Crawler thématique

- MBA
- Salah Dahamni
- Daniele Pitrolo

Objet

Le présent cahier des charges a pour objet la définition des exigences liées au programme *crawler thématique*.

Celui-ci est une application web destinée à l'analyse lexicale statistique de pages fournies par les utilisateurs avec une modalité récursive.

Livrables

Les livrables du projet sont :

- Cahier des charges (ce document)
- Code source du logiciel
- Documentation client
- Documentation utilisateur

Portée

Le programme permet de gérer les évènements suivants :

- Accueil de l'utilisateur
- Analyse de la page fournie
- Récursion de l'analyse
- Présentation du résultat
- Sauvegarde du résultat

Des éléments du programme seront réutilisés à partir de solutions déjà existantes et ne feront pas l'objet de développement dans le cadre du projet :

- serveur web
- bibliothèque de crawling
- base de données
- bibliothèque pour l'intéraction avec celle-ci

Les contraintes suivantes sont hors du champ de la portée du programme :

- offline first
- mobile first
- haute disponibilité
- compatbilité avec d'autres navigateurs que la dernière version de Firefox
- montée en charge
- scalabilité
- longevité
- tolerance aux pannes

Contraintes sur la conception de la solution

À prendre en compte

Plusieurs facteurs risquent de reduire la validité des résultats. La présence d'images et de vidéos, que l'application ne saurait pas prendre en compte, ainsi que des textes courts, ou bien des pages de longueur inégale, où l'on toucherait aux limites de l'analyse textuelle avec une simple statistique gaussienne.

Ces éléments seront pris en compte au cours du développement du produit.

Accueil de l'utilisateur

Exigence CRA1.a

Type d'exigence : Exigence Fonctionnelle

Description : Le système doit permettre à l'utilisateur de saisir une URL qui fera l'objet des traitements successifs.

Justification : La recherche s'applique sur une URL donnée par l'utilisateur.

Origine : Julien Roussel.

Critère de satisfaction : Quoique bienvenues, de minimes éléments de design ne sont pas nécessaires. L'affichage mobile n'est pas indispensable.

Contentement du maître d'ouvrage (1 bas - 5 haut) : 2

Mécontentement du maître d'ouvrage (1 bas - 5 haut) : 5

Exigences dépendantes : CRA2.a, CRA2.b, CRA2.c, CRA3.a, CRA6.a

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-05-08

Analyse de la page fournie

Exigence CRA2.a

Type d'exigence : Exigence Fonctionnelle

Description : Le système doit consulter la page présente à l'URL donnée et y identifier les mots les plus fréquents.

Justification : Déterminer les mots les plus fréquents sur la page.

Origine : Julien Roussel.

Critère de satisfaction : Capacité d'identifier les mots les plus fréquents.

Contentement du maître d'ouvrage : 5

Mécontentement du maître d'ouvrage : 5

Exigences dépendantes : CRA1.a

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-05-08

Exigence CRA2.b

Type d'exigence : Exigence Fonctionnelle

Description : Le système doit retourner les résultats si le résultat de l'analyse de l'URL est disponible dans la base de données de l'application, sans analyser la page à nouveau.

Justification : Proposer à l'utilisateur les résultats déjà calculés et fournir une réponse plus rapidement et avec moins de charge pour le système.

Origine : Daniele Pitrolo.

Critère de satisfaction : Une même recherche effectuée deux fois de suite sera traitée de manière bien plus rapide lors de la deuxième fois.

Contentement du maître d'ouvrage : 5

Mécontentement du maître d'ouvrage : 1

Exigences dépendantes : CRA6.a, CRA7.a

Exigences conflictuelles : /

Documents relatifs : /

Historique : * 2016-05-08 * 2016-08-08 Cette exigence est supprimée. Les pages web sont en effet sujettes à une variabilité du contenu difficile à déterminer

(certaines ne changeront jamais, d'autres évoluent dans l'espace de quelques minutes). Plutôt que gérer cette variété, trop complexe, ou bien introduire sciemment une erreur de vetusté des données, l'application effectuera une analyse à chaque fois qu'un lecteur le demande.

Exigence CRA3.a

Type d'exigence : Exigence Fonctionnelle

Description : Le système doit identifier les URL présentes à l'intérieur de la page soumise à l'analyse.

Justification : Étape nécessaire pour itérer l'analyse sur les pages associées.

Origine : Julien Roussel.

Critère de satisfaction : l'analyse dresse una liste des adresses des éléments `a` de la page ; les adresses présentes plusieurs fois ne sont comptées qu'une seule fois.

Contentement du maître d'ouvrage : 3

Mécontentement du maître d'ouvrage : 5

Exigences dépendantes : CRA5.a, CRA5.b, CRA5.c, CRA6.a

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-05-08

Récursion

Exigence CRA4.a

Type d'exigence : Exigence Fonctionnelle

Description : Le système doit itérer la recherche du point CRA2.a sur les URL du point CRA3.a.

Justification : Pour identifier les mots clefs de la même façon sur les pages associées de l'URL fournie par l'utilisateur.

Origine : Intitulé du sujet.

Critère de satisfaction : Le critère de satisfaction de l'exigence CRA2.a est remplie pour tous les éléments de l'exigence CRA3.a.

Contentement du maître d'ouvrage : 5

Mécontentement du maître d'ouvrage : 5

Exigences dépendantes : CRA2.a, CRA3.a

Exigences conflictuelles : /

Documents relatifs : /

Historique : * 2016-05-08 * 2016-08-08 L'algorithme de crawling est entièrement récursif et peut être adapté à d'autres niveaux de récursion.

