Recherche d'Information (RI)
Projet

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

Practice 1, rendu n°1

Projet RI : practice 1, rendu n°1 du 12/09

- **practice1_report.txt : aidez moi !!!**

- **<u>Rapide débrief de chaque équipe :</u>**
  - Confirmation composition équipes.
  - Langage, librairies, etc.
  - Qu'avez-vous fait ?
    - » Exo 1 : fichier inverse et matrice d'incidence à la main.
    - » Exo 2 : parsing, inverted file ?
    - » Exo 3 : boolean queries ?
  - Difficultés ?

- **<u>Équipes :</u>**
  - 1. InessAliFatihaMohamed
  - 2. GhilasIdriss
  - 3. FlorianJeoffreyJocelynArthur
  - 4. MohamedWilliam
  - 5. AlexandreIanOpheliePierre
  - 6. MahmoudMohammedEmrick
  - ~~7. Mohamed~~

- **<u>Practice n°2 :</u>** date limite 2/10.
- **Attention plagiat !**

Université Jean Monnet Saint-Étienne

Faculté des Sciences et Techniques Saint-Étienne

Practice 2, rendu n°2

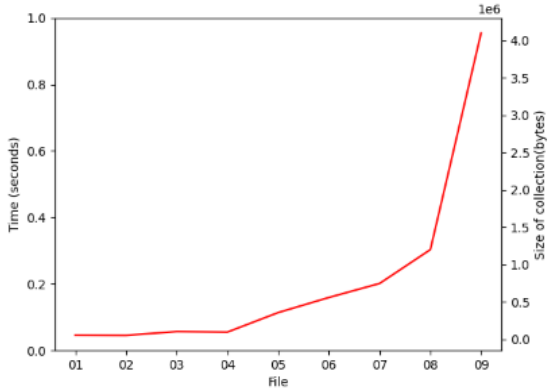
Projet RI : practice 2, rendu n°2 du 2/10

- **Rappel du « Practice 2 ».**

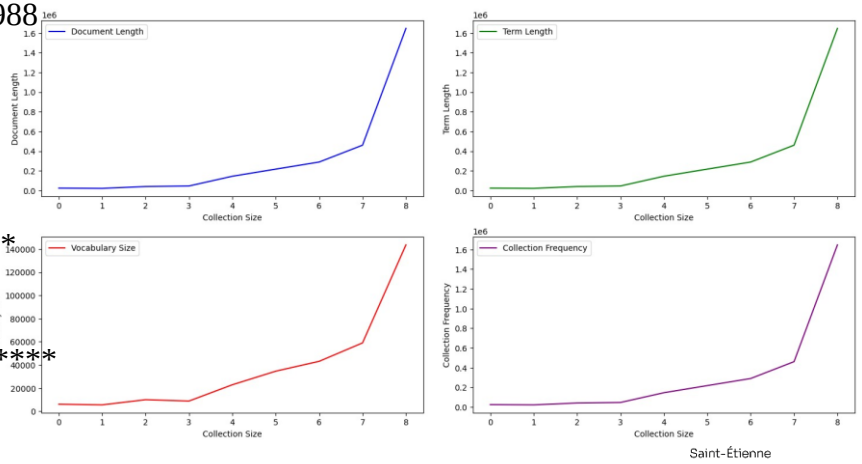
- **Rapide débrief de chaque équipe :**
  - Confirmation composition équipes.
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 fichiers, graphique size / seconds.
    - » Exo 2 : stats : doc length (#terms), term length (#char), voc size, cf total (#occs) ; graphiques.
    - » Exo 3 : stats (9ème fichier), avec stop-words
    - » Exo 4 : stats (9ème fichier), avec stop-words + Porter
  - Difficultés ? Questions ?

Projet RI : practice 2, rendu n°2 du 2/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Confirmation composition équipes.
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 fichiers, graphique size / seconds.
    - » Exo 2 : stats : doc length (#terms), term length (#char), voc size, cf total (#occs) ; graphiques.
    - » Exo 3 : stats (9ème fichier), avec stop-words
    - » Exo 4 : stats (9ème fichier), avec stop-words + Porter
  - Difficultés ? Questions ?



- 1. AliInessFatihaMohamed (Ali BOUGASSAA, Iness BOUABID, Fatiha KHALYFA, Mohamed BADAGUE)
  - Rendu 7,34Mo (?)
  - Exo1 : Tps index : 0.03 -> 1.37sec.
  - Exo2 : calculs ok ? Qu'est-ce qu'on compte ?
  - Exo3 :
    - » with / without ?
    - » ****with stop words****
    - » Document Length: 1646988, Term Length: 1646988
    - » Vocabulary Size: 143546
    - » Collection Frequency of Terms: 1646988
    - » ***without stop words****
    - » 1309110 – 1309110 – 143490 - 1309110
  - Exo4 :
    - » *****with stemming and with stop words*******
    - » Time efficiency : 22.019947052001953
    - » 1323070 – 1323070 – 108842 – 1323070
    - » *****with stemming and without stop words********
    - » Time efficiency :31.66462469100952
    - » 1646988 – 1646988 – 108860 – 1646988

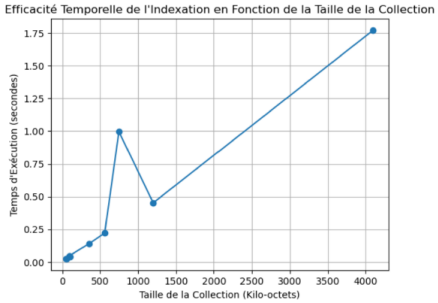


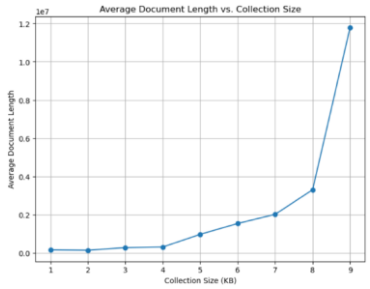
Université Jean Monnet Saint-Étienne

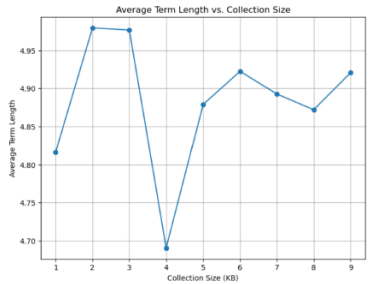
Projet RI : practice 2, rendu n°2 du 2/10

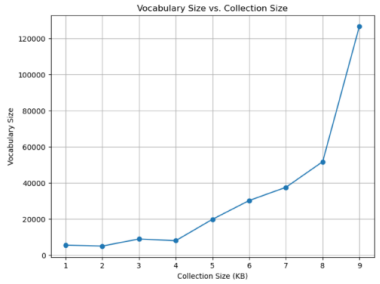
- **<u>Rapide débrief de chaque équipe :</u>**
  - Confirmation composition équipes.
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 fichiers, graphique size / seconds.
    - » Exo 2 : stats : doc length (#terms), term length (#char), voc size, cf total (#occs) ; graphiques.
    - » Exo 3 : stats (9ème fichier), avec stop-words
    - » Exo 4 : stats (9ème fichier), avec stop-words + Porter
  - Difficultés ? Questions ?

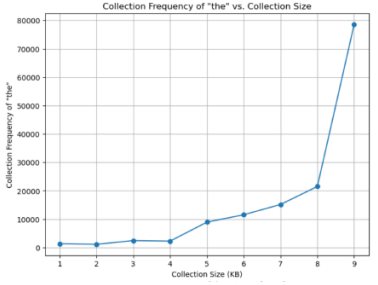
- 1 bis. InessAliMohamedFatiha (???)
  - Rendu 2,56Mo (?)
  - Exo1 : Tps index : … -> 1.75sec. 4Mo ?
  - Exo2 : calculs ok ? Avg doc length ? Cf total ? Unités ?
  - Exo3 :
    - » (graphiques non demandés)
    - » Doc Length diminue, avg term length augmente, voc size diminue.
  - Exo4 :
    - » Doc length (?), term length (?), voc size diminue, cf total (?)





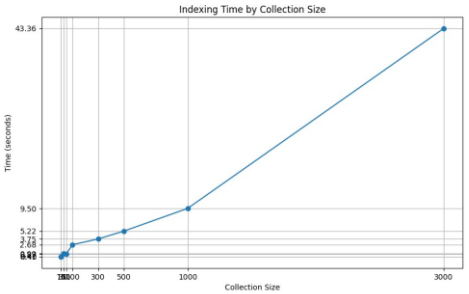




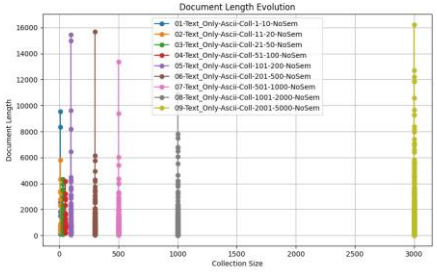


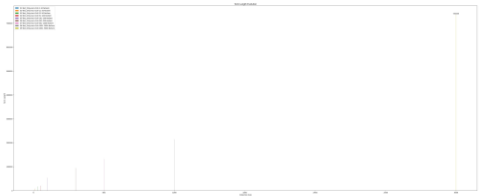
Projet RI : practice 2, rendu n°2 du 2/10

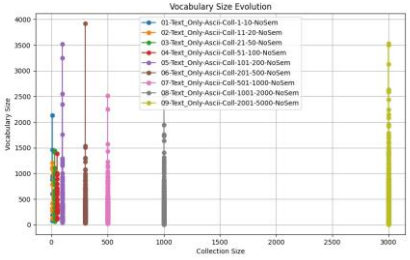
- **<u>Rapide débrief de chaque équipe :</u>**
  - Confirmation composition équipes.
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 fichiers, graphique size / seconds.
    - » Exo 2 : stats : doc length (#terms), term length (#char), voc size, cf total (#occs) ; graphiques.
    - » Exo 3 : stats (9ème fichier), avec stop-words
    - » Exo 4 : stats (9ème fichier), avec stop-words + Porter
  - Difficultés ? Questions ?

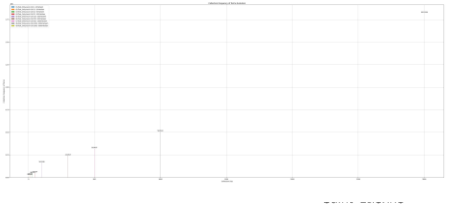


- 2. IdrissGhilas (Idriss BENGUEZZOU, Ghilas MEZIANE)
  - Rendu 9,36Mo (?)
  - Exo1 : Tps index : … -> 43sec
  - Exo2 : calculs à revoir.
  - Exo3 :
    - » 3 antidico.
    - » 39 sec ?
    - » Calculs à revoir.
  - Exo4 :
    - » 2 stemmers.
    - » 74 sec
    - » Calculs à revoir.



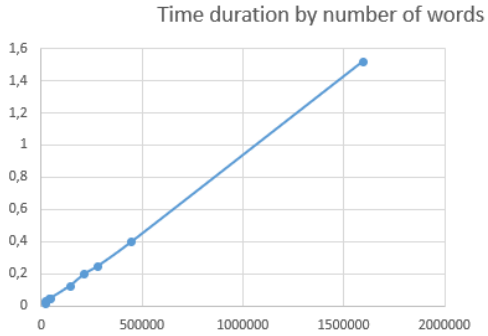






Projet RI : practice 2, rendu n°2 du 2/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Confirmation composition équipes.
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 fichiers, graphique size / seconds.
    - » Exo 2 : stats : doc length (#terms), term length (#char), voc size, cf total (#occs) ; graphiques.
    - » Exo 3 : stats (9ème fichier), avec stop-words
    - » Exo 4 : stats (9ème fichier), avec stop-words + Porter
  - Difficultés ? Questions ?

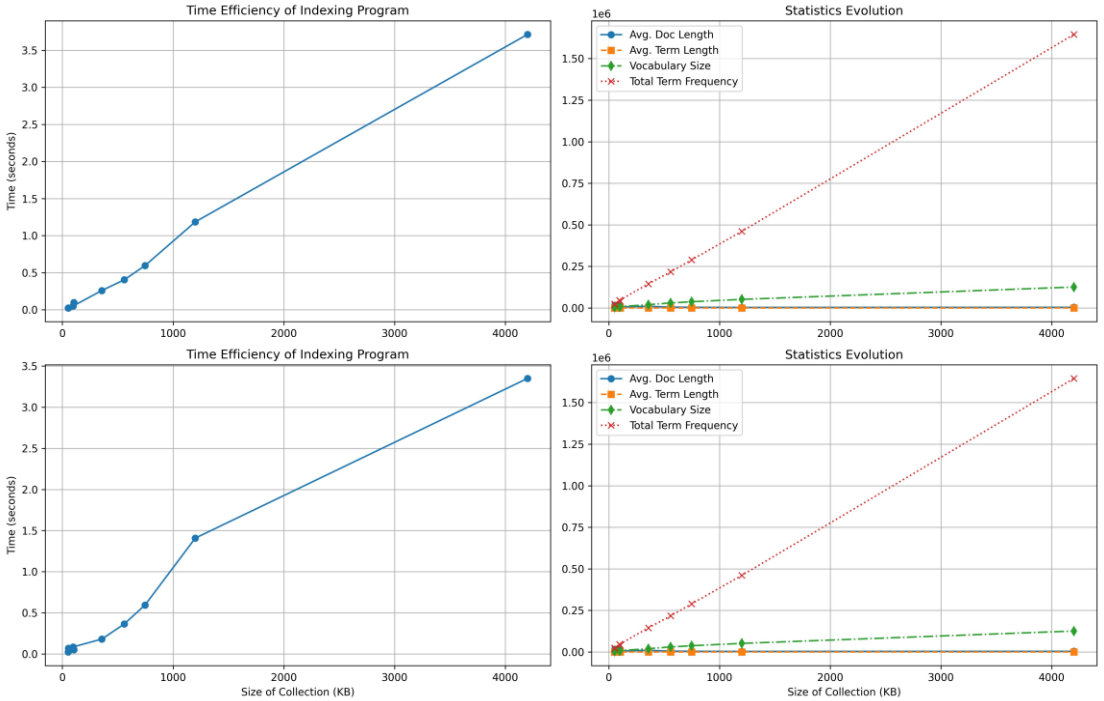


- 3. FlorianJeoffreyJocelynArthur (Florian DOFFEMONT, Jeoffrey PEREIRA, Jocelyn HAUF, Arthur MICOL)
  - Rendu 7,26Mo (?)
  - Exo1 : Tps index : … -> 1,5sec
  - Exo2 : doc length 3878 char - 531 terms, term length 8,74 char, voc size 157 000 ?, cf total (???)
  - Exo3 :
    - » Antidico ?
    - » 585 sec
    - » 3878 (?) / 383 – 8,75 – 157 000 – (???)
  - Exo4 :
    - » Stemmer ?
    - » 651 sec
    - » 3878 (?) / 383 (?) – 8,45 – 140 000 – (???)

Projet RI : practice 2, rendu n°2 du 2/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Confirmation composition équipes.
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 fichiers, graphique size / seconds.
    - » Exo 2 : stats : doc length (#terms), term length (#char), voc size, cf total (#occs) ; graphiques.
    - » Exo 3 : stats (9ème fichier), avec stop-words
    - » Exo 4 : stats (9ème fichier), avec stop-words + Porter
  - Difficultés ? Questions ?

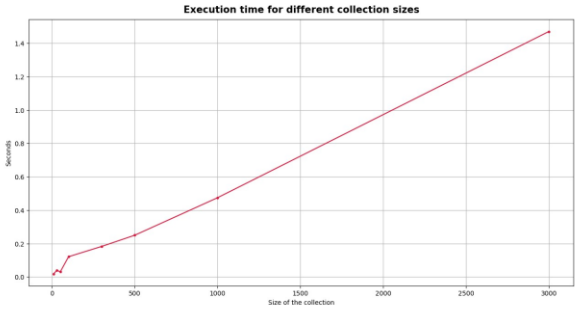
- 4. MohamedWilliam (Mohammed ROUABAH, William MAILLARD)
  - Rendu 12,7Mo (?)
  - Exo1 :
    - » Tps index : 0,03 -> 3,71sec.
    - » Graphs basic / poo ?
  - Exo2 : ?
  - Exo3 : ?
  - Exo4 : ?

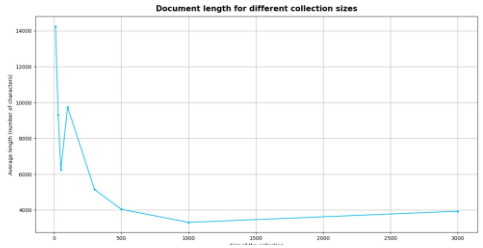


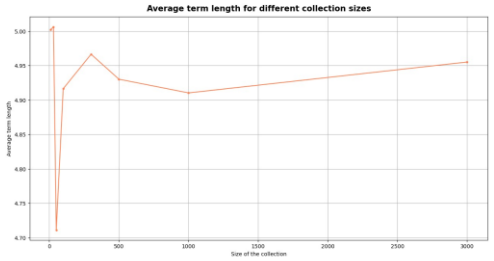
Projet RI : practice 2, rendu n°2 du 2/10

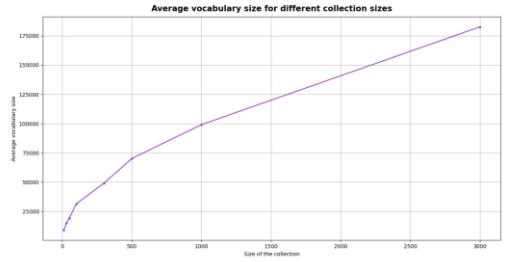
- **<u>Rapide débrief de chaque équipe :</u>**
  - Confirmation composition équipes.
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 fichiers, graphique size / seconds.
    - » Exo 2 : stats : doc length (#terms), term length (#char), voc size, cf total (#occs) ; graphiques.
    - » Exo 3 : stats (9ème fichier), avec stop-words
    - » Exo 4 : stats (9ème fichier), avec stop-words + Porter
  - Difficultés ? Questions ?

