Recherche d'Information (RI)

Mathias Géry
Mathias.Gery@univ-st-etienne.fr
Laboratoire Hubert Curien, UMR CNRS 5516
Université Jean Monnet Saint-Étienne

Projet RI

- Sujet : Moteur de recherche de documents structurés.
  - Développer un Système de RIS.
  - Tâche INEX : « Relevant In Context » (RIC).
  - Collection : sous-ensemble de la collection INEX 2009.
  - Ressources téléchargeables sur cours en ligne.
  - Étapes intermédiaires, à chaque fois :
    - Envoi de « runs » RIC INEX,
    - Vérification ➔ évaluation ➔ résultats.
    - **Principe 1 :** j'évalue seulement les runs reçus **sur Claroline** en temps et en heure.
    - **Principe 2 :** j'évalue seulement les runs valides.
  - Rendu final : runs + rapport.

- Questions (précises SVP) : Mathias.Gery@univ-st-etienne.fr
  Mathias.Gery@~~pers.~~univ-st-etienne.fr

Projet RI : calendrier

- 14/11 : équipes de 4 étudiants (par mail) :
  - Nom d'équipe : concaténation des 4 prénoms, sans espace ni accent.
  - Ex : JoeJackWilliamAverell
- 14/11 : envoi des premiers runs INEX (étapes 1 et 2).
  - ➔ 18/11 : vérification et évaluation ➔ 1ers résultats (en cours).
- J-2 de chaque cours : envoi des runs suivants (étapes 3-10).
  - ➔ Chaque cours : résultats des runs (06/12, 13/12 et 5/1/2017).
- 11/1 : envoi des runs définitifs.
  - ➔ 17/1 : évaluation finale des runs.
- 25/1 : deadline rendu rapport.

Rappel : exercice avec Michel Beigbeder

- Write a program in the language of your choice (perl, sh, lex, C, java, C++, python, etc.) which index documents contained in a file.
- (…)
- Several collection files of increasing size are available on:

  https://ead-sciences.univ-st-etienne.fr/claroline/course/index.php?cid=RI2016

  - 55k 01-Text_Only-Ascii-Coll-1-10-NoSem.gz
  - 52k 02-Text_Only-Ascii-Coll-11-20-NoSem.gz
  - 103k 03-Text_Only-Ascii-Coll-21-50-NoSem.gz
  - 96k 04-Text_Only-Ascii-Coll-51-100-NoSem.gz
  - 357k 05-Text_Only-Ascii-Coll-101-200-NoSem.gz
  - 559k 06-Text_Only-Ascii-Coll-201-500-NoSem.gz
  - 747k 07-Text_Only-Ascii-Coll-501-1000-NoSem.gz
  - 1.2M 08-Text_Only-Ascii-Coll-1001-2000-NoSem.gz
  - 4.1M 09-Text_Only-Ascii-Coll-2001-5000-NoSem.gz

Projet RI : collection

- Collection M2DSC7Q_2016-17 :
  - 7 requêtes d'INEX 2009.
  - 9804 articles XML (sur 2,6 millions).
  - Version texte (même format que celle utilisée avec Michel Beigbeder),
    - 76,4Mo
    - Text_Only-Ascii-Coll-MWI-NoSem.gz
  - Version XML (pour RIS !).
    - 237Mo
    - XML-Coll-MWI-withSem.tar.gz

7 Requêtes :
2009011 : olive oil health benefit
2009036 : notting hill film actors
2009067 : probabilistic models in information retrieval
2009073 : web link network analysis
2009074 : web ranking scoring algorithm
2009078 : supervised machine learning algorithm
2009085 : operating system +mutual +exclusion

Projet RI : étape 1

- Étape 1/10 (premiers runs « texte ») :
  - Construire 5 runs (maxi).
  - Avec la version texte des documents.
  - Avec la fonction de pondération *ltn* ou *ltc*.
  - Sans structure ➔ un run = 1500 articles complets **par requête**.
  - Objectif : runs valides syntaxiquement, sans recouvrement, etc. Peu importe la performance.
  - Remarque : anti-dico, lemmatisation, etc., sont optionnels.
- Étape 2/10 (premiers runs « XML-articles ») :
  - Idem, 5 runs (maxi), mais avec la collection au format XML.