Exigence CRA5.a

Type d'exigence : Exigence Fonctionnelle

Description : Le système doit itérer la recherche de l'exigence CRA3.a sur les URL obtenues de l'exigence CRA3.a.

Justification : Afin d'avoir une possibilité de restituer la ramification des pages, il est nécessaire que chaque URL obtenue soit analysée à son tour pour retrouver les URL qu'elle contient.

Origine : Julien Roussel.

Critère de satisfaction : L'URL donnée (URL1) est analysée pour obtenir les URL2a, b, c, d… Chacune de ces URL est à son tour analysée pour retrouver les URL3aa, URL3ab, URL3ac; URL3ba, URL3bb, URL3bc; URL3ca… L'analyse est exhaustive.

Contentement du maître d'ouvrage : 5

Mécontentement du maître d'ouvrage : 5

Exigences dépendantes : CRA2.a, CRA3.a, CRA4.a, CRA5.b

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-05-08

Exigence CRA5.b

Type d'exigence : Exigence Fonctionnelle

Description : Le système doit limiter la récursion du point CRA4.a à trois niveaux: URL fournie par l'utilisateur + deux niveaux de recherche.

Justification : Limiter la récursion est nécessaire pour pouvoir conclure l'opération d'analyse, et dans des temps suffisamment restreints.

Origine : Julien Roussel.

Critère de satisfaction : Après avoir analysé l'URL fournie par l'utilisateur, les URL qui y sont contenues, celles contenues dans celles-ci, l'application arrête

la recherche d'URL et l'analyse lexicale des pages, afin de passer à la présentation des résultats.

Contentement du maître d'ouvrage : 4

Mécontentement du maître d'ouvrage : 4

Exigences dépendantes : CRA2.a, CRA3.a, CRA4.a, CRA5.a

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-05-08

Exigence CRA5.c

Type d'exigence : Exigence Fonctionnelle

Description : Le système doit limiter la recherche du point CRA4.a à n éléments (à déterminer via des tests).

Justification : Une limitation horizontale de la récursion est nécessaire afin de limiter le nombre de pages analysées (qui impactent les temps d'attente et les performances de l'application).

Origine : Daniele Pitrolo (suggestion de Georges Grosz).

Critère de satisfaction :

Contentement du maître d'ouvrage : 4

Mécontentement du maître d'ouvrage : 4

Exigences dépendantes : CRA5.d

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-04-29

Présentation du résultat

Exigence CRA2.c

Type d'exigence : Exigence Fonctionnelle

Description : Le système doit produire un message d'erreur si l'URL n'est pas valide

Justification : Informer l'utilisateur que l'analyse ne peut être effectuée.

Origine : MBA.

Critère de satisfaction : Si l'URL fournie par l'utilisateur n'est pas valide la saisie ne peut être validée.

Contentement du maître d'ouvrage : 2

Mécontentement du maître d'ouvrage : 3

Exigences dépendantes : CRA2.a

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-05-08

Exigence CRA6.a

Type d'exigence : Exigence Fonctionnelle

Description : Le système doit permettre l'affichage des résultats d'analyse de l'URL fournie et de ses pages associées.

Justification : Rendre possible à l'utilisateur de prendre connaissance de l'analyse effectuée; suivre l'évolution des mots-clefs à travers plusieurs pages.

Origine : Julien Roussel.

Critère de satisfaction :

Contentement du maître d'ouvrage : 4

Mécontentement du maître d'ouvrage : 5

Exigences dépendantes : CRA2.a, CRA2.b, CRA3.a, CRA4.a, CRA5.a, CRA6.b

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-05-08

Exigence CRA6.b

Type d'exigence : Exigence Fonctionnelle

Description : Les pages des niveaux ultérieurs d'analyse ne sont incluses parmi les résultats que s'ils présentent une continuité lexicale suffisante avec la première page (le niveau sera déterminé en cours de développement par le biais de tests).

Justification : Affichage des résultats uniquement s'ils sont pertinents par rapport à l'analyse de l'URL fournie.

Origine : Julien Roussel.

Critère de satisfaction : La présentation des résultats permet de suivre jusqu'à quel point existe une continuité lexicale entre les pages.

Contentement du maître d'ouvrage : 4

Mécontentement du maître d'ouvrage : 5

Exigences dépendantes : CRA6.a

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-05-08

Sauvegarde du résultat

Exigence CRA7.a

Type d'exigence : Exigence fonctionnelle

Description : Le système doit stocker les résultats dans une base de données associée.

Justification : Pouvoir suivre a posteriori les changements éventuels dans la continuité lexicale entre les pages analysées.

Origine : Julien Roussel.

Critère de satisfaction : Le système peut proposer les résultats en moins de temps et avec moins de charge qu'en les calculant.

Contentement du maître d'ouvrage : 4

Mécontentement du maître d'ouvrage : 2

Exigences dépendantes : CRA2.b, CRA6.a

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-05-08

Contraintes

Exigence CRA8.a

Type d'exigence : Contrainte

Description : Le système doit utiliser Python3 et beautifulsoup4.

Justification : Nécessité pédagogique.

Origine : Julien Roussel.

Critère de satisfaction : /

Contentement du maître d'ouvrage : 1

Mécontentement du maître d'ouvrage : 5

Exigences dépendantes : /

Exigences conflictuelles : /

Documents relatifs : /

Historique : 2016-05-20

Références

Ce cahier des charges contient des éléments dérivés du modèle Volere: http://www.volere.co.uk