- 5. AlexandreIanOpheliePierre (Alexandre MARINE, Ian RIVIERE, Ophelie MARECHAL, Pierre BONNEFOY)
  - Rendu 0,3Mo (c'est possible !)
  - Exo1 : Tps index : … -> 1,4sec.
  - Exo2 : doc length 4000 char, term length 4,95 char, voc size 180 000, cf total (???)
  - Exo3 :
    - » Stats (graphs non demandés) ?
    - » Unité coll size ?
    - » Tps index >13sec.
    - » ~3300 – 5,7 – ~180 000 - ?
  - Exo4 :
    - » Stats (graphs non demandés) ?
    - » ~3300 – 5,7 – ~180 000 - ?







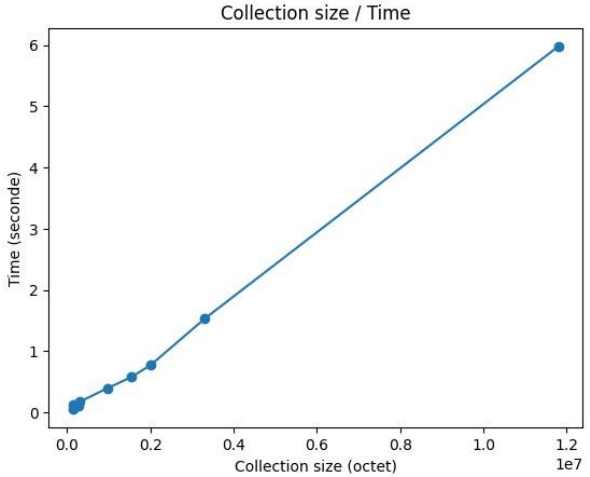


Université Jean Monnet Saint-Étienne

Projet RI : practice 2, rendu n°2 du 2/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Confirmation composition équipes.
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 fichiers, graphique size / seconds.
    - » Exo 2 : stats : doc length (#terms), term length (#char), voc size, cf total (#occs) ; graphiques.
    - » Exo 3 : stats (9ème fichier), avec stop-words
    - » Exo 4 : stats (9ème fichier), avec stop-words + Porter
  - Difficultés ? Questions ?

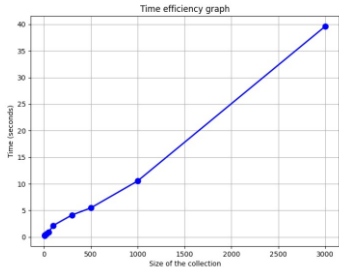
- 6. MahmoudMohammedEmrick (Mahmoud NASHAR, Mohammed BOUTOUIZERA, Emrick PONCET)
  - Rendu 20,7Mo (?)
  - Exo1 : Tps index : … -> 6sec.
  - Exo2 :
    - » Unités ?
    - » doc length 549, term length 7,2, voc size 126 000, cf total 1,65 millions
  - Exo3 :
    - » 369 - 5,6 - 125 700 - 1,1 millions
  - Exo4 :
    - » 549 - ~0 - 108 000 - 1,65 millions (???)

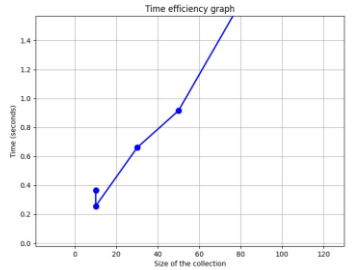


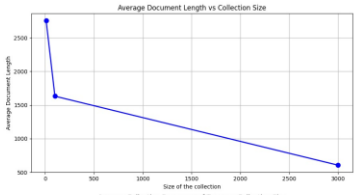
Projet RI : practice 2, rendu n°2 du 2/10

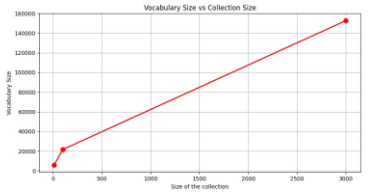
- **<u>Rapide débrief de chaque équipe :</u>**
  - Confirmation composition équipes.
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 fichiers, graphique size / seconds.
    - » Exo 2 : stats : doc length (#terms), term length (#char), voc size, cf total (#occs) ; graphiques.
    - » Exo 3 : stats (9ème fichier), avec stop-words
    - » Exo 4 : stats (9ème fichier), avec stop-words + Porter
  - Difficultés ? Questions ?

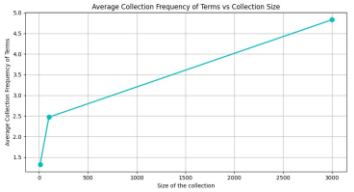
- 7. SaadZakariaBadreddineIgnacio (Saad RIFFI TEMSAMANI, Zakaria ANKIRA, Badreddine ENNOUAJI, Ignacio PEREZ)
  - Rendu 17,5Mo (?)
  - practice2_report.txt : explications code (???)
  - Exo1 : Tps index : … -> 40sec ? Size coll ?
  - Exo2 :
    - » Unités ?
    - » doc length ~550 ?, term length ??, voc size ~155 000, cf total ???
  - Exo3 :
    - » ~450 ? - ?? - ~155 000 - ??
  - Exo4 :
    - » ~450 ? - ?? - ~145 000 - ??











Projet RI : practice 3

- **Practice n°3 :**
  - Date limite 16/10 puis 23/10.
  - (On va enchaîner 3 séances)
  - Même noms d'équipe !
  - practice3_report.txt synthétique et complet !
  - Taille du rendu raisonnable ➔ faites le ménage !

Practice 3, rendu n°3

Projet RI : practice 3, rendu n°3 du 16/10

- **Rappel du « Practice 3 ».**

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 804 documents, 76Mo.
    - » Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size, cf total (#occs).
    - » Exo 3 : stats (stopwords + stemmer).
    - » Exo 4 : indexation ltn. Tests.
    - » Exo 5 : ranked retrieval ltn.
  - Difficultés ? Questions ?

Projet RI : practice 3, rendu n°3 du 16/10

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 804 documents, 76Mo.
    - » Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size, cf total (#occs).
    - » Exo 3 : stats (stopwords + stemmer).
    - » Exo 4 : indexation ltn. Tests.
    - » Exo 5 : ranked retrieval ltn.
  - Difficultés ? Questions ?

- 1. InessAliMohamedFatiha (Iness BOUABID, Ali BOUGASSAA, Mohamed BADAGUE, Fatiha KHALYFA)
  - Exo1 : ok. Tps index 12,49s. Tests ?
  - Exo2 : ok. Tps index ? Tests ?
    - » vocabulary size of 277,357
    - » average document length of 1,193.34
    - » average term length of 5.09
    - » collection term frequency of 11,699,504
  - Exo3 : ok. Tps index 19min ? Tests ?
    - » slight decrease in vocabulary size to 277,207
    - » average document length being 1 indicates that each word is treated as a separate "document,"
    - » average term length increasing to 6.19 is sensible after stemming.
    - » cf decreasing to 8,070,303 is consistent with the removal of stop words.
  - Exo4 : ok. Tps index ? Tests ?
  - Exo5 : ?

Projet RI : practice 3, rendu n°3 du 16/10

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 804 documents, 76Mo.
    - » Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size, cf total (#occs).
    - » Exo 3 : stats (stopwords + stemmer).
    - » Exo 4 : indexation ltn. Tests.
    - » Exo 5 : ranked retrieval ltn.
  - Difficultés ? Questions ?

- 2. IdrissGhilas (Idriss BENGUEZZOU, Ghilas MEZIANE)
  - Stats .json : bof
  - Exo1 : ok. Tests ?
  - Exo2 : ok. Tps index 6,62s. Tests ?
    - » "avg_collection_lengths": 1342.99928,
    - » "avg_term_lengths_in_collection": 4.898170,
    - » "collection_vocabulary_sizes": 426019,
    - » "collection_frequency_of_terms": 13166765
  - Exo3 : ok. Tps index 8,65s. Tests ?
    - » "avg_collection_lengths": 954.5105059159526,
    - » "avg_term_lengths_in_collection": 4.879868630,
    - » "collection_vocabulary_sizes": 383327,
    - » "collection_frequency_of_terms": 9358021
  - Exo4 : ok ? Tps index ? Tests ?
  - Exo5 : ok ? Tests ?

Docno	Score
191457	25.0399
70421	24.0086
23724	21.8707
1482394	21.2219
1009996	21.145
503580	20.7415
6422823	20.5585
3476965	19.8425
363695	19.8133
18336216	19.5984

Projet RI : practice 3, rendu n°3 du 16/10

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 804 documents, 76Mo.
    - » Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size, cf total (#occs).
    - » Exo 3 : stats (stopwords + stemmer).
    - » Exo 4 : indexation ltn. Tests.
    - » Exo 5 : ranked retrieval ltn.
  - Difficultés ? Questions ?

- 3. FlorianJeoffreyJocelynArthur (Florian DOFFEMONT, Jeoffrey PEREIRA, Jocelyn HAUF, Arthur MICOL)
  - Exo1 : ok. Tests ?
  - Exo2 : ok. Tps index 24,21s ? Tests ?
    - » Size of the index : 456590
    - » Average number of term per document : 1150.97
    - » Average term size : 9.91
    - » cf ?
  - Exo3 : ok ? Tps index ? Tests ?
    - » Size of the index : 151645
    - » Average number of term per document : 672.4026
    - » Average term size : 7.011599459263412
  - Exo4 : ok ? Tps index ? Tests ?
  - Exo5 : ?

Projet RI : practice 3, rendu n°3 du 16/10

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 804 documents, 76Mo.
    - » Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size, cf total (#occs).
    - » Exo 3 : stats (stopwords + stemmer).
    - » Exo 4 : indexation ltn. Tests.
    - » Exo 5 : ranked retrieval ltn.
  - Difficultés ? Questions ?

- 4. MohamedWilliam (Mohammed ROUABAH, William MAILLARD)
  - Exo1 : ok. Tests ?
  - Exo2 : ok. Tps index 20,88s. Tests ?
    - » Average Document Length: 8138.59 (words)
    - » Average Term Length: 7.54 (characters)
    - » Vocabulary Size: 271639 (unique terms)
    - » Total Collection Frequency: 11674409 (terms)
  - Exo3 : ok. Nltk. Tps index 278s. Tests ?
    - » Average Document Length: 8138.59 (words)
    - » Average Term Length: 9.26 (characters)
    - » Vocabulary Size: 388307 (unique terms)
    - » Total Collection Frequency: 7699851 (terms)
  - Exo4 : ok ? Tps index ? Tests ?
  - Exo5 : ok ? Tests ?

ID: 23724, Score: 26.00147685
ID: 448834, Score: 22.736500
ID: 18336216, Score: 22.33489586
ID: 207747, Score: 20.9216839
ID: 363695, Score: 19.4839200
ID: 719095, Score: 19.4042650
ID: 1803281, Score: 19.31324954
ID: 8967626, Score: 19.19253496
ID: 2086074, Score: 19.18737372
ID: 149289, Score: 18.8040122

Projet RI : practice 3, rendu n°3 du 16/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 804 documents, 76Mo.
    - » Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size, cf total (#occs).
    - » Exo 3 : stats (stopwords + stemmer).
    - » Exo 4 : indexation ltn. Tests.
    - » Exo 5 : ranked retrieval ltn.
  - Difficultés ? Questions ?

- 5. AlexandreIanOpheliePierre (Alexandre MARINE, Ian RIVIERE, Ophelie MARECHAL, Pierre BONNEFOY)
  - Pas de practice3_report.txt
  - Exo1 : ?
  - Exo2 : ?
  - Exo3 : ?
  - Exo4 : ?
  - Exo5 : ?

Projet RI : practice 3, rendu n°3 du 16/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 804 documents, 76Mo.
    - » Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size, cf total (#occs).
    - » Exo 3 : stats (stopwords + stemmer).
    - » Exo 4 : indexation ltn. Tests.
    - » Exo 5 : ranked retrieval ltn.
  - Difficultés ? Questions ?

- 6. MahmoudMohammedEmrick (Mahmoud NASHAR, Mohammed BOUTOUIZERA, Emrick PONCET)
  - Exo1 : ok. Tests ?
  - Exo2 : ok. Tps index 34.75s. Tests ?
    - » Document length : 1195.262953
    - » Longeur de Terme 7.53653129
    - » Taille de vocabulaire 273820
    - » Nombre de mots (Collection Frequency): 11718358
  - Exo3 : ok. Tps index 187s. Tests ?
    - » Document length : 749.37423500612
    - » Longeur de Terme 4.256079810008597e-05
    - » Taille de vocabulaire 234958
    - » Nombre de mots (Collection Frequency): 7346865
  - Exo4 : ok ? Tps index ? Tests ?
  - Exo5 : ok ? Tests ?

'33120': 3.0250306773525697,
'23724': 2.839933576577888,
'187946': 2.74192443238748,
'15308316': 2.723988425035286,
'475964': 2.7161540167569562,
'7602386': 2.6601064938544146,
'454351': 2.6139657089953072,
'45809': 2.604238961122085,
'6901703': 2.5949056369031145,
'2979338': 2.5862275209542207,

Projet RI : practice 3, rendu n°3 du 16/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ?
    - » Exo 1 : indexation 9 804 documents, 76Mo.
    - » Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size, cf total (#occs).
    - » Exo 3 : stats (stopwords + stemmer).
    - » Exo 4 : indexation ltn. Tests.
    - » Exo 5 : ranked retrieval ltn.
  - Difficultés ? Questions ?

- 7. SaadZakariaBadreddineIgnacio (Saad RIFFI TEMSAMANI, Zakaria ANKIRA, Badreddine ENNOUAJI, Ignacio PEREZ)
  - Pas de practice3_report.txt
  - Exo1 : ok ? Tests ?
  - Exo2 : ok. Tps index 8,04s. Tests ?
    - » avrege length of a document : 1193.3398 Word/Document
    - » size of the vocabulary : 277357 Word
    - » avrege collection frequency : 15.903243
  - Exo3 : ok. Tps index ? Tests ?
    - » avrege length of a document : 483.264585
    - » size of the vocabulary : 238593
    - » avrege collection frequency : 9.92376
  - Exo4 : ? Tps index ? Tests ?
  - Exo5 : ? Tests ?

Projet RI : practice 3, fin

- **Practice n°3, fin :**
  - Date limite 16/10 puis 23/10.
    - » ltn, ltc, BM25.
    - » On va encore enchaîner 2 séances ➔ regarder practice 4 pour le 26/10.
    - » Ensuite : 3 semaines ➔ practice 4 pour le 13/11.
  - practice3_report.txt synthétique et complet !
    - » Pas d'info dispersées.
    - » Stats, temps d'indexation, ranked lists, …
    - » Tests !
  - Taille du rendu raisonnable ➔ faites le ménage !

Practice 3, rendu n°4

Projet RI : practice 3, rendu n°4 du 23/10

- **Rappel du « Practice 3 ».**

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : indexation 9 804 documents, 76Mo. Temps ?
  - Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size (#terms), cf total (#occs).
  - Exo 3 : stats (stopwords + stemmer). Temps ?
  - Exo 4 : indexation ltn. Temps ?
  - Exo 5 : ranked retrieval ltn (web ranking scoring algorithm). Top 10 ? Temps ?
  - Exo 6-7 : idem, ltc.
  - Exo 8-9 : idem, BM25.

Projet RI : practice 3, rendu n°4 du 23/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : indexation 9 804 documents, 76Mo. Temps ?
  - Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size (#terms), cf total (#occs).
  - Exo 3 : stats (stopwords + stemmer). Temps ?
  - Exo 4 : indexation ltn. Temps ?
  - Exo 5 : ranked retrieval ltn (web ranking scoring algorithm). Top 10 ? Temps ?
  - Exo 6-7 : idem, ltc.
  - Exo 8-9 : idem, BM25.
- 1. InessAliMohamedFatiha (Iness BOUABID, Ali BOUGASSAA, Mohamed BADAGUE, Fatiha KHALYFA)
  - Exo1 : Temps 12,49s.
  - Exo2 : Temps ?
    - Average Document Length: 8160
    - Average Term Length: 6.19
    - Collection Vocabulary Size: 277207
    - Total Collection Frequency of Terms (without Stop Words): 8050695
  - Exo3 : Temps ? (was: 19min)
    - Average Document Length: 8160
    - Average Term Length: 5.27
    - Collection Vocabulary Size: 238668
    - Total Collection Frequency of Terms: 8050695
  - Exo4 : Temps ?
  - Exo5 : 9 min. query --> United States patent case law
  - Exo6-9 : ?