- Runs à déposer sur cours en ligne au plus tard le 14/11 (pour que je puisse les vérifier/évaluer pour le cours du 18/11).
- Conseil : commencez dès aujourd'hui !

Projet RI : étape 1

- Nom du fichier :
  - NomEquipe_n°etape_n°run_pondération_granularité_paramètres.txt
    - N° étape sur 2 chiffres,
    - N° run unique sur 2 chiffres,
    - Pondération = un code court, ex. ltn, ltc, bm25, bm25f, etc. ou “perso”.
    - Granularité $\in$ {elements, articles}
    - Paramètres : description synthétique des paramètres
  - Exemple : JoeJackWilliamAverell_01_08_bm25_articles_k1.2b0.8.txt
  - Dépôt de plusieurs runs : un seul fichier compressé sans sous-répertoire.
- Vérification et évaluation des runs valides déposés dans les temps.
- Collection au format XML ➔ utiliser un Parser XML :
  - Java : par exemple SAXParserFactory,
  - Perl : XML::LibXML : un exemple est fourni.

Composition des équipes

- Équipes validées au 17/11 :
  - DianaYassineAlhassanHamza : Diana Ramirez, Hamza Baaziz, Yassine Bendari, Alhassane Bah
  - MarlyatouImene : Imène BOUKHALFA, Marlyatou DIALLO
  - IbrahimNoor : Ibrahim BEZRHOUD, Nour el houda ROUATBI
  - PierreJulienDimitriLudovic : Pierre Gourseaud, Julien Da Costa Vaz, Dimitri Escalle, Ludovic Tichit.
  - Total 12 étudiants.
- Runs de :
  - DianaYassineAlhassanHamza
  - IbrahimNoor
  - MarlyatouImene

Projet RI : résultats rendu 16/11

1ère série de runs => cette fois-ci j'ai pris le temps de repérer vos erreurs, d'évaluer aussi les runs envoyés par mail, etc.

- **Remarques :**
  - 12 runs pour 4 équipes (possible : 40 runs)
  - 1 seul fichier contenant tous les runs, sans sous-répertoire !
  - Pas de code source avec vos runs !
  - Scores > 1 : autorisés, scores < 0 : autorisés, tant que ordre décroissant.
  - Pour plus d'infos : cf. les sorties d'inex_eval sur cours en ligne :
    - » resultats_rendu1.tar

Rappel format :

2010001 Q0 364275 12 0.9765 Emse514 /article[1]/bdy[1]/sec[6]/p[1]

Projet RI : résultats rendu 16/11

Rappel format :

