

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 ➔ 1^{ers} 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 Requetes :

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 $\in \{\text{elements, articles}\}$
 - 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
 - Marlyatoulmene : 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
 - Marlyatoulmene

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 :

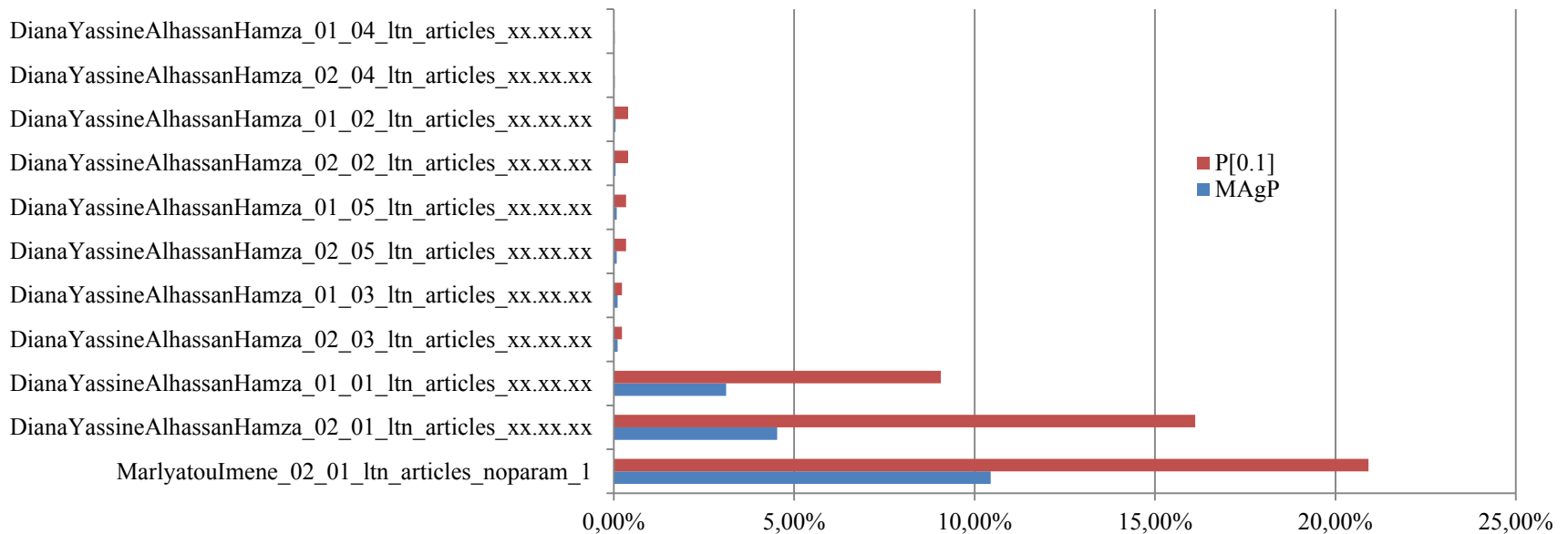
2010001 Q0 364275 12 0.9765 Emse514 /article[1]/bdy[1]/sec[6]/p[1]

- **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]

Marlyatoulmene : 1 run

- 10 500 résultats : ok.

Projet RI : résultats rendu 16/11



Equipe	MAgP	P[0.1]
Marlyatoulmene_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...

Composition des équipes

- Équipes validées au 06/12 :
 - DianaYassineAlhassanHamza : Diana Ramirez, Hamza Baaziz, Yassine Bendari, Alhassane Bah
 - Marlyatoulmene : 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
 - Marlyatoulmene
 - PierreJulienDimitriLudovic

Projet RI : résultats rendu 04/12

- **Remarques :**

- 10 runs pour 4 équipes (possible : 40 runs)
- **Attention : 1 seul fichier contenant tous les runs, sans sous-répertoire !**
- Pour plus d'infos : cf. les sorties d'inex_eval sur cours en ligne :
 - » `resultats_rendu2.tar`

Rappel format :