Projet RI : practice 3, rendu n°4 du 23/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : indexation 9 804 documents, 76Mo. Temps ?
  - Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size (#terms), cf total (#occs).
  - Exo 3 : stats (stopwords + stemmer). Temps ?
  - Exo 4 : indexation ltn. Temps ?
  - Exo 5 : ranked retrieval ltn (web ranking scoring algorithm). Top 10 ? Temps ?
  - Exo 6-7 : idem, ltc.
  - Exo 8-9 : idem, BM25.
- 2. IdrissGhilas (Idriss BENGUEZZOU, Ghilas MEZIANE)
  - Exo1 : Temps ?
    - » calcul du DF lors de l'indexation
    - » était faux, nous comptions le nombre de documents contenant le terme au lieu de compter la fréquence du terme dans la collection
  - Exo2-3 : Temps ? Stats ?
  - Exo4-5 : Temps 11,7s.
  - Exo6-7 : Temps ?
  - Exo8-9 : Temps 6,4s + 4,9s.
    - » si le df est supérieur au N alors le log est négatif

Docno	Score
18336216	2.05135
6082436	2.0168
719095	1.97729
448834	1.96674
15460921	1.88828
503580	1.87465
13884688	1.86562
1803281	1.83125
5780326	1.82327
47293	1.82103

Docno	Score
5883652	0.0671641
19219678	0.0638628
16186070	0.060381
5124584	0.0602939
752465	0.0594784
18546432	0.0590079
1796590	0.0580079
12162988	0.0575184
2363262	0.0574286
18395000	0.056031

Docno	Score
2758484	1.49486
6415715	1.47706
5780326	1.47563
18336216	1.47235
503580	1.46822
752465	1.46367
13781156	1.45909
1796590	1.45446
15460921	1.45017
9335283	1.44798

Projet RI : practice 3, rendu n°4 du 23/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : indexation 9 804 documents, 76Mo. Temps ?
  - Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size (#terms), cf total (#occs).
  - Exo 3 : stats (stopwords + stemmer). Temps ?
  - Exo 4 : indexation ltn. Temps ?
  - Exo 5 : ranked retrieval ltn (web ranking scoring algorithm). Top 10 ? Temps ?
  - Exo 6-7 : idem, ltc.
  - Exo 8-9 : idem, BM25.
- 3. FlorianJeoffreyJocelynArthur (Florian DOFFEMONT, Jeoffrey PEREIRA, Jocelyn HAUF, Arthur MICOL)
  - Exo1 : Temps 35s.
  - Exo2 :
    - Size of the index : 201381
    - Average number of term per document : 904
    - Average term size : 7.6
  - Exo3 : Temps 151s.
    - Size of the index : 151645
    - Average number of term per document : 672
    - Average term size : 7.0
  - Exo4-5 : ? + 0,002s.
  - Exo6-7 : Temps 5,96s + 0,003s.
  - Exo8-9 : Temps 7,09s + 0,003s.

1 - 23724 , Score : 27.0507
2 - 448834 , Score : 24.5113
3 - 70421 , Score : 23.6490
4 - 1482394, Score : 22.8329
5 - 6082436, Score : 20.8496
6 - 12431 , Score : 20.7694
7 - 207747 , Score : 20.7563
8 - 47293 , Score : 20.6837
9 - 587642 , Score : 20.5546
10 - 187946 , Score : 20.5004

1 - 9981737 , Score : 0.4258
2 - 18096221, Score : 0.3879
3 - 8080270 , Score : 0.3861
4 - 2363262 , Score : 0.3725
5 - 632489 , Score : 0.3377
6 - 8908590 , Score : 0.3281
7 - 10416781, Score : 0.3276
8 - 6422823 , Score : 0.3251
9 - 1851223 , Score : 0.3164
10 - 3503154, Score : 0.3098

1 - 3503154 , Score : 12.2187
2 - 465576 , Score : 11.8733
3 - 23724 , Score : 11.7664
4 - 18096221, Score : 11.7181
5 - 18543218, Score : 11.5304
6 - 1793571 , Score : 11.3432
7 - 6422823 , Score : 11.2309
8 - 1482394 , Score : 10.6060
9 - 3940868 , Score : 10.5938
10 - 752465 , Score : 10.5213

Projet RI : practice 3, rendu n°4 du 23/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : indexation 9 804 documents, 76Mo. Temps ?
  - Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size (#terms), cf total (#occs).
  - Exo 3 : stats (stopwords + stemmer). Temps ?
  - Exo 4 : indexation ltn. Temps ?
  - Exo 5 : ranked retrieval ltn (web ranking scoring algorithm). Top 10 ? Temps ?
  - Exo 6-7 : idem, ltc.
  - Exo 8-9 : idem, BM25.
- 4. MohamedWilliam (Mohammed ROUABAH, William MAILLARD)
  - Exo1 : Temps 123s + 15s ?
  - Exo2 :
    - » Average Document Length: 7301 (words)
    - » Average Term Length: 9.5 (characters)
    - » Vocabulary Size: 428461 (unique terms)
    - » Total Collection Frequency: 11212672 (terms)
  - Exo3 : Temps 405s + 14s ?
    - » nltk_stopwords_stemmer_Lemmizer
    - » Average Document Length: 5209 (words)
    - » Average Term Length: 9.2 (characters)
    - » Vocabulary Size: 387899 (unique terms)
    - » Total Collection Frequency: 7699851 (terms)
  - Exo4-9 : Temps ?

33120, Score: 12.815783519765617
23724, Score: 11.738831802363338
15308316, Score: 11.056182605640544
187946, Score: 10.954102315616797
475964, Score: 10.923951920015716
7602386, Score: 10.612140522126357
454351, Score: 10.359067751823723
45809, Score: 10.307404791505602
3687926, Score: 10.233869785125622
6901703, Score: 10.163819365031584

33120, Score: 0.08302599042431535
23724, Score: 0.0737903271094118
3687926, Score: 0.06886577841176224
187946, Score: 0.06885312002889946
15308316, Score: 0.06821087521267954
1555022, Score: 0.06802736440357869
7602386, Score: 0.06668121154915203
475964, Score: 0.06500047765723194
43651, Score: 0.06404009114289672
4059023, Score: 0.06361985393582249

10013985, Score: 3.956491541626959
10013, Score: 3.956491541626959
1026437, Score: 3.956491541626959
10414589, Score: 3.956491541626959
10416781, Score: 3.956491541626959
10912523, Score: 3.956491541626959
1093623, Score: 3.956491541626959
11255865, Score: 3.956491541626959
11257519, Score: 3.956491541626959
11354525, Score: 3.956491541626959

Projet RI : practice 3, rendu n°4 du 23/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : indexation 9 804 documents, 76Mo. Temps ?
  - Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size (#terms), cf total (#occs).
  - Exo 3 : stats (stopwords + stemmer). Temps ?
  - Exo 4 : indexation ltn. Temps ?
  - Exo 5 : ranked retrieval ltn (web ranking scoring algorithm). Top 10 ? Temps ?
  - Exo 6-7 : idem, ltc.
  - Exo 8-9 : idem, BM25.
- 5. AlexandreIanOpheliePierre (Alexandre MARINE, Ian RIVIERE, Ophelie MARECHAL, Pierre BONNEFOY)
  - Exo1 : Temps 12,7s.
  - Exo2 :
    - » size of the vocabulary 286437
    - » average term length 5.11
    - » average document length 6073
  - Exo3 : Temps 199s.
    - » size of the vocabulary 247470
    - » average term length 5.49
    - » average document length 3881
  - Exo4-9 : Temps ?

23724	73.7818437055454	8080270	0.47352982677938515
448834	62.286284100685194	18096221	0.46748269018742333
70421	59.665926260163495	6422823	0.40136281362556503
1482394	59.24227733481932	1851223	0.39352315285753225
187946	58.725175414415006	4839372	0.3905607206898279
12431	57.36184477602266	10416781	0.37857735740520937
363695	55.4445547417689	11486091	0.37257413024138314
191457	53.834200820428535	18622152	0.36294615546944875
1009996	52.59557030089466	2363262	0.34702542778501494
207747	52.04930800687434	632489	0.3434566673082907

8080270	0.47352982677938515
18096221	0.46748269018742333
6422823	0.40136281362556503
1851223	0.39352315285753225
4839372	0.3905607206898279
10416781	0.37857735740520937
11486091	0.37257413024138314
18622152	0.36294615546944875
2363262	0.34702542778501494
632489	0.3434566673082907

Projet RI : practice 3, rendu n°4 du 23/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : indexation 9 804 documents, 76Mo. Temps ?
  - Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size (#terms), cf total (#occs).
  - Exo 3 : stats (stopwords + stemmer). Temps ?
  - Exo 4 : indexation ltn. Temps ?
  - Exo 5 : ranked retrieval ltn (web ranking scoring algorithm). Top 10 ? Temps ?
  - Exo 6-7 : idem, ltc.
  - Exo 8-9 : idem, BM25.
- 6. MahmoudMohammedEmrick (Mahmoud NASHAR, Mohammed BOUTOUIZERA, Emrick PONCET)
  - Exo1 : Temps 33s.
  - Exo2 :
    - » Longeur de Terme 7.53
    - » Taille de vocabulaire 273820
    - » Collection Frequency: 11718358
    - » AVDL: 1195
  - Exo3 : Temps 17s.
    - » Longeur de Terme 7.06
    - » Taille de vocabulaire 234958
    - » Collection Frequency: 7346865
    - » AVDL: 749
  - Exo4-9 : Temps ?

'33120': 3.0250306773525697,
'23724': 2.839933576577888,
'187946': 2.74192443238748,
'15308316': 2.723988425035286,
'475964': 2.7161540167569562,
'7602386': 2.6601064938544146,
'454351': 2.6139657089953072,
'45809': 2.604238961122085,
'6901703': 2.5949056369031145,
'2979338': 2.5862275209542207,
'1851223': 0.12673467320552947,
'6422823': 0.12352021304954944,
'10433312': 0.12246584997150799,
'15383889': 0.1204904826606353,
'6025658': 0.12006663801212568,
'8080270': 0.11834225392029149,
'782541': 0.11740440436354678,
'1395835': 0.1155997547974226,
'18096221': 0.11417790008975004,
'2045930': 0.11361406966772017,

'23724': 2.126645229969653,
'187946': 2.119923513837759,
'15308316': 2.1170165501985783,
'7602386': 2.1129303430334523,
'6422823': 2.1033494789580836,
'2979338': 2.101094608820903,
'33120': 2.0999784198396427,
'43651': 2.0911051144446304,
'17885012': 2.089503799104267,
'23080': 2.0887196643598798,

Projet RI : practice 3, rendu n°4 du 23/10

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : indexation 9 804 documents, 76Mo. Temps ?
  - Exo 2 : stats (no stopwords, no stemmer) : doc length (#terms), term length (#char), voc size (#terms), cf total (#occs).
  - Exo 3 : stats (stopwords + stemmer). Temps ?
  - Exo 4 : indexation ltn. Temps ?
  - Exo 5 : ranked retrieval ltn (web ranking scoring algorithm). Top 10 ? Temps ?
  - Exo 6-7 : idem, ltc.
  - Exo 8-9 : idem, BM25.
- 7. SaadZakariaBadreddineIgnacio (Saad RIFFI TEMSAMANI, Zakaria ANKIRA, Badreddine ENNOUAJI, Ignacio PEREZ)
  - practice3_report.txt : doc mais pas de résultats.
  - Exo1 : Temps 7,6s.
  - Exo2 :
    - » avrege length of a document: 1193 Word
    - » size of the vocabulary: 277357 Word
    - » avrege collection frequency of a term: 15
  - Exo3 : Temps 10s ?
    - » avrege length of a document: 790 Word
    - » size of the vocabulary: 238593 Word
    - » avrege collection frequency of a term: 14
  - Exo4-5 : Temps +1,8s ; 0,001s.
  - Exo6-7 : Temps +2,2s ; 0,003s.
  - Exo8-9 : Temps +1,4s ; 0,000s.

33120	3.0278401317310686	1851223	0.1255929743305016
23724	2.8421313929664507	10433312	0.1221244561945371
187946	2.7440398345239476	6422823	0.12209189501413689
15308316	2.725853292491381	15383889	0.11995551325786184
475964	2.717567313635172	6025658	0.11885286990199934
7602386	2.6621364044240288	8080270	0.11814575662175647
454351	2.615268788074331	782541	0.11558029045859741
45809	2.6055420402011094	18096221	0.11423486591947421
6901703	2.595787811656785	2045930	0.11317343202305825
2979338	2.587951504358598	1395835	0.11315753741982004

6422823	1.8333567348651036
6843345	1.7765786214086678
23724	1.7733072272601442
6431400	1.7664142920599912
1851223	1.7621542609739214
4979732	1.74486998133898
187946	1.7425445388903105
1592887	1.7396002754729223
43651	1.7318125255532824
15860279	1.7316314026170714

Projet RI : practice 4

- **<u>Practice n°3 :</u>** continuer tests, debug, … → base fiable.

- **<u>Practice n°4 :</u>**
  - Date limite 14/11.
  - Même collection (9804 docs textuels), même indexeur.
  - Génération de « runs » pour 7 requêtes fournies (compétition de RI) :
    - » Pondérations ltn, ltc, BM25.
    - » Avec ou sans stemmer, anti-dico.
    - » Maxi 50 runs.
    - » **Produire des runs syntaxiquement corrects.**
  - practice4_report.txt

Practice 4, rendu n°5

Projet RI : practice 4, rendu n°5 du 14/11

- **Practice n°3 :** continuer tests, debug, … → base fiable.

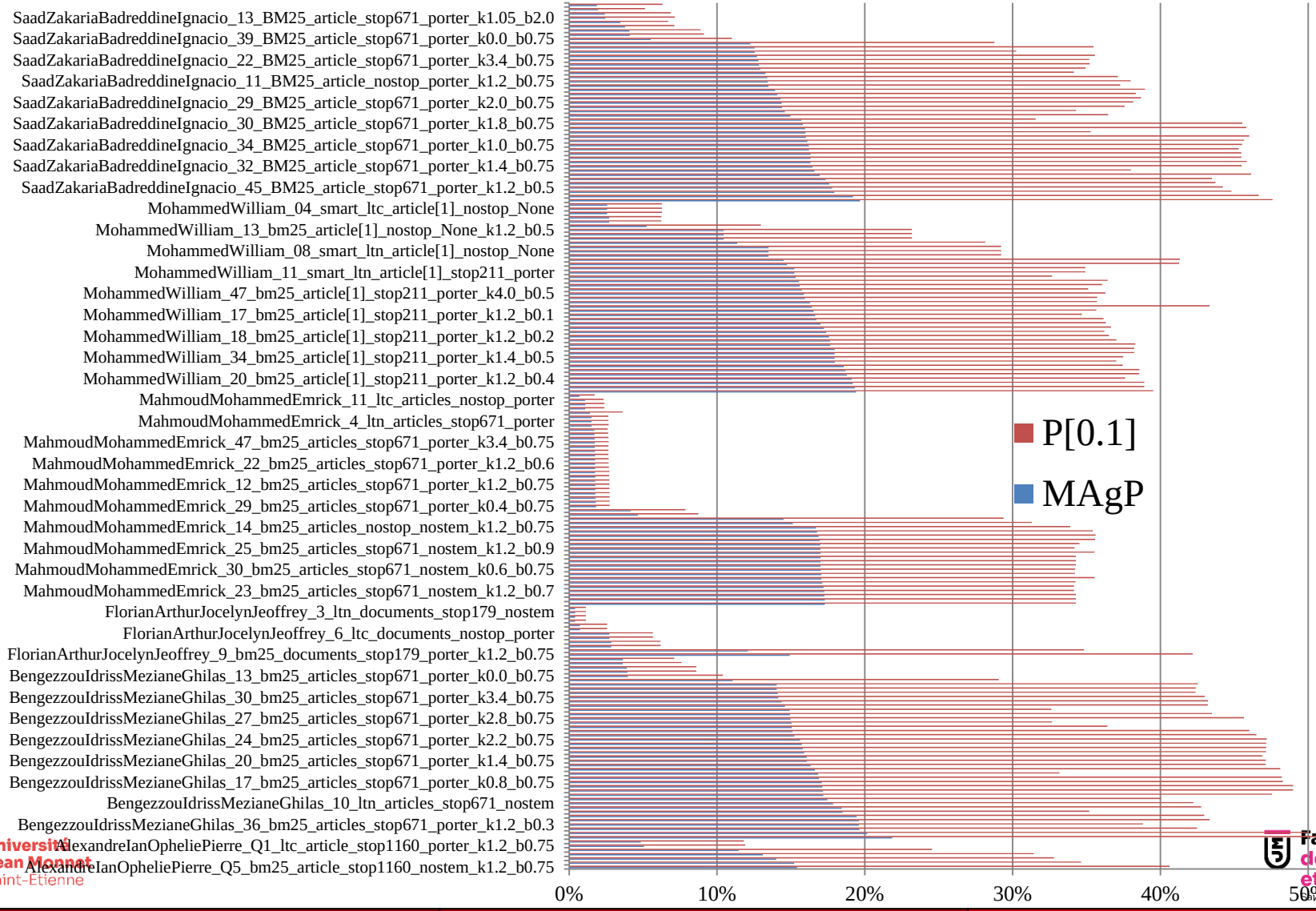
- **Practice n°4 :**
  - Date limite 14/11.
  - Même collection (9804 docs textuels), même indexeur.
  - Génération de « runs » pour 7 requêtes fournies (compétition de RI) :
    - » Pondérations ltn, ltc, BM25.
    - » Avec ou sans stemmer, anti-dico.
    - » Maxi 50 runs.
    - » **Produire des runs syntaxiquement corrects.**
  - practice4_report.txt
  - Exo 4 : 12 runs « baseline » :
    - » requêtes : ignorez les "+"
    - » tokenization: termes sans chiffres ni caractères spéciaux
    - » stop-list: la liste stop-words-english4.txt de ce fichier
    - » Stemming: Porter
    - » paramètres de BM25: k1=1,2 et b=0,75

Projet RI : practice 4, rendu n°5 du 14/11

- **Remarques :**
  - 205 runs pour 7 équipes (sur 7*50 = 350 runs possibles)
  - 204 résultats
  - Pour plus d'infos : cf. les sorties d'*inex_eval* sur cours en ligne :
    - » resultats_runs1.tar

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : 1 run ltn
  - Exo 2 : 1 run ltc
  - Exo 3 : 1 run BM25
  - Exo 4 : 12 runs « baseline » = 3 (weighting) * 2 (stop-list) * 2 (stemmer)
  - Exo 5 : **n** runs « BM25 tuning »

Projet RI : practice 4, rendu n°5 du 14/11



Projet RI : practice 4, rendu n°5 du 14/11

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : 1 run ltn
  - Exo 2 : 1 run ltc
  - Exo 3 : 1 run BM25
  - Exo 4 : 12 runs « baseline » = 3 (weighting) * 2 (stop-list) * 2 (stemmer)
  - Exo 5 : n runs « BM25 tuning »
- **Rappel format :**
  - VictorAlbertJulesIsaac_12_bm25_elements_sec_p_stop344_nostem_k1.2_b0.75.txt
  - <qid> Q0 <article> <rank> <rsv> <run_id> <xml_path>
  - 2010001 Q0 364275 12 0.9765 Mathias514 /article[1]/bdy[1]/sec[6]

- 1. InessAliMohamedFatiha : 0 run, 0 résultat
  - 104 runs ltn ? 15 runs ltc ? 4 runs BM25 ?
  - 1 seul rép !
  - 1 run = 1 fichier (7 requêtes * 1 500 résultats = 10 500 lignes)
  - InessFatihaAliMohammed_Execution_1_2009011_bm25_documents_nostem_nostop.txt
  - 2009085 Q6 10003934 document 1 0.000000 TeamFAMI \document
  - Rang croissant, score décroissant
  - Temps de traitement ?
  - Exos 1-3 : ok ?
  - Exos 4-5 : nok ?

