SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Ticket #1664. POC moteur de recherche d'images

BM. Bui-Xuan

Banque d'images et moteur de recherche: Dans ce ticket, on appelle banque d'images toute base de données de taille assez conséquente contenant des URLs vers des fichiers d'images. Une banque d'images peut contenir des dizaines de milliers d'URLs, rendant une recherche manuelle impossible. Un moteur de recherche dans une base de données est un webapp permettant à l'utilisateur d'accéder plus rapidement à un URL par recherche de mot-clef.

Objectif du ticket: Fournir un POC d'un moteur de recherche d'images, avec à minima les 4 fonctionnalités suivantes. Une phase de test de montée en charge est à prévoir.

1 Fonctionnalités principales

Le POC sera à livrer sous la forme d'un WebAPI, où les endpoints doivent obligatoirement inclure les fonctionnalités suivantes:

- **Une fonctionnalité explicite de "Recherche":** Recherche d'URL par mot-clef exact. A la requête `.../myImageFromKeyword/<str:name>/`, le webapp retourne la liste de tous les URLs contenant exactement `<str:name>` comme un des mots-clefs dans leurs méta-données, qu'on appellera également table d'index dans ce qui suit.
- **Une fonctionnalité explicite de "Recherche avancée":** Recherche dite "avancée" des URLs par RegEx. A la requête `.../myImageFromRegEx/<str:name>/`, le webapp retourne la liste de tous les URLs contenant au moins un index qui vérifie l'expression régulière `<str:name>` (ici, `<str:name>` est compris comme un RegEx).
- **Une fonctionnalité implicite de classement:** Classement de la liste d'URLs retournée par les deux types de requête précédents. La réponse aux deux types de requête précédents doit être triée par un certain critère de pertinence: on peut penser, pour chaque URL trouvé, à calculer la position de `<str:name>` dans l'URL, divisée par la taille de la table d'index de l'URL par exemple. On est libre de proposer sa façon de produire de tel classement, à condition que le classement n'est pas aléatoire!
- **Une fonctionnalité explicite de "Recherche multiple":** Recherche d'une suite d'URLs par suite de mots-clefs. A la requête `.../myImagesFromManyKeywords/<str:name>/`, le webapp retourne une liste comprenant d'autant URLs que de mots-clefs présents dans `<str:name>`: à chaque mot clef demandé il doit exister un URL distinct contenant le mot clef dans ses méta-données. Ici, les mots-clefs présents dans `<str:name>` doivent être séparés par des séparateurs de type *magic number*: par exemple, on peut supposer par convention que %1664 est le séparateur; alors, la requête `.../myImagesFromManyKeywords/sea%20bream%1664oyster%1664king%20crab%1664caviar/` représentera la liste de 4 mots-clefs `sea_bream`, `oyster`, `king_crab` et `caviar`: le *magic number* %20 représentera alors le caractère d'espacement "_" tandis que le *magic number* %1664 représentera les séparateurs de mots-clefs dans la requête. A cette requête, les 4 URLs retournés doivent satisfaire la condition

suivante: les méta-données du premier URL contient le mot-clef `sea_bream`, les méta-données du deuxième URL contient le mot-clef `oyster`, les méta-données du troisième URL contient le mot-clef `king_crab`, les méta-données du quatrième URL contient le mot-clef `caviar`. De plus, les 4 URLs doivent être disctincts deux à deux.

Afin de tester le POC, la livraison doit comprendre une base de données assez conséquente d'URLs: une phase de récolte de données sera à prévoir. Le POC doit obligatoirement être livré avec un rapport complet des résultats de test de montée en charge, où on veillera à expliciter les conditions ainsi que les modalités du test.

Rapport: Un document (8-12 pages) de retour d'expérience est attendu avant 12h45 le 22/11, jour de livraison du POC.

Présentation par vidéo picth: Une vidéo présentant ce projet est également attendue le jour de la livraison, le 22/11. La présentation est de type *highlight*, et doit durer exactement entre 2 et 5 minutes.

Contraintes:

- A réaliser en individuel.
- Archiver la totalité du rendu en un seul fichier compressé contenant le document de rapport sur la conduite du projet, la vidéo de présentation, et tout ce dont on juge utile à la compréhension du déroulé de réalisation de ce POC sans toute fois dépasser la dizaine de Méga-octet.
- Envoyer ce fichier à `buixuan@lip6.fr`, 3 emails maximum par rendu. L'utilisation des hébergeurs en ligne (drive, git, Slack, etc) est strictement interdite. La nomination de préférence du fichier est gpstl-rendu-final-NOM-PRENOM.piki, où piki peut être un élément de $\{tgz, zip, rar, 7z, etc\}$. Ce format du nom de fichier est important pour un classement automatique des rendus de projet dans le pauvre PC de l'évaluateur des projets de l'UE...
- Deadline: 22 Novembre 2022, 12h45, cachet de serveur de messagerie faisant foi. Pénalité de retard : malus de $2^{h/24}$ points pour h heures de retard.