateur parcourant le graphe du Web en cliquant aléatoirement, sur un des liens apparaissant sur chaque page. Ainsi, une page a un PageRank d'autant plus important qu'est grande la somme des PageRanks des pages qui pointent vers elle (elle comprise, s'il y a des liens internes).

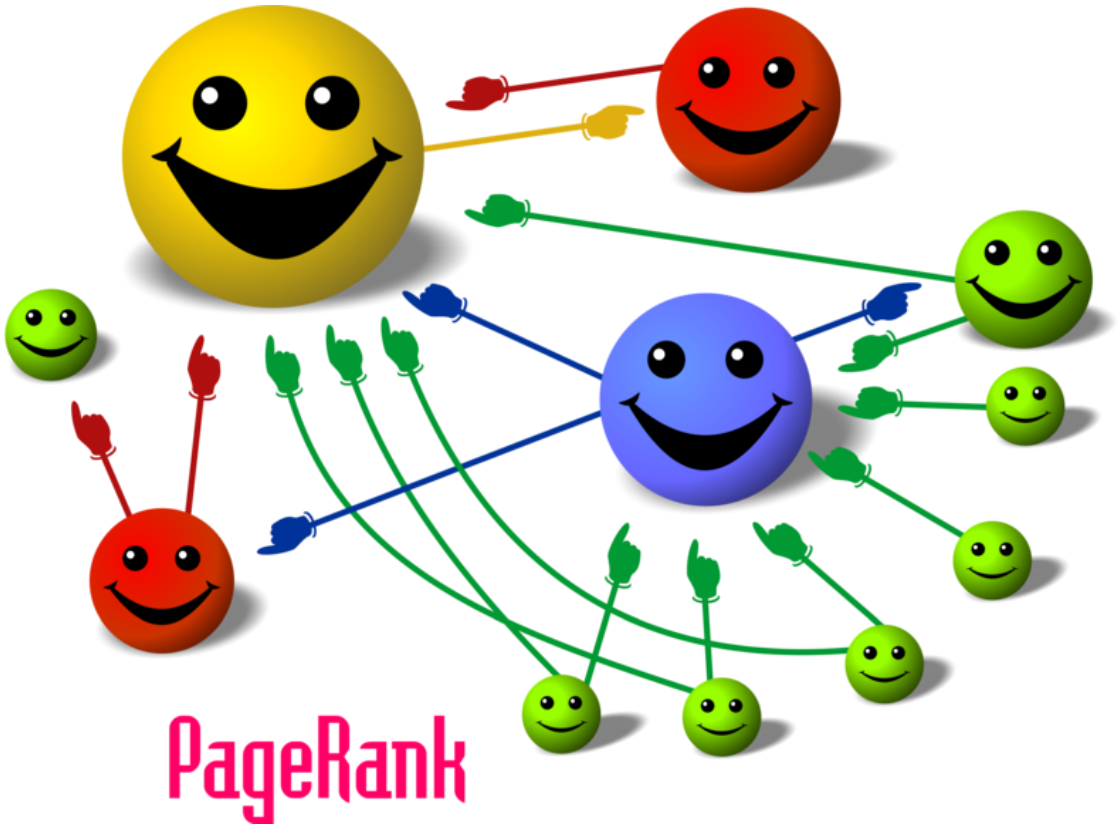


Illustration du PageRank

Il existe de nombreuses autres méthodes pour calculer la pertinence des pages, et ce sont des secrets bien gardés des différents moteurs de recherche. C'est même surtout sur le classement de pertinence des pages que se battent les moteurs de recherche, plus que sur le nombre de pages indexées.

2.4. La recherche :

Lorsque l'utilisateur d'un moteur de recherche remplit le formulaire de recherche, il spécifie les mots qu'il cherche, la requête est envoyée au moteur de recherche qui supprime les stop-words et consulte son index pour chacun des mots puis affine la recherche en triant les pages selon leurs classements.
Il retourne ensuite une liste de résultats contenant liens vers des pages, avec soit le début du texte de la page, soit le texte spécifié par le créateur de la page grâce aux balises spécifiques, appelées méta-tags, ou encore l'extrait de la page qui contient les mots recherchés.

Chapitre 3

Mise en oeuvre

3.1 Présentation des outils utilisés :

3.1.1 Langage Python :

Python est un langage de programmation objet, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ; il est ainsi similaire Perl, Ruby, Scheme, Smalltalk et Tcl.

Le langage Python est placé sous une licence libre proche de la licence BSD3 et fonctionne sur la plupart des plates-formes informatiques, des supercalculateurs aux ordinateurs centraux4, de Windows à Unix avec notamment GNU/Linux en passant par macOS, ou encore Android, iOS, et aussi avec Java ou encore .NET. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser.

3.1.2 Framework Django :

Django est un framework open-source de développement web en Python. Il a pour but de rendre le développement web 2.0 simple et rapide. Pour cette raison, le projet a pour slogan « Le framework web pour les perfectionnistes sous pression ». Développé en 2003 pour le journal local de Lawrence (Kansas), Django a été publié sous licence BSD à partir de juillet 2005.