Projet RI : practice 4, rendu n°5 du 14/11

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : 1 run ltn
  - Exo 2 : 1 run ltc
  - Exo 3 : 1 run BM25
  - Exo 4 : 12 runs « baseline » = 3 (weighting) * 2 (stop-list) * 2 (stemmer)
  - Exo 5 : n runs « BM25 tuning »
- **Rappel format :**
  - VictorAlbertJulesIsaac_12_bm25_elements_sec_p_stop344_nostem_k1.2_b0.75.txt
  - <qid> Q0 <article> <rank> <rsv> <run_id> <xml_path>
  - 2010001 Q0 364275 12 0.9765 Mathias514 /article[1]/bdy[1]/sec[6]

- 2. BengezzouIdrissMezianeGhilas : 43 runs, 43 resultats
  - 43 x 10500 lignes, 43 x 0 small irrelevant nodes
  - practice4_report.txt quasi vide
  - 2009011 Q0 22478 1 4.817061504453655 16 /article[1]
  - Temps de traitement ?
  - Exos 1-3 : ok ?
  - Exo 4 (12 runs "baseline") : ok ?
    - » math.log10((N - df + b) / (df + b))
    - » ltn 0.1786, antidico, ~~Porter~~
    - » ltc 0.0395, ~~antidico~~, Porter
    - » BM25 0.2186, antidico, ~~Porter~~
  - Exo 5 : 0.1964, antidico671, Porter, k1=1.2, b=0.4

Projet RI : practice 4, rendu n°5 du 14/11

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : 1 run ltn
  - Exo 2 : 1 run ltc
  - Exo 3 : 1 run BM25
  - Exo 4 : 12 runs « baseline » = 3 (weighting) * 2 (stop-list) * 2 (stemmer)
  - Exo 5 : n runs « BM25 tuning »
- **Rappel format :**
  - VictorAlbertJulesIsaac_12_bm25_elements_sec_p_stop344_nostem_k1.2_b0.75.txt
  - <qid> Q0 <article> <rank> <rsv> <run_id> <xml_path>
  - 2010001 Q0 364275 12 0.9765 Mathias514 /article[1]/bdy[1]/sec[6]

- 3. FlorianArthurJocelynJeoffrey : 12 runs, 12 resultats
  - 3 x 9000, 3 x 10393, 6 x 10500 lignes
  - 12 x 0 small irrelevant nodes
  - Pas de practice4_report.txt
  - Temps de traitement ?
  - Exos 1-3 : ok ?
  - Exo 4 (12 runs "baseline") : nok ?
    - » stop179 ?
    - » ltn 0.0284, antidico, Porter
    - » ltc 0.0284, antidico, Porter
    - » BM25 0.1493, antidico, Porter
  - Exo 5 : nok ?

Projet RI : practice 4, rendu n°5 du 14/11

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : 1 run ltn
  - Exo 2 : 1 run ltc
  - Exo 3 : 1 run BM25
  - Exo 4 : 12 runs « baseline » = 3 (weighting) * 2 (stop-list) * 2 (stemmer)
  - Exo 5 : n runs « BM25 tuning »
- **Rappel format :**
  - VictorAlbertJulesIsaac_12_bm25_elements_sec_p_stop344_nostem_k1.2_b0.75.txt
  - <qid> Q0 <article> <rank> <rsv> <run_id> <xml_path>
  - 2010001 Q0 364275 12 0.9765 Mathias514 /article[1]/bdy[1]/sec[6]

- 4. MohammedWilliam : 45 runs, 45 resultats
  - 45 x 10500 lignes, 45 x 0 small irrelevant nodes
  - practice4_report.txt : bcp de sorties, peu de commentaires. Exos 4&5 ?
  - Temps de traitement ?
  - Exos 1-3 : ok ?
  - Exo 4 (12 runs "baseline") : nok ?
    - stop211 ?
    - ltn 0.1524, antidico, Porter
    - ltc 0.0272, antidico, Porter
    - BM25 ?
  - Exo 5 : 0.1941, antidico211, Porter, k1=0.2, b=0.5

Projet RI : practice 4, rendu n°5 du 14/11

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : 1 run ltn
  - Exo 2 : 1 run ltc
  - Exo 3 : 1 run BM25
  - Exo 4 : 12 runs « baseline » = 3 (weighting) * 2 (stop-list) * 2 (stemmer)
  - Exo 5 : n runs « BM25 tuning »
- **Rappel format :**
  - VictorAlbertJulesIsaac_12_bm25_elements_sec_p_stop344_nostem_k1.2_b0.75.txt
  - <qid> Q0 <article> <rank> <rsv> <run_id> <xml_path>
  - 2010001 Q0 364275 12 0.9765 Mathias514 /article[1]/bdy[1]/sec[6]

- 5. AlexandreIanOpheliePierre : 7 runs, 7 resultats
  - 7 x 10500 lignes, 7 x 0 small irrelevant nodes
  - practice4_report.txt ok
  - "We still appear to have a problem with the …"
  - Index 215sec, lnt-ltc-bm25 ~2sec, querying 1,5sec
  - Exos 1-3 : ok ?
  - Exo 4 (12 runs "baseline") : nok ?
    - » stop1160 ?
    - » index 16sec ?
    - » ltn 0.1311, antidico, ~~Porter~~
    - » ltc 0.0507, antidico, ~~Porter~~
    - » BM25 0.1542, antidico, ~~Porter~~
  - Exo 5 : nok ?

Projet RI : practice 4, rendu n°5 du 14/11

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : 1 run ltn
  - Exo 2 : 1 run ltc
  - Exo 3 : 1 run BM25
  - Exo 4 : 12 runs « baseline » = 3 (weighting) * 2 (stop-list) * 2 (stemmer)
  - Exo 5 : n runs « BM25 tuning »
- **Rappel format :**
  - VictorAlbertJulesIsaac_12_bm25_elements_sec_p_stop344_nostem_k1.2_b0.75.txt
  - <qid> Q0 <article> <rank> <rsv> <run_id> <xml_path>
  - 2010001 Q0 364275 12 0.9765 Mathias514 /article[1]/bdy[1]/sec[6]

- 6. MahmoudMohammedEmrick : 50 runs, 50 resultats
  - 50 x 10500 lignes, 50 x 0 small irrelevant nodes
  - practice4_report.txt light
  - Temps de traitement : 20 min au total
  - Exos 1-3 : ok ?
  - Exo 4 (12 runs "baseline") : ok ?
    - » ltn 0.1513, antidico, ~~Porter~~
    - » ltc 0.0465, antidico, ~~Porter~~
    - » BM25 0.1731, antidico, ~~Porter~~
  - Exo 5 : 0.1731, stop671, ~~Porter~~, k1=1.2, b=0.75

Projet RI : practice 4, rendu n°5 du 14/11

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Tests ? Difficultés ? Questions ?
  - Exo 1 : 1 run ltn
  - Exo 2 : 1 run ltc
  - Exo 3 : 1 run BM25
  - Exo 4 : 12 runs « baseline » = 3 (weighting) * 2 (stop-list) * 2 (stemmer)
  - Exo 5 : n runs « BM25 tuning »
- **Rappel format :**
  - VictorAlbertJulesIsaac_12_bm25_elements_sec_p_stop344_nostem_k1.2_b0.75.txt
  - <qid> Q0 <article> <rank> <rsv> <run_id> <xml_path>
  - 2010001 Q0 364275 12 0.9765 Mathias514 /article[1]/bdy[1]/sec[6]

- 7. SaadZakariaBadreddineIgnacio : 48 runs, 47 resultats
  - 47 x 10500, 1 x 20999 lignes
  - 1 x 1, 47 x 0 small irrelevant nodes
  - Pas de practice4_report.txt
  - Temps de traitement : ?
  - Exos 1-3 : ok ?
  - Exo 4 (12 runs "baseline") : ok ?
    - ltn 0.1497, antidico, ~~Porter~~
    - ltc 0.0412, antidico, Porter
    - BM25 0.1968, antidico, ~~Porter~~
  - Exo 5 : 0.1968, stop671, ~~Porter~~, k1=1.2, b=0.75

Projet RI : practice 4, rendu n°5 du 14/11

- **Meilleurs scores**
  - 3 années précédentes : MAgP : 0,2770 ; P[0.1] : 0,6247
  - Practice 4, rendu n°5 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	Q5_bm25_article_stop1160_nostem_k1.2_b0.75	0,1542	Q5_bm25_article_stop1160_nostem_k1.2_b0.75	0,4062
BengezzouIdrissMezianeGhilas	12_bm25_articles_stop671_nostem	0,2186	12_bm25_articles_stop671_nostem	0,5427
FlorianArthurJocelynJeoffrey	9_bm25_documents_stop179_porter_k1.2_b0.75	0,1493	9_bm25_documents_stop179_porter_k1.2_b0.75	0,4218
MahmoudMohammedEmrick	13_bm25_articles_stop671_nostem_k1.2_b0.75	0,1731	50_bm25_articles_stop671_nostem_k4.0_b0.75	0,3561
MohammedWilliam	28_bm25_article[1]_stop211_porter_k0.2_b0.5	0,1941	23_bm25_article[1]_stop211_porter_k1.2_b0.7	0,4333
SaadZakariaBadreddineIgnacio	10_BM25_article_stop671_nostem_k1.2_b0.75	0,1968	10_BM25_article_stop671_nostem_k1.2_b0.75	0,4759

Prochain rendu (n°6)

- Date limite : 20/11. Priorités :
  - 1) Toutes les équipes : 24 runs « baseline ltn / ltc / bm25 » :
    - Practice 4, exo 4 : 12 runs « baseline » :
      - 3 pondérations * 2 (dico / nodico) * 2 (stemmer / nostemmer)
      - Practice_03_data
      - requêtes : ignorez les "+"
      - tokenization: termes sans chiffres ni caractères spéciaux
      - stop-list: la liste stop-words-english4.txt de ce fichier
      - Stemming: Porter
      - paramètres de BM25: k1=1,2 et b=0,75
    - Practice 5 : 12 runs « baseline » :
      - Idem avec Practice_05_data
      - Vérifier : baseline Practice 3 = baseline Practice 5
  - 2) Practice 5 :
    - 2.1) Exo 1 : indexer XML, stats collection.
    - 2.2) Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)
- Ensuite (5/12, 12/12 et 3/1), Practice 5 :
  - Améliorer BM25 « éléments ».
  - Wilkinson & Robertson (BM25F).

Practice 5, rendu n°6

Projet RI : practice 5, rendu n°6 du 20/11

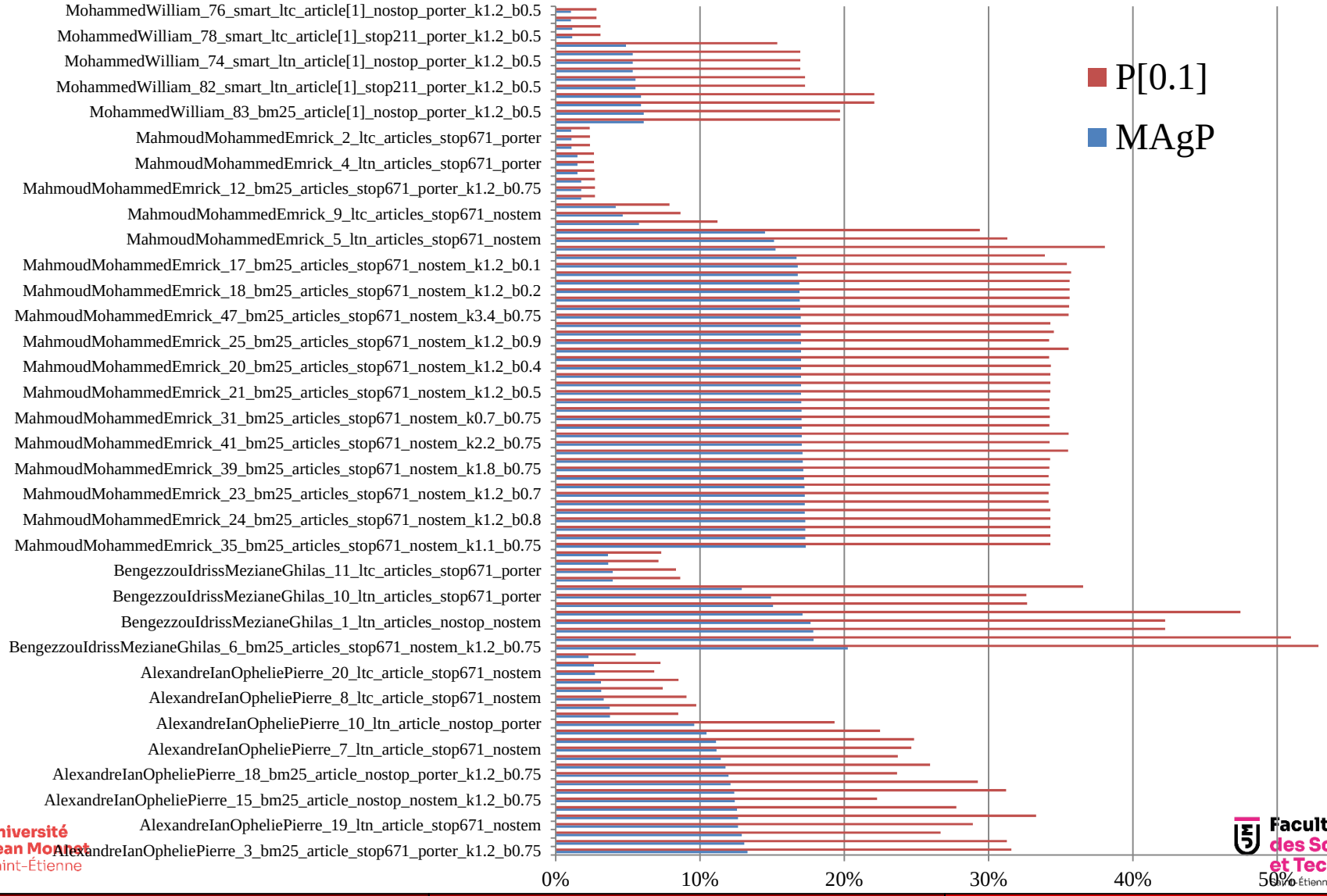
- 1) 24 runs « baseline ltn / ltc / bm25 » :
  - Practice 4, exo 4 : 12 runs « baseline » :
    - 12 = |{ltn,ltc,bm25}| * |{nostop,stop671}| * |{nostem,porter}|
    - Practice_03_data
    - requêtes : ignorez les "+"
    - tokenization: termes sans chiffres ni caractères spéciaux
    - stop-list: la liste stop-words-english4.txt de ce fichier
    - Stemming: Porter
    - paramètres de BM25: k1=1,2 et b=0,75
  - Practice 5 : 12 runs « baseline » :
    - Idem avec Practice_05_data
    - Vérifier : baseline Practice 3 = baseline Practice 5
- 2) Practice 5 :
  - 2.1) Exo 1 : indexer XML, stats collection.
  - 2.2) Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)
- 3) Ensuite (5/12, 12/12 et 3/1), Practice 5 :
  - Améliorer BM25 « éléments ».
  - Wilkinson & Robertson (BM25F).