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

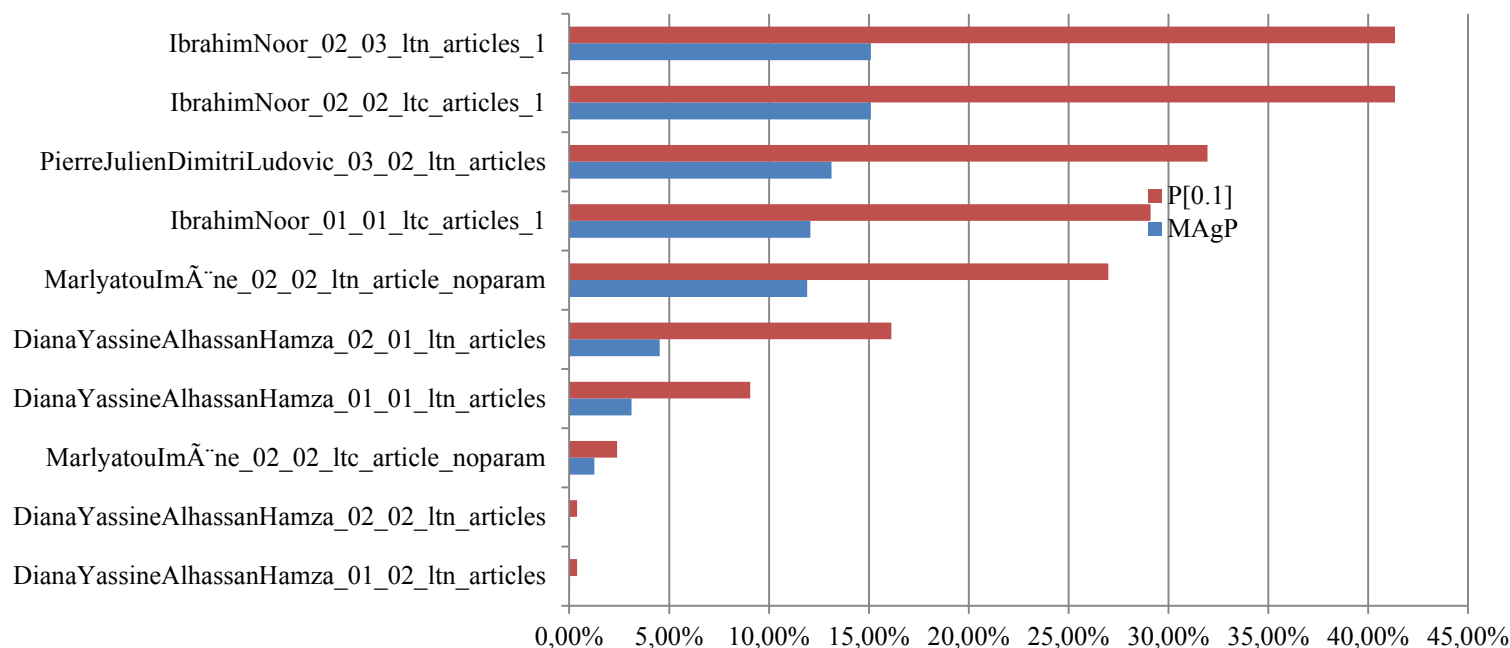
Projet RI : résultats rendu 04/12

Rappel format :

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

- **Total : 8 runs valides**
- **DianaYassineAlhassanHamza : 4 runs valides**
 - 10 500 résultats : ok
- **IbrahimNoor : 3 runs valides**
 - 10 500 résultats : ok
 - Même ranking avec ltn et ltc ?
- **PierreJulienDimitriLudovic : 1 run valide**
 - 10 500 résultats : ok
- **MarlyatouImene : 2 runs invalides**
 - 10 500 résultats , Small irrelevant nodes: 21 et 29
 - Présence de double-espaces dans les runs.

Projet RI : résultats rendu 04/12



Equipe	MAgP	P[0.1]
IbrahimNoor_02_02_ltc_articles_1	15,09%	41,35%
IbrahimNoor_02_03_ltn_articles_1	15,09%	41,35%
PierreJulienDimitriLudovic_03_02_ltn_articles	13,13%	31,96%
IbrahimNoor_01_01_ltc_articles_1	12,07%	29,10%
MarlyatouImÃ~ne_02_02_ltn_article_noparam	11,91%	26,99%
DianaYassineAlhassanHamza_02_01_ltn_articles	4,53%	16,12%
DianaYassineAlhassanHamza_01_01_ltn_articles	3,12%	9,06%
MarlyatouImÃ~ne_02_02_ltc_article_noparam	1,25%	2,39%
DianaYassineAlhassanHamza_02_02_ltn_articles	0,04%	0,39%
DianaYassineAlhassanHamza_01_02_ltn_articles	0,04%	0,39%

Projet RI : étapes suivantes (XML)

- Étape 1/10 (premiers runs « texte »).
- Étape 2/10 (premiers runs « XML-articles ») :
- Étape 3/10 (runs « texte » performants, Q1 ou Q2 ou Q3 du TD) :
 - Construire un run « performant ».
 - Avec la version texte des documents et une pondération de votre choix.
- Étape 4/10 (premiers runs « XML-éléments », Q4-Q5-Q6 du TD) :
 - Avec la version XML des documents.
 - Avec les méthodes de Q4, Q5 et Q6 (essayer plusieurs paramétrages).
 - Choisir les types d'éléments retournés, par ex. : *article*, *sec*, *p*.
- Étape 5/10 (runs « XML-Robertson », Q7-Q8) :
 - Idem étape 4, mais avec les méthodes Q7 et Q8.
- Étape 6/10 (runs « XML-éléments » performants) :
 - Avec la méthode de votre choix.