2010001 Q0 364275 12 0.9765 Emse514 /article[1]/bdy[1]/sec[6]/p[1]

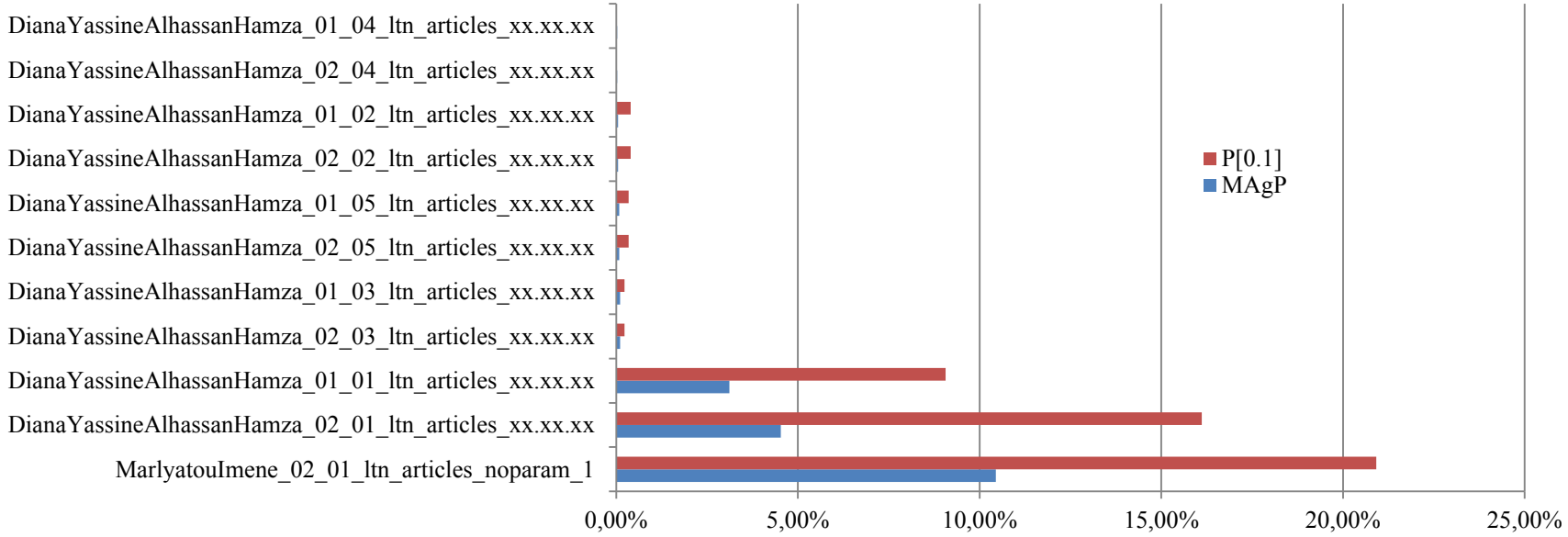
- **Total : 1 run valide…**

- **DianaYassineAlhassanHamza : 10 runs**
  - Ne pas mettre de ligne d'en-tête !
  - 10 500 résultats, Small irrelevant nodes: 10500
  - \article[1] -> /article[1]

- **IbrahimNoor : 1 run**
  - 25 résultats seulement (pour 2 requêtes)
  - article[1] -> /article[1]
  - ERROR *** (IbrahimNoor_01_01_bm25_articles_1.trec++) passage overlaps with previously retrieved passages in topic 2009011 article id: 1006214 in line: 2
  - 2009011 Q0 1006214 1 4992.06 IbrahimNour /article[1]
  - 2009011 Q0 1006214 1 33953.86 IbrahimNour /article[1]

- **MarlyatouImene : 1 run**
  - 10 500 résultats : ok.

Projet RI : résultats rendu 16/11



Equipe	MAgP	P[0.1]
MarlyatouImene_02_01_ltn_articles_noparam_1	10,45%	20,91%
DianaYassineAlhassanHamza_02_01_ltn_articles_xx.xx.xx	4,53%	16,12%
DianaYassineAlhassanHamza_01_01_ltn_articles_xx.xx.xx	3,12%	9,06%
DianaYassineAlhassanHamza_02_03_ltn_articles_xx.xx.xx	0,11%	0,22%
DianaYassineAlhassanHamza_01_03_ltn_articles_xx.xx.xx	0,11%	0,22%
DianaYassineAlhassanHamza_02_05_ltn_articles_xx.xx.xx	0,08%	0,34%
DianaYassineAlhassanHamza_01_05_ltn_articles_xx.xx.xx	0,08%	0,34%
DianaYassineAlhassanHamza_02_02_ltn_articles_xx.xx.xx	0,04%	0,39%
DianaYassineAlhassanHamza_01_02_ltn_articles_xx.xx.xx	0,04%	0,39%
DianaYassineAlhassanHamza_02_04_ltn_articles_xx.xx.xx	0,03%	0,00%
DianaYassineAlhassanHamza_01_04_ltn_articles_xx.xx.xx	0,03%	0,00%

Rendu 4 décembre

- Prochain rendu :
  - 4 décembre 2016.
  - Maxi 20 runs / équipe.
  - Au choix parmi étapes 1 (texte seul) et 2 (XML).
  - Attention : je n'aurais pas le temps de corriger les erreurs des runs…