Projet RI : practice 5, rendu n°6 du 20/11

- **Remarques :**
  - 100 runs pour 7 équipes (sur 7*50 = 350 runs possibles)
  - 0 run « elements »
  - 100 résultats
  - Fichiers déposés mardi ➔ trop tard…
  - Pour plus d'infos : cf. les sorties d'*inex_eval* sur cours en ligne :
    - resultats_runs2.tar

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

Projet RI : practice 5, rendu n°6 du 20/11



Projet RI : practice 5, rendu n°6 du 20/11

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 1. InessAliMohamedFatiha : 0 run, 0 résultat
  - Practice 4, exo 4 : 12 runs baseline ?
  - Practice 5, exo 1 :
    - » 427267 docs ?
    - » Average document length: Practice 05: 25.77, Practice 03: 1,193.33
    - » Vocabulary size: Practice 05: 9,804, Practice 03: 426,019
  - Practice 5, 12 runs baseline ?
  - Practice 5, exos 2-3-4 : ?

Projet RI : practice 5, rendu n°6 du 20/11

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 2. BengezzouIdrissMezianeGhilas : 12 runs, 12 resultats
  - 12 x 10500 lignes, 12 x 0 small irrelevant nodes
  - Practice 4, exo 4 : 12 runs baseline ?
  - Practice 5, exo 1 :
    - » indexing time 6-8 sec, weighting time 2-4sec
    - » avg_lengths 1424, voc size 374837, cf 13967237
  - Practice 5, 12 runs baseline ?
    - » ltn 0.1786, antidico, ~~Porter~~
    - » ltc 0.0396, ~~antidico~~, Porter
    - » BM25 0.2025, antidico, ~~Porter~~
  - Practice 5, exos 2-3-4 : ?

Projet RI : practice 5, rendu n°6 du 20/11

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 3. FlorianArthurJocelynJeoffrey : 0 run, 0 resultat
  - ?

Projet RI : practice 5, rendu n°6 du 20/11

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 4. MohammedWilliam : 14 runs, 14 resultats
  - 3 x 10498, 11 x 10499 lignes, 14 x 0 small irrelevant nodes
  - Practice 4, exo 4 : 12 runs baseline ?
  - Practice 5, exo 1 :
    - » total time 6msec (???)
    - » 9982 doc (???), avg doc length: 216 (words), avg term length: 4.3 (characters), voc size: 39604 (unique terms)
  - Practice 5, 12 runs baseline ?
    - » ltn-ltc-bm25 : 0.01 – 0.06 MAgP
    - » 79_smart_ltn_article\[1\]_nostop_porter_k1.2_b0.5.txt
    - » 2009011 Q0 243 3.5057462095837897 MohammedWilliam79 /article[1]
  - Practice 5, exos 2-3-4 : ?

Projet RI : practice 5, rendu n°6 du 20/11

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 5. AlexandreIanOpheliePierre : 24 runs, 24 resultats
  - 24 x 10500 lignes, 24 x 0 small irrelevant nodes
  - Practice 4, exo 4 : 12 runs baseline ?
    - » No stem, no stop
    - » size voc 199291 words
    - » avg doc length 1091 words/doc
    - » index : 7.02s, ltn : 1.5s, ltc : 1.4s, bm25 : 1.9s
    - » ltn 0,1115, antidico, ~~Porter~~
    - » ltc 0,0377, antidico, Porter
    - » BM25 0,1330, antidico, Porter
  - Practice 5, exo 1 :
    - » size voc 193461 words
    - » avg doc length 588 words/doc
    - » index : 9.4s, ltn : 0.9s, ltc : 0.9s, bm25 : 1.2s
  - Practice 5, 12 runs baseline ?
    - » ltn 0,1308, antidico, Porter
    - » ltc 0,0375, antidico, Porter
    - » BM25 0,1291, antidico, Porter
  - Practice 5, exos 2-3-4 : ?

Projet RI : practice 5, rendu n°6 du 20/11

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 6. MahmoudMohammedEmrick : 50 runs, 50 resultats
  - 50 x 10500 lignes, 50 x 0 small irrelevant nodes
  - Practice 4, exo 4 : 12 runs baseline ?
    - » Temps : 5.0 seconds
    - » doc len : 1113
    - » voca len : 206913
    - » collec len : 10913214
  - Practice 5, exo 1 :
    - » 6.4 seconds
    - » doc len : 1113
    - » voca len : 206690
    - » collec len : 10916103
  - Practice 5, 12 runs baseline ?
  - Practice 5, exos 2-3-4 : ?

Projet RI : practice 5, rendu n°6 du 20/11

- **<u>Rapide débrief de chaque équipe :</u>**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 7. SaadZakariaBadreddineIgnacio : 0 run, 0 resultat
  - ?

Projet RI : practice 5, rendu n°6 du 20/11

- **Meilleurs scores**
  - 3 années précédentes : MAgP : 0,2770 ; P[0.1] : 0,6247
  - Practice 4, rendu n°5 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	Q5_bm25_article_stop1160_nostem_k1.2_b0.75	0,1542	Q5_bm25_article_stop1160_nostem_k1.2_b0.75	0,4062
BengezzouIdrissMezianeGhilas	12_bm25_articles_stop671_nostem	0,2186	12_bm25_articles_stop671_nostem	0,5427
FlorianArthurJocelynJeoffrey	9_bm25_documents_stop179_porter_k1.2_b0.75	0,1493	9_bm25_documents_stop179_porter_k1.2_b0.75	0,4218
MahmoudMohammedEmrick	13_bm25_articles_stop671_nostem_k1.2_b0.75	0,1731	50_bm25_articles_stop671_nostem_k4.0_b0.75	0,3561
MohammedWilliam	28_bm25_article[1]_stop211_porter_k0.2_b0.5	0,1941	23_bm25_article[1]_stop211_porter_k1.2_b0.7	0,4333
SaadZakariaBadreddineIgnacio	10_BM25_article_stop671_nostem_k1.2_b0.75	0,1968	10_BM25_article_stop671_nostem_k1.2_b0.75	0,4759

  - Practice 5, rendu n°6 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,133	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	6_bm25_articles_stop671_nostem_k1.2_b0.75	0,2025	6_bm25_articles_stop671_nostem_k1.2_b0.75	0,5287
MahmoudMohammedEmrick	35_bm25_articles_stop671_nostem_k1.1_b0.75	0,1734	26_bm25_articles_stop671_nostem_k1.2_b1	0,3806
MohammedWilliam	83_bm25_article[1]_nostop_porter_k1.2_b0.5	0,0612	85_bm25_article[1]_stop211_porter_k1.2_b0.5	0,2208

Prochain rendu (n°7)

- Date limite : 5/12. Priorités :
  - 1) 24 runs « baseline ltn / ltc / bm25 » :
    - Practice 4, exo 4 : 12 runs « baseline » :
      - 12 = |{ltn,ltc,bm25}| * |{nostop,stop671}| * |{nostem,porter}|
      - Practice_03_data
      - requêtes : ignorez les "+"
      - tokenization: termes sans chiffres ni caractères spéciaux
      - stop-list: la liste stop-words-english4.txt de ce fichier
      - Stemming: Porter
      - paramètres de BM25: k1=1,2 et b=0,75
    - Practice 5 : 12 runs « baseline » :
      - Idem avec Practice_05_data
      - Vérifier : baseline Practice 3 = baseline Practice 5
  - 2) Practice 5 :
    - 2.1) Exo 1 : indexer XML, stats collection.
    - 2.2) Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)
- Ensuite (12/12 et 3/1), Practice 5 :
  - Améliorer BM25 « éléments ».
  - Wilkinson & Robertson (BM25F).

Practice 5v2, rendu n°7

Projet RI : practice 5v2, rendu n°7 du 5/12

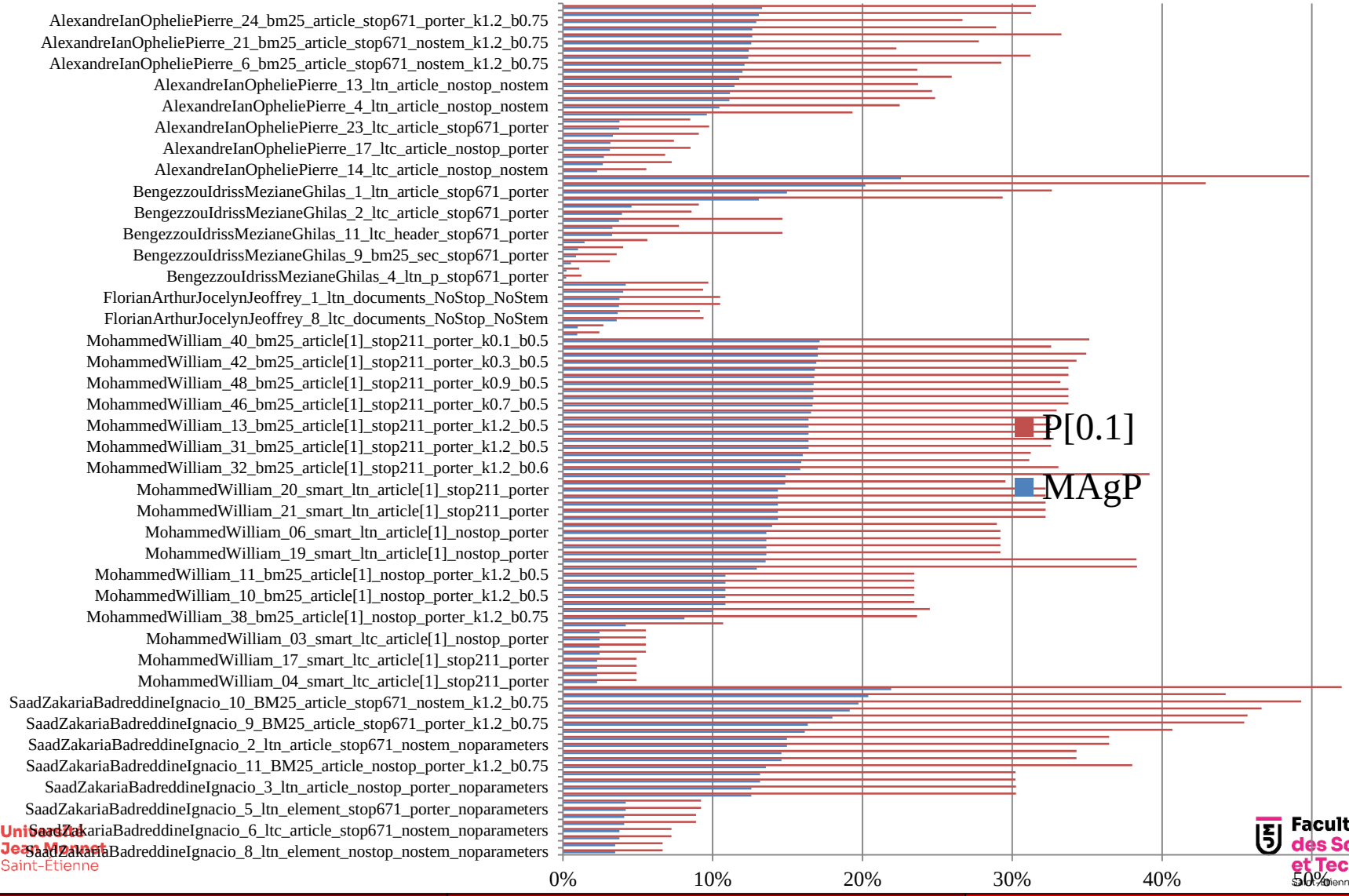
- Date limite : 5/12. Priorités :
  - 1) 24 runs « baseline ltn / ltc / bm25 » :
    - Practice 4, exo 4 : 12 runs « baseline » :
      - 12 = |{ltn,ltc,bm25}| * |{nostop,stop671}| * |{nostem,porter}|
      - Practice_03_data
      - requêtes : ignorez les "+"
      - tokenization: termes sans chiffres ni caractères spéciaux
      - stop-list: la liste stop-words-english4.txt de ce fichier
      - Stemming: Porter
      - paramètres de BM25: k1=1,2 et b=0,75
    - Practice 5 : 12 runs « baseline » :
      - Idem avec Practice_05_data
      - Vérifier : baseline Practice 3 = baseline Practice 5
  - 2) Practice 5 :
    - 2.1) Exo 1 : indexer XML, stats collection.
    - 2.2) Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)
- Ensuite (12/12 et 3/1), Practice 5 :
  - Améliorer BM25 « éléments ».
  - Wilkinson & Robertson (BM25F).

Projet RI : practice 5v2, rendu n°7 du 5/12

- **Remarques :**
  - 190 runs pour 7 équipes (sur 7*50 = 350 runs possibles)
  - 15 runs « elements » (une seule équipe)
  - 190 résultats
  - Pour plus d'infos : cf. les sorties d'*inex_eval* sur cours en ligne :
    - resultats_runs3.tar

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

Projet RI : practice 5v2, rendu n°7 du 5/12



Projet RI : practice 5v2, rendu n°7 du 5/12

- **Débrief :**
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 1. InessAliMohamedFatiha : 70 runs, 70 résultats
  - 70 x 1500 lignes, 70 x 1500 small irrelevant nodes
  - InessAliMohamedFatiha_Run-39658611_smart_ltn_articles_nostop_stemmed.txt
  - 2009073 Q0 8740 1 3.5616 39658611 /article[8741]
  - 2009073 Q0 9087 2 3.3305 39658611 /article[9088]
  - 2009073 Q0 6377 3 3.3021 39658611 /article[6378]
  - Fin Practice 4 ?
  - Rappel format : <qid> Q0 <article> <rank> <rsv> <run_id> <xml_path>
    2010001 Q0 364275 12 0.9765 Mathias514 /article[1]/bdy[1]/sec[6]

Projet RI : practice 5v2, rendu n°7 du 5/12

- **Débrief :**
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 2. BengezzouIdrissMezianeGhilas : 15 runs, 15 résultats
  - 6 x 10500, 3 x 7995, 3 x 4412, 3 x 8823 lignes
  - 15 x 0 small irrelevant nodes
  - Practice 4&5, 2*12 runs baseline : non.
  - Practice 5, exo 1 : debug pre-processing
    - » indexing_time: 9.2
    - » avg_collection_lengths: 641
    - » vocabulary_sizes: 207965
    - » cf_of_terms: 6293619
  - Practice 5, exos 2-3-4 :
    - » 15 runs « éléments » ?
    - » Stockage XPath dans fichier inverse json
    - » Index 9 sec, req 2-3 sec
    - » Calcul du df ?

	MAgP	P[0.1]
15_bm25_bdy_stop671_porter	22,58%	49,84%
3_bm25_article_stop671_porter	20,18%	42,93%
1_ltn_article_stop671_porter	14,95%	32,65%
13_ltn_bdy_stop671_porter	13,08%	29,36%
14_ltc_bdy_stop671_porter	4,58%	9,07%
2_ltc_article_stop671_porter	3,94%	8,59%
12_bm25_header_stop671_porter	3,74%	14,66%
10_ltn_header_stop671_porter	3,31%	7,74%
11_ltc_header_stop671_porter	3,28%	14,65%
6_bm25_p_stop671_porter	1,43%	5,64%
8_ltc_sec_stop671_porter	1,01%	4,02%
9_bm25_sec_stop671_porter	0,87%	3,58%
5_ltc_p_stop671_porter	0,54%	3,13%
7_ltn_sec_stop671_porter	0,23%	1,09%
4_ltn_p_stop671_porter	0,22%	1,25%

Projet RI : practice 5v2, rendu n°7 du 5/12

- **Débrief :**
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 3. FlorianArthurJocelynJeoffrey : 8 runs, 8 résultats
  - 1 x 10459 lignes, 7 x 10500 lignes
  - 8 x 0 small irrelevant nodes
  - Practice 4&5, 2*12 runs baseline : non.
  - Practice 5, exo 1 :
    - » Indexation 150sec
    - » Practice 3 ≠ Practice 5 ?
  - Practice 5, exos 2-3-4 ?

	MAgP	P[0.1]
4_ltn_documents_Stop365_Porter	0,95%	2,44%
3_ltn_documents_NoStop_Porter	0,98%	2,69%
8_ltc_documents_NoStop_NoStem	3,59%	9,38%
6_ltc_documents_Stop635_NoStem	3,65%	9,16%
2_ltn_documents_Stop365_NoStemmer	3,74%	10,50%
1_ltn_documents_NoStop_NoStem	3,77%	10,49%
7_ltc_documents_NoStop_Porter	4,02%	9,36%
5_ltc_documents_Stop635_Porter	4,19%	9,71%

Projet RI : practice 5v2, rendu n°7 du 5/12

- **Débrief :**
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 4. MohammedWilliam : 49 runs, 49 résultats
  - 8 x 10500 lignes, 41 x 10499 lignes
  - 49 x 0 small irrelevant nodes
  - 41 runs sur 49 avec des id doc inexistants.
  - Practice 4&5, 2*12 runs baseline : non.
  - Practice 5, exo 1 : work in progress…
  - Practice 5, exos 2-3-4 : ?