Rendu 11 décembre

- Prochain rendu :
 - 11 décembre 2016.
 - Maxi 20 runs / équipe, étapes 1 à 6.
 - Cette fois-ci tout le monde devrait pouvoir déposer 20 runs valides avec $\text{MAgP} > 10$!

(rappel) 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 ➔ 1^{ers} 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.

Projet RI : résultats rendu 11 décembre

- **Remarques :**

- 31 runs de 3 équipes (sur : $20 * 4 = 80$ runs).
- 1 seul fichier contenant tous les runs, sans sous-répertoire !
- Pour plus d'infos : cf. les sorties d'inex_eval sur cours en ligne :
 - » resultats_rendu3.tar

Rappel format :

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

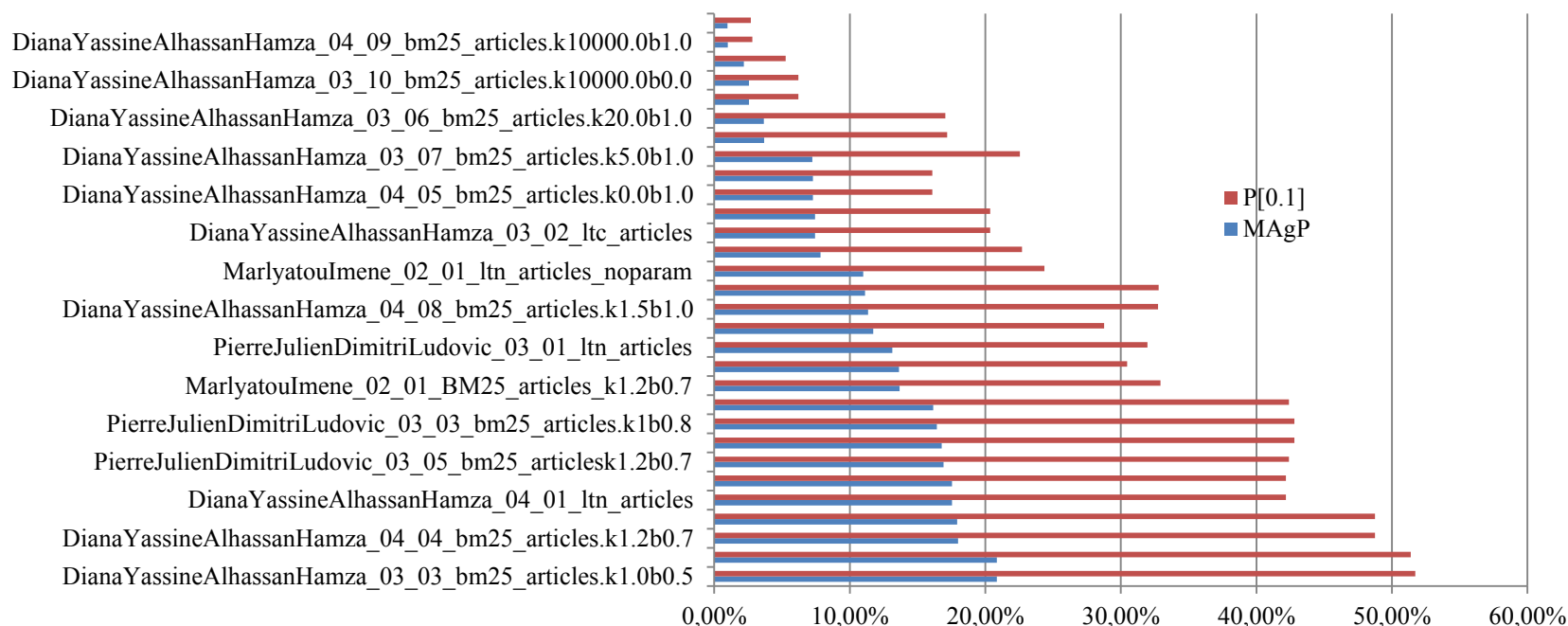
Projet RI : résultats rendu 11 décembre

Rappel format :

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

- **Total : 31 runs, dont 30 valides.**
- **DianaYassineAlhassanHamza : 20 runs valides**
 - 10 500 résultats : ok
- **IbrahimNoor : 0 run**
- **PierreJulienDimitriLudovic : 6 runs dont 5 valides**
 - 10 500 résultats : ok
- **Marlyatoulmene : 5 runs valides**
 - 10 500 résultats : ok