Depuis juin 2008, la Django Software Foundation s'occupe du développement et de la promotion du framework. En plus de cette promotion régulière, des conférences entre développeurs et utilisateurs de Django sont organisées deux fois par an depuis 2008. Nommées DjangoCon, une se déroule en Europe et l'autre aux États-Unis.

Plusieurs sites grand public sont désormais fondés sur le framework, dont Pinterest, Instagram ou encore Mozilla.

3.1.3 HTML :

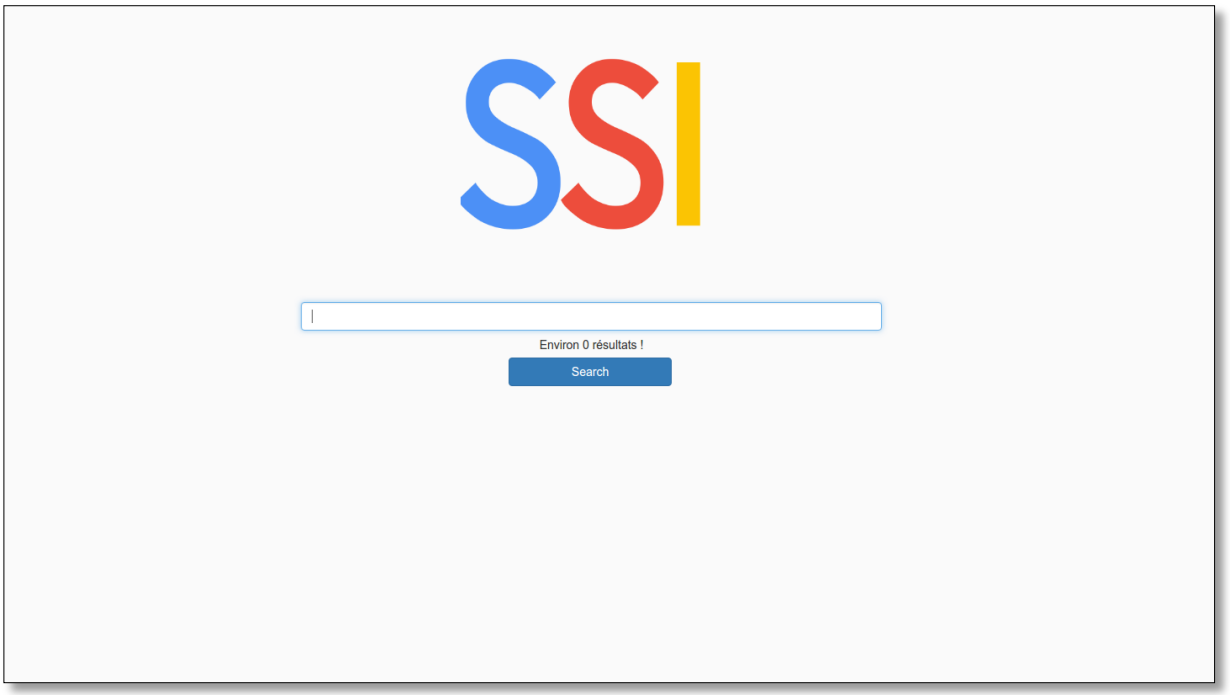
L'Hypertext Markup Language, généralement abrégé HTML, est le format de données conçu pour représenter les pages web. C'est un langage de balisage qui permet d'écrire de l'hypertexte, d'où son nom. HTML permet également de structure sémantiquement et de mettre en forme le contenu des pages, d'inclure des ressources multimédias dont des images et des formulaires de saisie.

3.1.4 CSS :

Le rôle du CSS (Cascading Style Sheets) est de gérer l'apparence de la page web tout en gardant l'information séparée des détails de sa présentation.

Afin de manipuler la présentation, nous avons utilisé des feuilles de style CSS. Voici les raisons : D'une part, il permet d'alléger le code source écrit en HTML, puisque tout ce qui est relatif à la présentation est géré dans un fichier séparé. Ce qui entraîne donc un chargement plus rapide des pages, qui est après manipulé par la feuille de style. Et d'autre part, il permet de nous retrouver plus facilement dans notre code et ainsi facilite les modifications à effectuer, puisqu'au lieu d'avoir à modifier toutes les pages unes à unes, nous avons juste à modifier le fichier CSS.

3.2 Prises d'écran :



Page Index du moteur de recherche

Affichage des résultats d'une recherche

Conclusion

Ce Projet de réalisation d'un moteur de recherche se révèle être une expérience très bénéfique.

Il a été l'occasion idoine pour enrichir nos connaissances et nos compétences en matière du développement Python et développer notre aptitude de travail en équipe.

Webographie

Moteurs de recherche : principes de fonctionnement :

https://www.sites.univ-rennes2.fr/urfist/ressources/moteurs-de-recherche-principes-de-fonctionnement/

Comment ça marche un moteur de recherche ?

http://sebsauvage.net/comprendre/recherche/index.html

http://www.commentcamarche.net/contents/1321-moteur-de-recherche

PageRank

https://fr.wikipedia.org/wiki/PageRank

Les Moteurs de recherche

https://fr.wikipedia.org/wiki/Moteur_de_recherche

Documentation Djano

http://www.tutorialspoint.com/django/

https://code.djangoproject.com/wiki/Tutorials