	MAgP	P[0.1]
40_bm25_article[1]_stop211_porter_k0.1_b0.5	17,13%	35,14%
30_bm25_article[1]_stop211_porter_k1.2_b0.4	17,02%	32,60%
41_bm25_article[1]_stop211_porter_k0.2_b0.5	17,02%	34,94%
42_bm25_article[1]_stop211_porter_k0.3_b0.5	16,93%	34,29%
43_bm25_article[1]_stop211_porter_k0.4_b0.5	16,81%	33,75%
47_bm25_article[1]_stop211_porter_k0.8_b0.5	16,78%	33,76%
48_bm25_article[1]_stop211_porter_k0.9_b0.5	16,73%	33,22%
45_bm25_article[1]_stop211_porter_k0.6_b0.5	16,72%	33,75%
44_bm25_article[1]_stop211_porter_k0.5_b0.5	16,72%	33,75%
46_bm25_article[1]_stop211_porter_k0.7_b0.5	16,67%	33,75%
49_bm25_article[1]_stop211_porter_k1.0_b0.5	16,57%	32,97%
31_bm25_article[1]_stop211_porter_k1.2_b0.5	16,40%	32,60%
25_bm25_article[1]_stop211_porter_k1.2_b0.5	16,40%	32,60%
24_bm25_article[1]_stop211_porter_k1.2_b0.5	16,40%	32,60%

Projet RI : practice 5v2, rendu n°7 du 5/12

- **Débrief :**
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 5. AlexandreIanOpheliePierre : 24 runs, 24 résultats
  - 24 x 10500 lignes
  - 24 x 0 small irrelevant nodes
  - Practice 4&5, 2*12 runs baseline :
    - » Idem Practice 5v1 ?
  - Practice 5, exo 1 : work in progress?
  - Practice 5, exos 2-3-4 : ?

	MAgP	P[0.1]
3_bm25_article_stop671_porter_k1.2_b0.75	13,30%	31,58%
22_ltn_article_stop671_porter	13,08%	31,27%
24_bm25_article_stop671_porter_k1.2_b0.75	12,91%	26,67%
19_ltn_article_stop671_nostem	12,65%	28,92%
12_bm25_article_nostop_porter_k1.2_b0.75	12,65%	33,29%
21_bm25_article_stop671_nostem_k1.2_b0.75	12,58%	27,77%
15_bm25_article_nostop_nostem_k1.2_b0.75	12,41%	22,27%
9_bm25_article_nostop_nostem_k1.2_b0.75	12,37%	31,22%
6_bm25_article_stop671_nostem_k1.2_b0.75	12,12%	29,27%
18_bm25_article_nostop_porter_k1.2_b0.75	11,99%	23,65%
16_ltn_article_nostop_porter	11,77%	25,95%
13_ltn_article_nostop_nostem	11,45%	23,71%

Projet RI : practice 5v2, rendu n°7 du 5/12

- **Débrief :**
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 6. MahmoudMohammedEmrick : 0 run, 0 résultat
  - Practice 4&5, 2*12 runs baseline : non.
  - Practice 5, exo 1 : ?
  - Practice 5, exos 2-3-4 : work in progress?

Projet RI : practice 5v2, rendu n°7 du 5/12

- **Débrief :**
  - Practice 4, exo 4 : 12 runs « baseline ».
  - Practice 5 :
    - 12 runs : baseline Practice 3 = baseline Practice 5 ?
    - Exo 1 : indexer XML, stats collection.
    - Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)

- 7. SaadZakariaBadreddineIgnacio : 24 runs, 24 résultats
  - 24 x 10500 lignes
  - 13 x 0, 11 x 12-38 small irrelevant nodes
  - Practice 4, 12 runs baseline : non.
  - Practice 5, exo 1 : 12 runs baseline ?
  - Practice 5, exos 2-3-4 : work in progress?
  - Runs éléments :
    - » /bdy[1]/sec[4]/p[1]
    - » /article[1] (seulement)

	MAgP	P[0.1]
11_BM25_article_nostop_porter_k1.2_b0.75	13,55%	38,01%
4_ltn_element_nostop_nostem_noparameters	14,59%	34,30%
4_ltn_article_nostop_nostem_noparameters	14,59%	34,30%
2_ltn_article_stop671_nostem_noparameters	14,96%	36,47%
3_ltn_element_stop671_nostem_noparameters	14,96%	36,47%
11_ltn_element_nostop_porter_k1.2_b0.75	16,15%	40,70%
9_BM25_article_stop671_porter_k1.2_b0.75	16,35%	45,49%
9_ltn_element_stop671_porter_k1.2_b0.75	18,00%	45,70%
12_BM25_article_nostop_nostem_k1.2_b0.75	19,15%	46,66%
10_BM25_article_stop671_nostem_k1.2_b0.75	19,74%	49,31%
12_ltn_element_nostop_nostem_k1.2_b0.75	20,38%	44,26%
10_ltn_element_stop671_nostem_k1.2_b0.75	21,92%	52,00%

Université Jean Monnet Saint-Étienne

Prochain rendu (n°8)

- Practice5v3 : consignes inchangées (mais : on n'est pas en avance…).
- Date limite : 12/12. Priorités :
  - 1) 24 runs « baseline ltn / ltc / bm25 » : Practice 4 (texte), Practice 5 (XML). Vérifier : baseline Practice 3 = baseline Practice 5
  - 2) Practice 5 :
    - 2.1) Exo 1 : indexer XML, stats collection.
    - 2.2) Exos 2-3-4 : produire un run « éléments » qui fonctionne ! (i.e. : sans overlap, sans entrelacement, etc.)
    - 2.3) Exos 2-3-4 : ltn, ltc, BM25 au niveau « éléments », avec une stratégie Fetch & Browse (ou autre !).
    - 2.4) BM25 « éléments » performant.
    - 2.5) Wilkinson & Robertson (BM25F).
    - 2.6) LNU, BM25L, etc.

Practice 5v3, rendu n°8

Projet RI : practice 5v3, rendu n°8 du 12/12

- Practice5v3 : consignes inchangées (mais : on n'est pas en avance…).
- Date limite : 12/12. Priorités :
  - 1) 24 runs « baseline ltn / ltc / bm25 » : vérifier : baseline Practice 4 (texte) = baseline Practice 5 (XML). Rappel :
    - 24 = |{ltn,ltc,bm25}| * |{nostop,stop671}| * |{nostem,porter}| * |{Practice_03, Practice_05}|
    - Granularity = article
    - Queries: ignore the "+"
    - Tokenization: terms without digits or special characters
    - Stop-list: stop-words-english4.txt in this list : https://storage.googleapis.com/google-code-archive-downloads/v2/code.google.com/stop-words/stop-words-collection-2011.11.21.zip
    - Stemming: Porter
    - BM25 parameters: k1=1.2 ; b=0.75
  - 2) Practice 5 :
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.).
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - BM25 « éléments » performant.
    - Wilkinson & Robertson (BM25F).
    - LNU, BM25L, etc.

Projet RI : practice 5v3, rendu n°8 du 12/12

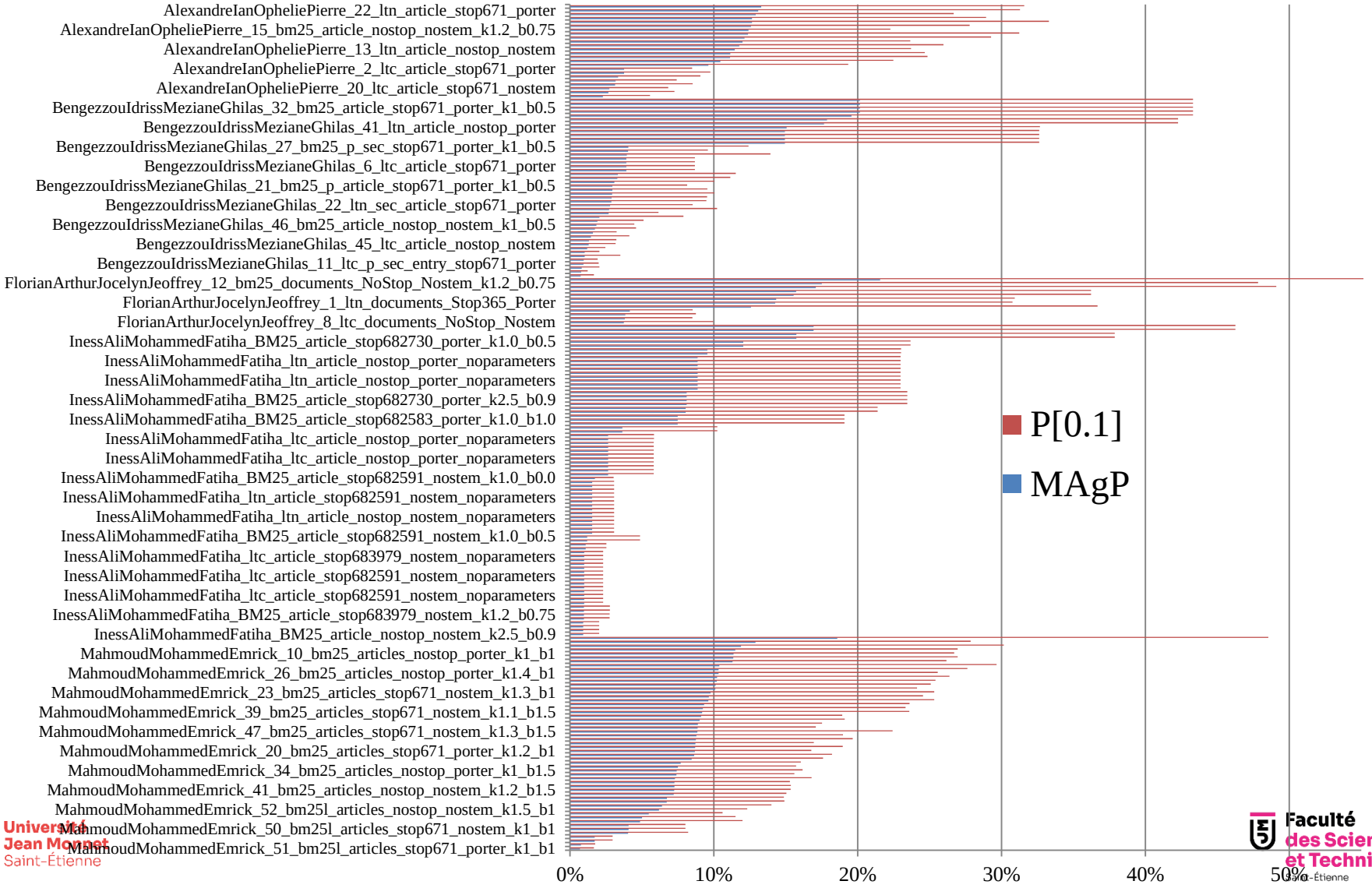
- **Remarques :**
  - 281 runs pour 6 équipes (sur 7*100 = 700 runs possibles).
  - J'ai pu traiter des runs tardifs (cette fois-ci).
  - 143 runs « elements » (3 équipes)
  - 217 résultats
  - Pour plus d'infos : cf. les sorties d'*inex_eval* sur cours en ligne :
    - resultats_runs4.tar

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
  - Practice 5 :
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

Université Jean Monnet Saint-Étienne

Faculté des Sciences et Techniques Saint-Étienne

Projet RI : practice 5v3, rendu n°8 du 12/12



Projet RI : practice 5v3, rendu n°8 du 12/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 1. InessAliMohamedFatiha : 80 runs, 80 résultats
  - 39 x 10500 lignes, 39 x 0 small irrelevant nodes
  - 41 x 9000 lignes, 41 x 0 small irrelevant nodes
  - InessAliMohammedFatiha9_ltc_article_nostop_nostem_noparameters
  - InessAliMohammedFatiha8_ltc_article_nostop_porter_noparameters
  - InessAliMohammedFatiha_ltn_article_stop682583_porter_noparameters
  - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - » BM25_article_stop792682_porter_k1.2_b0.75 16,92% 46,25%
    - » ltn_article_stop792682_porter_noparameters 15,74% 37,86%
    - » ltc_article_stop792682_porter_noparameters 3,63% 10,25%
  - Indexation XML ?
  - 0 run "elements" ?
  - Exos 2-3-4 ?
  - Wilkinson & Robertson (BM25F), LNU, BM25L, etc. ?

Projet RI : practice 5v3, rendu n°8 du 12/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 2. BengezzouIdrissMezianeGhilas : 46 runs, 46 résultats
  - 3 x 10278 lignes, 3 x 10424 lignes, 40 x 10500 lignes
  - 46 x 0 small irrelevant nodes
  - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - » 4_bm25_article_stop671_porter_k1_b0.5 20,16% 43,30%
    - » 38_ltn_article_stop671_nostem17,86% 42,27%
    - » 42_ltc_article_nostop_porter 4,04% 9,58%
  - Indexation XML ok.
  - 24 runs "elements" ?
    - » Plusieurs calculs de df
    - » Plusieurs granularités
    - » 27_bm25_p_sec_stop671_porter_k1_b0.5 4,05% 12,41%
  - Exos 2-3-4 ?
  - Wilkinson & Robertson (BM25F), LNU, BM25L, etc. ?

Projet RI : practice 5v3, rendu n°8 du 12/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 3. FlorianArthurJocelynJeoffrey : 12 runs, 12 résultats
  - 12 x 10500 lignes, 12 x 0 small irrelevant nodes
  - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - » 11_bm25_documents_Stop365_NoStem_k1.2_b0.75 21,55% 57,39%
    - » 3_ltn_documents_Stop365_NoStem 15,71% 36,22%
    - » 6_ltc_documents_NoStop_Porter 4,15% 8,52%
    - » Stats ➔ Practice 4 ≠ Practice 5 ?
  - Indexation XML ok ?
  - 0 run "elements" ?
  - Exos 2-3-4 ?
  - Wilkinson & Robertson (BM25F), LNU, BM25L, etc. ?

Projet RI : practice 5v3, rendu n°8 du 12/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 4. MohammedWilliam : 64 runs, 0 résultats
  - 64 x 10500 lignes, 64 x 0 small irrelevant nodes
  - ERROR *** (84_bm25_article[1]_stop211_nostem_k3.7_b0.9) article results are interleaved in topic 2009011 article id: 22479
    - » 2009011 Q0 62784 3 12.31424755740994 MohammedWilliam84 /article[1]
    - » 2009011 Q0 22479 4 11.925402358943323 MohammedWilliam84 /article[1]
    - » 2009011 Q0 439973 5 11.688700476616141 MohammedWilliam84 /article[1]
    - » 2009011 Q0 22479 6 11.22450455695122 MohammedWilliam84 /article[1]/plant[1]
  - 70_smart_lnu_article[1]_stop211_nostem_slope0.3.txt... Loaded 10500 results ... Oops! Skipping malformed result element
  - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - » « cohérent »
  - Indexation XML ok.
  - 64 runs "elements" ?
    - » 13_smart_ltn_article[1]_nostop_porter__2K
  - Exos 2-3-4 ?
  - Wilkinson & Robertson (BM25F), LNU, BM25L, etc. : LNU ?

Projet RI : practice 5v3, rendu n°8 du 12/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 5. AlexandreIanOpheliePierre : 24 runs, 24 résultats
  - 24 x 10500 lignes, 24 x 0 small irrelevant nodes
  - 24 runs « baseline » : Practice 4 (texte) ≠ Practice 5 (XML) ?
    - » 3_bm25_article_stop671_porter_k1.2_b0.75 13,30% 31,58%
    - » 22_ltn_article_stop671_porter 13,08% 31,27%
    - » 2_ltc_article_stop671_porter 3,77% 8,50%
    - » Stats Practice 4 ≠ Stats Practice 5
  - Indexation XML ?
  - 0 run "elements" ?
  - Exos 2-3-4 ?
  - Wilkinson & Robertson (BM25F), LNU, BM25L, etc. ?

Projet RI : practice 5v3, rendu n°8 du 12/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 6. MahmoudMohammedEmrick : 55 runs, 55 résultats
  - 55 x 10500 lignes, 55 x 0 small irrelevant nodes
  - 2009011 Q0 460499 2 14995.951882077597 MahmoudMohammedEmrick article[1]/bdy[1]/sec[1]/p[3]/link[1]
  - (577 500 occurrences)
  - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - » 4_ltn_articles_nostop_nostem 18,57% 48,54%
    - » 22_bm25_articles_nostop_porter_k1.3_b1 11,49% 26,94%
    - » 48_bm25l_articles_nostop_nostem_k1_b1 7,30% 16,79%
    - » 7_ltc_articles_stop671_porter 5,47% 10,60%
  - Indexation XML ?
    - » parseur doit conserver le contenu interne des balises ?
  - 55 runs "elements" ?
    - » MahmoudMohammedEmrick_1_ltn_articles_stop671_porter.txt
  - Exos 2-3-4 ?
  - Wilkinson & Robertson (BM25F), LNU, BM25L, etc. ?
    - » « paramètres b, k et delta » : LNU ?

Projet RI : practice 5v3, rendu n°8 du 12/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 7. SaadZakariaBadreddineIgnacio : 0 run.
  - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
  - Indexation XML ?
  - Runs "elements" ?
  - Exos 2-3-4 ?
  - Wilkinson & Robertson (BM25F), LNU, BM25L, etc. ?