Projet RI : résultats rendu 11 décembre



Equipe	MAgP	P[0.1]
DianaYassineAlhassanHamza_03_03_bm25_articles.k1.0b0.5	20,86%	51,72%
DianaYassineAlhassanHamza_04_03_bm25_articles.k1.0b0.5	20,85%	51,40%
DianaYassineAlhassanHamza_04_04_bm25_articles.k1.2b0.7	17,99%	48,76%
DianaYassineAlhassanHamza_03_04_bm25_articles.k1.2b0.7	17,93%	48,76%
DianaYassineAlhassanHamza_04_01_ltn_articles	17,55%	42,18%
DianaYassineAlhassanHamza_03_01_ltn_articles	17,55%	42,18%
PierreJulienDimitriLudovic_03_05_bm25_articlesk1.2b0.7	16,90%	42,41%
PierreJulienDimitriLudovic_03_06_bm25_articlesk1.2b0.5	16,79%	42,81%
PierreJulienDimitriLudovic_03_03_bm25_articles.k1b0.8	16,43%	42,81%
PierreJulienDimitriLudovic_03_04_bm25_articlesk1.2b0.8	16,15%	42,41%
Marlyatoulmene_02_01_BM25_articles_k1.2b0.7	13,68%	32,92%
Marlyatoulmene_02_01_BM25_articles_k1.0b0.5	13,63%	30,47%
PierreJulienDimitriLudovic_03_01_ltn_articles	13,13%	31,95%
Marlyatoulmene_02_01_BM25_articles_k0.5b0.5	11,73%	28,77%
DianaYassineAlhassanHamza_04_08_bm25_articles.k1.5b1.0	11,36%	32,74%
DianaYassineAlhassanHamza_03_08_bm25_articles.k1.5b1.0	11,13%	32,79%
Marlyatoulmene_02_01_ltn_articles_noparam	10,99%	24,35%
DianaYassineAlhassanHamza_04_07_bm25_articles.k5.0b1.0	7,84%	22,70%
DianaYassineAlhassanHamza_03_02_ltc_articles	7,43%	20,35%
DianaYassineAlhassanHamza_04_02_ltc_articles	7,43%	20,36%
DianaYassineAlhassanHamza_04_05_bm25_articles.k0.0b1.0	7,29%	16,08%
DianaYassineAlhassanHamza_03_05_bm25_articles.k0.0b1.0	7,29%	16,08%
DianaYassineAlhassanHamza_03_07_bm25_articles.k5.0b1.0	7,23%	22,56%
DianaYassineAlhassanHamza_04_06_bm25_articles.k20.0b1.0	3,68%	17,18%
DianaYassineAlhassanHamza_03_06_bm25_articles.k20.0b1.0	3,65%	17,04%
DianaYassineAlhassanHamza_04_10_bm25_articles.k10000.0b0.0	2,56%	6,22%
DianaYassineAlhassanHamza_03_10_bm25_articles.k10000.0b0.0	2,56%	6,20%
Marlyatoulmene_02_01_ltc_articles_noparam	2,19%	5,26%
DianaYassineAlhassanHamza_04_09_bm25_articles.k10000.0b1.0	0,99%	2,81%
DianaYassineAlhassanHamza_03_09_bm25_articles.k10000.0b1.0	0,98%	2,69%

Projet RI : étapes suivantes (XML)

- Étape 1/10 (premiers runs « texte »).
- Étape 2/10 (premiers runs « XML-articles ») :
- Étape 3/10 (runs « texte » performants, Q1 ou Q2 ou Q3 du TD) :
 - Construire un run « performant ».
 - Avec la version texte des documents et une pondération de votre choix.
- Étape 4/10 (premiers runs « XML-éléments », Q4-Q5-Q6 du TD) :
 - Avec la version XML des documents.
 - Avec les méthodes de Q4, Q5 et Q6 (essayer plusieurs paramétrages).
 - Choisir les types d'éléments retournés, par ex. : *article*, *sec*, *p*.
- Étape 5/10 (runs « XML-Robertson », Q7-Q8) :
 - Idem étape 4, mais avec les méthodes Q7 et Q8.
- Étape 6/10 (runs « XML-éléments » performants) :
 - Avec la méthode de votre choix.

Rendu 3 janvier 2017

- Prochain rendu :
 - 3 janvier 2017.
 - Maxi 20 runs / équipe, étapes 1 à 6.
 - Attention : c'est déjà l'avant-dernier dépôt de run !

(rappel) 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 ➔ 1^{ers} 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.