Projet RI : practice 5v3, rendu n°8 du 12/12

- **Meilleurs scores**
  - 3 années précédentes : MAgP : 0,2770 ; P[0.1] : 0,6247
  - Practice 4, rendu n°5 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	Q5_bm25_article_stop1160_nostem_k1.2_b0.75	0,1542	Q5_bm25_article_stop1160_nostem_k1.2_b0.75	0,4062
BengezzouIdrissMezianeGhilas	12_bm25_articles_stop671_nostem	0,2186	12_bm25_articles_stop671_nostem	0,5427
FlorianArthurJocelynJeoffrey	9_bm25_documents_stop179_porter_k1.2_b0.75	0,1493	9_bm25_documents_stop179_porter_k1.2_b0.75	0,4218
MahmoudMohammedEmrick	13_bm25_articles_stop671_nostem_k1.2_b0.75	0,1731	50_bm25_articles_stop671_nostem_k4.0_b0.75	0,3561
MohammedWilliam	28_bm25_article[1]_stop211_porter_k0.2_b0.5	0,1941	23_bm25_article[1]_stop211_porter_k1.2_b0.7	0,4333
SaadZakariaBadreddineIgnacio	10_BM25_article_stop671_nostem_k1.2_b0.75	0,1968	10_BM25_article_stop671_nostem_k1.2_b0.75	0,4759

  - Practice 5, rendu n°6 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,133	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	6_bm25_articles_stop671_nostem_k1.2_b0.75	0,2025	6_bm25_articles_stop671_nostem_k1.2_b0.75	0,5287
MahmoudMohammedEmrick	35_bm25_articles_stop671_nostem_k1.1_b0.75	0,1734	26_bm25_articles_stop671_nostem_k1.2_b1	0,3806
MohammedWilliam	83_bm25_article[1]_nostop_porter_k1.2_b0.5	0,0612	85_bm25_article[1]_stop211_porter_k1.2_b0.5	0,2208

  - Practice 5v2, rendu n°7 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,1330	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	15_bm25_bdy_stop671_porter	0,2258	15_bm25_bdy_stop671_porter	0,4984
FlorianArthurJocelynJeoffrey	5_ltc_documents_Stop635_Porter	0,0419	2_ltn_documents_Stop365_NoStemmer	0,1050
MohammedWilliam	40_bm25_article[1]_stop211_porter_k0.1_b0.5	0,1713	33_bm25_article[1]_stop211_porter_k1.2_b0.7	0,3916
SaadZakariaBadreddineIgnacio	10_ltn_element_stop671_nostem_k1.2_b0.75	0,2192	10_ltn_element_stop671_nostem_k1.2_b0.75	0,5200

  - Practice 5v3, rendu n°8 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,1330	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	1_bm25_article_stop671_porter_k1_b0.5	0,2016	1_bm25_article_stop671_porter_k1_b0.5	0,4330
FlorianArthurJocelynJeoffrey	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,2155	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,5739
InessAliMohammedFatiha	BM25_article_stop792682_porter_k1.2_b0.75	0,1692	BM25_article_stop792682_porter_k1.2_b0.75	0,4625
MahmoudMohammedEmrick	4_ltn_articles_nostop_nostem	0,1857	4_ltn_articles_nostop_nostem	0,4854

Prochain rendu (n°9)

- Practice5v4 : consignes inchangées (mais : on n'est toujours pas en avance…).
- Date limite : 19/12. Priorités :
  - 1) 24 runs « baseline ltn / ltc / bm25 » : vérifier : baseline Practice 4 (texte) = baseline Practice 5 (XML). Rappel :
    - 24 = |{ltn,ltc,bm25}| * |{nostop,stop671}| * |{nostem,porter}| * |{Practice_03, Practice_05}|
    - Granularity = article
    - Queries: ignore the "+"
    - Tokenization: terms without digits or special characters
    - Stop-list: stop-words-english4.txt in this list : https://storage.googleapis.com/google-code-archive-downloads/v2/code.google.com/stop-words/stop-words-collection-2011.11.21.zip
    - Stemming: Porter
    - BM25 parameters: k1=1.2 ; b=0.75
  - 2) Practice 5 :
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.).
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - BM25 « éléments » performant.
    - Wilkinson & Robertson (BM25F).
    - LNU, BM25L, etc.
  - Rappel : maxi 100 runs.

Practice 5v4, rendu n°9

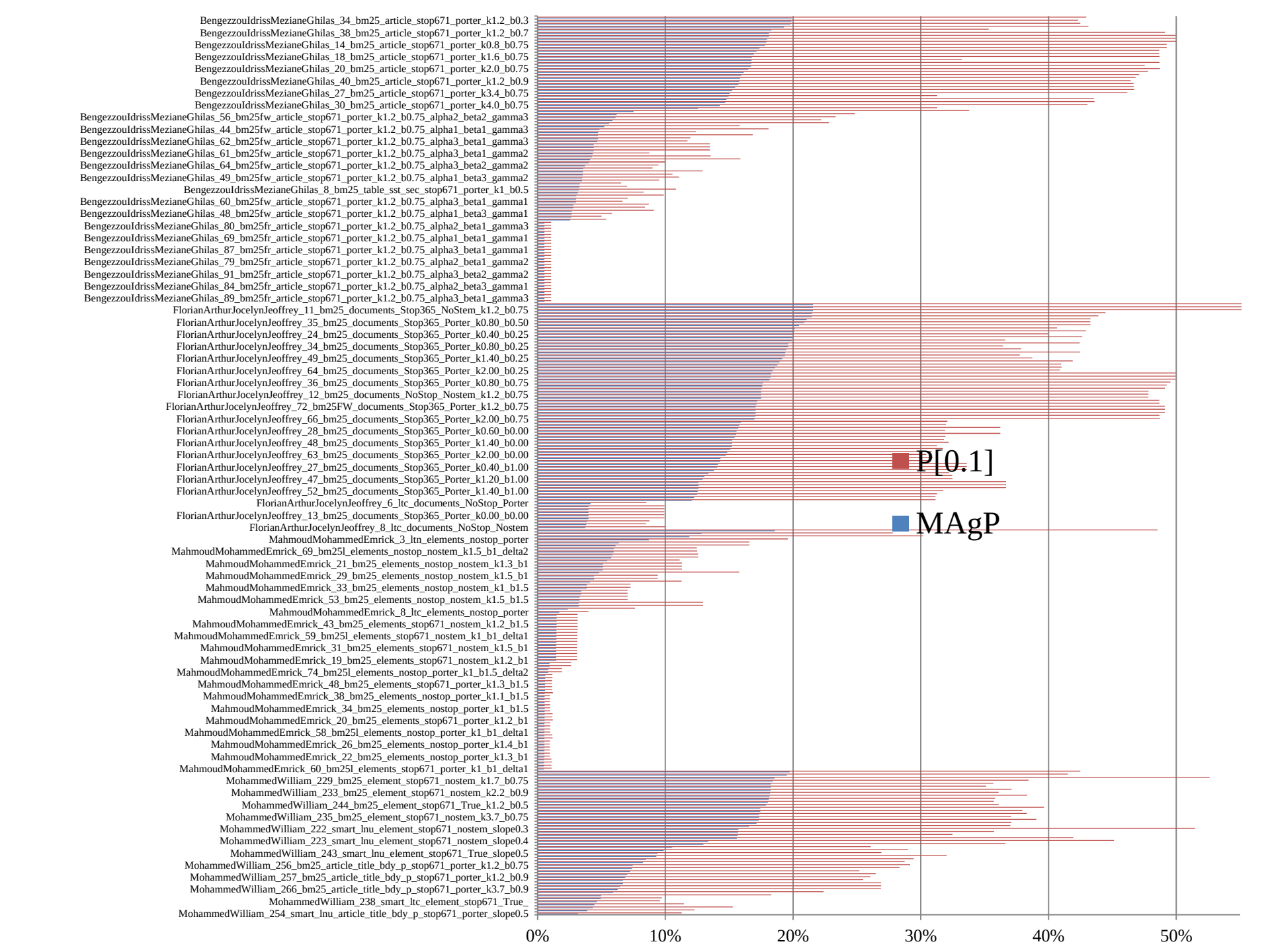
Projet RI : practice 5v4, rendu n°9 du 19/12

- Practice5v4 : consignes inchangées (mais : on n'est toujours pas en avance…).
- Date limite : 19/12. Priorités :
  - 1) 24 runs « baseline ltn / ltc / bm25 » : vérifier : baseline Practice 4 (texte) = baseline Practice 5 (XML). Rappel :
    - 24 = |{ltn,ltc,bm25}| * |{nostop,stop671}| * |{nostem,porter}| * |{Practice_03, Practice_05}|
    - Granularity = article
    - Queries: ignore the "+"
    - Tokenization: terms without digits or special characters
    - Stop-list: stop-words-english4.txt in this list : https://storage.googleapis.com/google-code-archive-downloads/v2/code.google.com/stop-words/stop-words-collection-2011.11.21.zip
    - Stemming: Porter
    - BM25 parameters: k1=1.2 ; b=0.75
  - 2) Practice 5 :
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.).
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - BM25 « éléments » performant.
    - Wilkinson & Robertson (BM25F).
    - LNU, BM25L, etc.
  - Rappel : maxi 100 runs.

Projet RI : practice 5v4, rendu n°9 du 19/12

- **Remarques :**
  - 298 runs pour 4 équipes (sur 7*100 = 700 runs possibles).
  - 137 runs « éléments » (3 équipes)
  - 298 résultats
  - Pour plus d'infos : cf. les sorties d'*inex_eval* sur cours en ligne :
    - resultats_runs5.tar

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
  - Practice 5 :
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.


BengezzouIdrissMezianeGhilas_34_bm25_article_stop671_porter_k1.2_b0.3
BengezzouIdrissMezianeGhilas_38_bm25_article_stop671_porter_k1.2_b0.7
BengezzouIdrissMezianeGhilas_14_bm25_article_stop671_porter_k0.8_b0.75
BengezzouIdrissMezianeGhilas_18_bm25_article_stop671_porter_k1.6_b0.75
BengezzouIdrissMezianeGhilas_20_bm25_article_stop671_porter_k2.0_b0.75
BengezzouIdrissMezianeGhilas_40_bm25_article_stop671_porter_k1.2_b0.9
BengezzouIdrissMezianeGhilas_27_bm25_article_stop671_porter_k3.4_b0.75
BengezzouIdrissMezianeGhilas_30_bm25_article_stop671_porter_k4.0_b0.75
BengezzouIdrissMezianeGhilas_56_bm25fw_article_stop671_porter_k1.2_b0.75_alpha2_beta2_gamma3
BengezzouIdrissMezianeGhilas_44_bm25fw_article_stop671_porter_k1.2_b0.75_alpha1_beta1_gamma3
BengezzouIdrissMezianeGhilas_62_bm25fw_article_stop671_porter_k1.2_b0.75_alpha3_beta1_gamma3
BengezzouIdrissMezianeGhilas_61_bm25fw_article_stop671_porter_k1.2_b0.75_alpha3_beta1_gamma2
BengezzouIdrissMezianeGhilas_64_bm25fw_article_stop671_porter_k1.2_b0.75_alpha3_beta2_gamma2
BengezzouIdrissMezianeGhilas_49_bm25fw_article_stop671_porter_k1.2_b0.75_alpha1_beta3_gamma2
BengezzouIdrissMezianeGhilas_8_bm25_table_sst_sec_stop671_porter_k1_b0.5
BengezzouIdrissMezianeGhilas_60_bm25fw_article_stop671_porter_k1.2_b0.75_alpha3_beta1_gamma1
BengezzouIdrissMezianeGhilas_48_bm25fw_article_stop671_porter_k1.2_b0.75_alpha1_beta3_gamma1
BengezzouIdrissMezianeGhilas_80_bm25fr_article_stop671_porter_k1.2_b0.75_alpha2_beta1_gamma3
BengezzouIdrissMezianeGhilas_69_bm25fr_article_stop671_porter_k1.2_b0.75_alpha1_beta1_gamma1
BengezzouIdrissMezianeGhilas_87_bm25fr_article_stop671_porter_k1.2_b0.75_alpha3_beta1_gamma1
BengezzouIdrissMezianeGhilas_79_bm25fr_article_stop671_porter_k1.2_b0.75_alpha2_beta1_gamma2
BengezzouIdrissMezianeGhilas_91_bm25fr_article_stop671_porter_k1.2_b0.75_alpha3_beta2_gamma2
BengezzouIdrissMezianeGhilas_84_bm25fr_article_stop671_porter_k1.2_b0.75_alpha2_beta3_gamma1
BengezzouIdrissMezianeGhilas_89_bm25fr_article_stop671_porter_k1.2_b0.75_alpha3_beta1_gamma3
FlorianArthurJocelynJeoffrey_11_bm25_documents_Stop365_NoStem_k1.2_b0.75
FlorianArthurJocelynJeoffrey_35_bm25_documents_Stop365_Porter_k0.80_b0.50
FlorianArthurJocelynJeoffrey_24_bm25_documents_Stop365_Porter_k0.40_b0.25
FlorianArthurJocelynJeoffrey_34_bm25_documents_Stop365_Porter_k0.80_b0.25
FlorianArthurJocelynJeoffrey_49_bm25_documents_Stop365_Porter_k1.40_b0.25
FlorianArthurJocelynJeoffrey_64_bm25_documents_Stop365_Porter_k2.00_b0.25
FlorianArthurJocelynJeoffrey_36_bm25_documents_Stop365_Porter_k0.80_b0.75
FlorianArthurJocelynJeoffrey_12_bm25_documents_NoStop_Nostem_k1.2_b0.75
FlorianArthurJocelynJeoffrey_72_bm25FW_documents_Stop365_Porter_k1.2_b0.75
FlorianArthurJocelynJeoffrey_66_bm25_documents_Stop365_Porter_k2.00_b0.75
FlorianArthurJocelynJeoffrey_28_bm25_documents_Stop365_Porter_k0.60_b0.00
FlorianArthurJocelynJeoffrey_48_bm25_documents_Stop365_Porter_k1.40_b0.00
FlorianArthurJocelynJeoffrey_63_bm25_documents_Stop365_Porter_k2.00_b0.00
FlorianArthurJocelynJeoffrey_27_bm25_documents_Stop365_Porter_k0.40_b1.00
FlorianArthurJocelynJeoffrey_47_bm25_documents_Stop365_Porter_k1.20_b1.00
FlorianArthurJocelynJeoffrey_52_bm25_documents_Stop365_Porter_k1.40_b1.00
FlorianArthurJocelynJeoffrey_6_ltc_documents_NoStop_Porter
FlorianArthurJocelynJeoffrey_13_bm25_documents_Stop365_Porter_k0.00_b0.00
FlorianArthurJocelynJeoffrey_8_ltc_documents_NoStop_Nostem
MahmoudMohammedEmrick_3_ltn_elements_nostop_porter
MahmoudMohammedEmrick_69_bm25l_elements_nostop_nostem_k1.5_b1_delta2
MahmoudMohammedEmrick_21_bm25_elements_nostop_nostem_k1.3_b1
MahmoudMohammedEmrick_29_bm25_elements_nostop_nostem_k1.5_b1
MahmoudMohammedEmrick_33_bm25_elements_nostop_nostem_k1_b1.5
MahmoudMohammedEmrick_53_bm25_elements_nostop_nostem_k1.5_b1.5
MahmoudMohammedEmrick_8_ltc_elements_nostop_porter
MahmoudMohammedEmrick_43_bm25_elements_stop671_nostem_k1.2_b1.5
MahmoudMohammedEmrick_59_bm25l_elements_stop671_nostem_k1_b1_delta1
MahmoudMohammedEmrick_31_bm25_elements_stop671_nostem_k1.5_b1
MahmoudMohammedEmrick_19_bm25_elements_stop671_nostem_k1.2_b1
MahmoudMohammedEmrick_74_bm25l_elements_nostop_porter_k1_b1.5_delta2
MahmoudMohammedEmrick_48_bm25_elements_stop671_porter_k1.3_b1.5
MahmoudMohammedEmrick_38_bm25_elements_nostop_porter_k1.1_b1.5
MahmoudMohammedEmrick_34_bm25_elements_nostop_porter_k1_b1.5
MahmoudMohammedEmrick_20_bm25_elements_stop671_porter_k1.2_b1
MahmoudMohammedEmrick_58_bm25l_elements_nostop_porter_k1_b1_delta1
MahmoudMohammedEmrick_26_bm25_elements_nostop_porter_k1.4_b1
MahmoudMohammedEmrick_22_bm25_elements_nostop_porter_k1.3_b1
MahmoudMohammedEmrick_60_bm25l_elements_stop671_porter_k1_b1_delta1
MohammedWilliam_229_bm25_element_stop671_nostem_k1.7_b0.75
MohammedWilliam_233_bm25_element_stop671_nostem_k2.2_b0.9
MohammedWilliam_244_bm25_element_stop671_True_k1.2_b0.5
MohammedWilliam_235_bm25_element_stop671_nostem_k3.7_b0.75
MohammedWilliam_222_smart_lnu_element_stop671_nostem_slope0.3
MohammedWilliam_223_smart_lnu_element_stop671_nostem_slope0.4
MohammedWilliam_243_smart_lnu_element_stop671_True_slope0.5
MohammedWilliam_256_bm25_article_title_bdy_p_stop671_porter_k1.2_b0.75
MohammedWilliam_257_bm25_article_title_bdy_p_stop671_porter_k1.2_b0.9
MohammedWilliam_266_bm25_article_title_bdy_p_stop671_porter_k3.7_b0.9
MohammedWilliam_238_smart_ltc_element_stop671_True_
MohammedWilliam_254_smart_lnu_article_title_bdy_p_stop671_porter_slope0.5
P[0.1]
MAgP
0%
10%
20%
30%
40%
50%

Projet RI : practice 5v4, rendu n°9 du 19/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 1. InessAliMohamedFatiha : 0 run, 0
  - Parsing XML ok?
  - Runs « éléments » : work in progress?

Projet RI : practice 5v4, rendu n°9 du 19/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 2. BengezzouIdrissMezianeGhilas : 95 runs, 95 résultats
  - 95 x 10500 lignes, 95 x 0 small irrelevant nodes
  - 9 runs éléments, 10500 éléments / run
  - Granularité des éléments ?
    - » Run n°6 : 10500 éléments / 10500, mais bdy[1] ~ /article[1] ?

    009011 Q0 2921902 1 24.420411910175122 BengezzouIdrissMezianeGhilas /article[1]/chemical[1]/bdy[1]

    2009011 Q0 212394 2 23.364931912809034 BengezzouIdrissMezianeGhilas /article[1]/fat[1]/ingredient[1]/material[1]/bdy[1]

    2009011 Q0 22594 3 23.35409255037675 BengezzouIdrissMezianeGhilas /article[1]/therapy[1]/bdy[1]
  - Runs similaires : corrélations ?
  - $BM25F_W$ et $BM25F_R$:
    - » alpha : title, bdy, category
    - » Choix des balises ?
    - » $alpha_i$ = (1,1,1) avec $BM25F_R$ ➔ 0.044… ?
  - Taille du vocabulaire conservé : 25 termes.

36_bm25_article_stop671_porter_k1.2_b0.5	0,1989	0,4295
6_bm25_bdy_stop671_porter_k1_b0.5	0,0752	0,3379
47_bm25fw_article_stop671_porter_k1.2_b0.75_alpha1_beta2_gamma3	0,0618	0,2486
3_bm25_st_sec_ss1_p_template_list_bdy_table_stop671_porter_k1_b0.5	0,0425	0,1356
71_bm25fr_article_stop671_porter_k1.2_b0.75_alpha1_beta1_gamma3	0,0053	0,0106

Projet RI : practice 5v4, rendu n°9 du 19/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 3. FlorianArthurJocelynJeoffrey : 75 runs, 75 résultats
  - 75 x 10500 lignes, 75 x 0 small irrelevant nodes
  - Practice 4 = Practice 5
  - 0 run éléments
  - BM25L, $BM25F_W$:
    - » Mais encore ?
    - » Quels paramètres pour obtenir le même run que BM25 ?
  - $BM25F_R$: ?

0_bm25L_documents_Stop365_NoStem_k1.2_b0.75	21,55%	57,39%
4_bm25FW_documents_Stop365_NoStem_k1.2_b0.75	21,55%	57,39%
11_bm25_documents_Stop365_NoStem_k1.2_b0.75	21,55%	57,39%
3_ltn_documents_Stop365_NoStem	15,71%	36,22%
6_ltc_documents_NoStop_Porter	4,15%	8,52%

Projet RI : practice 5v4, rendu n°9 du 19/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 4. MohammedWilliam : 48 runs, 48 résultats
  - 48 x 10500 lignes, 48 x 0 small irrelevant nodes
  - 48 runs éléments, 3196 éléments / run
  - Skipping malformed result element
  - Sélection 4 tags : 20 sec ➔ 0.2 sec
  - Fetch & Browse classique ➔ perte d'info ?
  - LNU meilleur que BM25 ?
  - Granularité des éléments ?
    - » Run n°220 : 6057 éléments / 10500, mais /article[1]/category[1] ~ /article[1] ?

220_smart_lnu_element_stop671_nostem_slope0.1	19,74%	42,49%
229_bm25_element_stop671_nostem_k1.7_b0.75	18,41%	38,43%
255_bm25_article_title_bdy_p_stop671_porter_k1.2_b0.5	9,29%	32,04%
250_smart_lnu_article_title_bdy_p_stop671_porter_slope0.1	5,91%	22,39%
219_smart_ltc_element_stop671_nostem_	4,93%	9,73%

2009011 Q0 22478 1 0.020857850952408885 MohammedWilliam220 /article[1]
2009011 Q0 22479 2 0.019667984967779304 MohammedWilliam220 /article[1]/plant[1]
2009011 Q0 1818836 3 0.018358844220367165 MohammedWilliam220 /article[1]/artifact[1]

Projet RI : practice 5v4, rendu n°9 du 19/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 5. AlexandreIanOpheliePierre : 0 run.
  - Runs « baseline » ?
  - Problème parsing XML ?

Projet RI : practice 5v4, rendu n°9 du 19/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 6. MahmoudMohammedEmrick : 80 runs, 80 résultats
  - 80 x 10500 lignes, 6 x 0 + 74 x 2-12 small irrelevant nodes
  - Correction de bugs ?
  - 80 runs éléments, 7939 éléments / run
  - Granularité des éléments ?
    - » Run n°4 : 4370 éléments / 10500, mais /article[1]/.../sec[n] ~ /article[1] ?

    2009011 Q0 1402262 22 14794.135059412498 MahmoudMohammedEmrick /article[1]/disease[1]/ailment[1]/bdy[1]/sec[2]

    2009011 Q0 187886 26 14754.009062628584 MahmoudMohammedEmrick /article[1]/behavior[1]/practice[1]/tradition[1]/service[1]/therapy[1]/bdy[1]/sec[1]

    2009011 Q0 21525 27 14743.95742081274 MahmoudMohammedEmrick /article[1]/bdy[1]
  - LTN meilleur que BM25 ?
  - $BM25F_W$, $BM25F_R$: ?

4_ltn_elements_nostop_nostem	18,57%	48,54%
73_bm25l_elements_nostop_nostem_k1_b1.5_delta2	6,37%	16,56%
9_bm25_elements_nostop_nostem_k1_b1	5,44%	11,12%
5_ltc_elements_nostop_nostem	4,81%	15,78%

Université Jean Monnet Saint-Étienne

Faculté des Sciences et Techniques Saint-Étienne

Projet RI : practice 5v4, rendu n°9 du 19/12

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - Practice 5 :
    - 24 runs « baseline » : Practice 4 (texte) = Practice 5 (XML) ?
    - Exo 1 : indexer XML, stats collection.
    - Exo 2 : 1er run « éléments » (sans overlap, sans entrelacement, etc.)
    - Exos 2-3-4 : ltn, ltc, BM25 « éléments » (Fetch & Browse ou autre !).
    - Wilkinson & Robertson (BM25F), LNU, BM25L, etc.

- 7. SaadZakariaBadreddineIgnacio :
  - Ben alors ?

Projet RI : practice 5v4, rendu n°9 du 19/12

- **Meilleurs scores**
  - 3 années précédentes : MAgP : 0,2770 ; P[0.1] : 0,6247
  - Practice 4, rendu n°5 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	Q5_bm25_article_stop1160_nostem_k1.2_b0.75	0,1542	Q5_bm25_article_stop1160_nostem_k1.2_b0.75	0,4062
BengezzouldrissMezianeGhilas	12_bm25_articles_stop671_nostem	0,2186	12_bm25_articles_stop671_nostem	0,5427
FlorianArthurJocelynJeoffrey	9_bm25_documents_stop179_porter_k1.2_b0.75	0,1493	9_bm25_documents_stop179_porter_k1.2_b0.75	0,4218
MahmoudMohammedEmrick	13_bm25_articles_stop671_nostem_k1.2_b0.75	0,1731	50_bm25_articles_stop671_nostem_k4.0_b0.75	0,3561
MohammedWilliam	28_bm25_article[1]_stop211_porter_k0.2_b0.5	0,1941	23_bm25_article[1]_stop211_porter_k1.2_b0.7	0,4333
SaadZakariaBadreddineIgnacio	10_BM25_article_stop671_nostem_k1.2_b0.75	0,1968	10_BM25_article_stop671_nostem_k1.2_b0.75	0,4759

  - Practice 5, rendu n°6 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,133	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouldrissMezianeGhilas	6_bm25_articles_stop671_nostem_k1.2_b0.75	0,2025	6_bm25_articles_stop671_nostem_k1.2_b0.75	0,5287
MahmoudMohammedEmrick	35_bm25_articles_stop671_nostem_k1.1_b0.75	0,1734	26_bm25_articles_stop671_nostem_k1.2_b1	0,3806
MohammedWilliam	83_bm25_article[1]_nostop_porter_k1.2_b0.5	0,0612	85_bm25_article[1]_stop211_porter_k1.2_b0.5	0,2208

  - Practice 5v2, rendu n°7 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,1330	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouldrissMezianeGhilas	15_bm25_bdy_stop671_porter	0,2258	15_bm25_bdy_stop671_porter	0,4984
FlorianArthurJocelynJeoffrey	5_ltc_documents_Stop635_Porter	0,0419	2_ltn_documents_Stop365_NoStemmer	0,1050
MohammedWilliam	40_bm25_article[1]_stop211_porter_k0.1_b0.5	0,1713	33_bm25_article[1]_stop211_porter_k1.2_b0.7	0,3916
SaadZakariaBadreddineIgnacio	10_ltn_element_stop671_nostem_k1.2_b0.75	0,2192	10_ltn_element_stop671_nostem_k1.2_b0.75	0,5200

  - Practice 5v3, rendu n°8 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,1330	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouldrissMezianeGhilas	1_bm25_article_stop671_porter_k1_b0.5	0,2016	1_bm25_article_stop671_porter_k1_b0.5	0,4330
FlorianArthurJocelynJeoffrey	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,2155	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,5739
InessAliMohammedFatiha	BM25_article_stop792682_porter_k1.2_b0.75	0,1692	BM25_article_stop792682_porter_k1.2_b0.75	0,4625
MahmoudMohammedEmrick	4_ltn_articles_nostop_nostem	0,1857	4_ltn_articles_nostop_nostem	0,4854

  - Practice 5v4, rendu n°9 :

Equipe	MAgP		P[0.1]	
BengezzouldrissMezianeGhilas	36_bm25_article_stop671_porter_k1.2_b0.5	0,1989	13_bm25_article_stop671_porter_k0.6_b0.75	0,5001
FlorianArthurJocelynJeoffrey	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,2155	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,5738
MahmoudMohammedEmrick	4_ltn_elements_nostop_nostem	0,1857	4_ltn_elements_nostop_nostem	0,4853
MohammedWilliam	220_smart_lnu_element_stop671_nostem_slope0.1	0,1974	221_smart_lnu_element_stop671_nostem_slope0.2	0,5261

Prochain rendu (n°10)

- Practice6
- Date limite : 9/1 puis 16/1 (dernier rendu).
  - 24 runs « baseline ltn / ltc / bm25 »
  - Run(s) « articles » optimisés (BM25, LNU, BM25L, etc.).
  - Practice 5 :
    - Run(s) « éléments » optimisés.
    - Run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 :
    - Runs « liens » : SNA metrics, popularity, PageRank, HITS, etc.
    - Runs « anchor text ».
  - Rappel : maxi 100 runs (dernier rendu : 200).

Practice 6, rendu n°10

Projet RI : practice 6, rendu n°10 du 9/1

- Practice6
- Date limite : 9/1 puis 16/1 (dernier rendu).
  - 24 runs « baseline ltn / ltc / bm25 »
  - Run(s) « articles » optimisés (BM25, LNU, BM25L, etc.).
  - Practice 5 :
    - Run(s) « éléments » optimisés.
    - Run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 :
    - Runs « liens » : SNA metrics, popularity, PageRank, HITS, etc.
    - Runs « anchor text ».
  - Rappel : maxi 100 runs (dernier rendu : 200).

Projet RI : practice 6, rendu n°10 du 9/1

- **Remarques :**
  - 330 runs pour 5 équipes (sur 7*100 = 700 runs possibles).
  - 108 runs « éléments » (4 équipes)
  - 268 résultats
  - Pour plus d'infos : cf. les sorties d'*inex_eval* sur cours en ligne :
    - resultats_runs6.tar

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

Université Jean Monnet Saint-Étienne

Faculté des Sciences et Techniques Saint-Étienne


AlexandreIanOpheliePierre_24_bm25_article_stop671_porter_k1.2_b0.75
AlexandreIanOpheliePierre_9_bm25_article_nostop_nostem_k1.2_b0.75
AlexandreIanOpheliePierre_7_ltn_article_stop671_nostem
AlexandreIanOpheliePierre_23_ltc_article_stop671_porter
AlexandreIanOpheliePierre_5_ltc_article_nostop_nostem
BengezzouIdrissMezianeGhilas_28_bm25_article_stop670_porter_k1.0_b0.65
BengezzouIdrissMezianeGhilas_42_bm25_article_stop670_porter_k1.2_b0.65
BengezzouIdrissMezianeGhilas_46_bm25_article_stop670_porter_k1.3_b0.5
BengezzouIdrissMezianeGhilas_26_bm25_article_stop670_porter_k1.0_b0.55
BengezzouIdrissMezianeGhilas_40_bm25_article_stop670_porter_k1.2_b0.55
BengezzouIdrissMezianeGhilas_43_bm25_article_stop670_porter_k1.2_b0.7
BengezzouIdrissMezianeGhilas_58_bm25_article_stop670_porter_k1.4_b0.75
BengezzouIdrissMezianeGhilas_45_bm25_article_stop670_porter_k1.2_b0.8
BengezzouIdrissMezianeGhilas_13_bm25_article_nostop_porter_k1_b0.5
BengezzouIdrissMezianeGhilas_88_bm25_article_event_title_stop670_porter_k1_b0.5
BengezzouIdrissMezianeGhilas_21_bm25_article_sec_header_stop670_porter_k1_b0.5
BengezzouIdrissMezianeGhilas_24_bm25_article_sec_p_list_header_ss1_table_row_stop670_porte..
BengezzouIdrissMezianeGhilas_100_bm25_bdy_sec_p_list_ss1_stop670_porter_k1_b0.5
BengezzouIdrissMezianeGhilas_6_ltc_article_stop670_porter
BengezzouIdrissMezianeGhilas_65_bm25fw_article_stop670_porter_k1.2_b0.75_alpha1.5_beta1.0..
BengezzouIdrissMezianeGhilas_63_bm25fw_article_stop670_porter_k1.2_b0.75_alpha1.0_beta1.5..
BengezzouIdrissMezianeGhilas_67_bm25fr_article_stop670_porter_k1.2_b0.75_alpha1.0_beta1.0_..
BengezzouIdrissMezianeGhilas_86_ANCE_msmarco_roberta_base_ance_firstp_article_stop670_p..
BengezzouIdrissMezianeGhilas_78_lnu_article_stop670_porter_slope0.5
BengezzouIdrissMezianeGhilas_83_lnu_article_stop670_porter_slope1.0
FlorianArthurJocelynJeoffrey_4_ltn_documents_NoStop_Nostem
FlorianArthurJocelynJeoffrey_13_bm25_documents_Stop365_NoStem_k1.2_b0.75
FlorianArthurJocelynJeoffrey_10_bm25_documents_NoStop_Porter_k1.2_b0.75
MahmoudMohammedEmrick_18_bm25L_articles_stop671_nostem_k1.2_b0.5_delta0.5
MahmoudMohammedEmrick_24_bm25_articles_nostop_nostem_k1.2_b0.5
MahmoudMohammedEmrick_23_ltc_articles_nostop_nostem
MahmoudMohammedEmrick_14_lnu_articles_nostop_porter_slope0.5
MahmoudMohammedEmrick_6_bm25L_articles_stop671_porter_k1.2_b0.75_delta0.5
MahmoudMohammedEmrick_11_bm25L_articles_nostop_porter_k1.2_b0.5_delta0.5
MohammedWilliam_510_bm25fw_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle0.5_..
MohammedWilliam_538_bm25fw_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle1_al..
MohammedWilliam_527_bm25fr_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle1_alp..
MohammedWilliam_534_bm25fw_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle1_al..
MohammedWilliam_523_bm25fr_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle0.5_a..
MohammedWilliam_678_bm25fw_article_stop671_porter_anchors__k1.2_b0.75_alphaarticle1_alp..
MohammedWilliam_657_bm25fw_article_stop671_porter_anchors__k1.2_b0.75_alphaarticle1_alp..
MohammedWilliam_670_bm25fr_article_stop671_porter_anchors__k1.2_b0.75_alphaarticle1_alph..
MohammedWilliam_669_bm25fw_article_stop671_porter_anchors__k1.2_b0.75_alphaarticle1_alp..
MohammedWilliam_634_bm25fr_article_stop671_porter_anchors__k1.2_b0.75_alphaarticle0.5_al..
MohammedWilliam_40_bm25fw_article_title_bdy_p_stop671_porter_k1.2_b0.75_alphaarticle0.5_..
MohammedWilliam_45_bm25fr_article_title_bdy_p_stop671_porter_k1.2_b0.75_alphaarticle1_alp..
MohammedWilliam_63_bm25fr_article_title_bdy_p_stop671_porter_k1.2_b0.75_alphaarticle1_alp..
MohammedWilliam_66_bm25fw_article_title_bdy_p_stop671_porter_k1.2_b0.75_alphaarticle1_al..
MohammedWilliam_608_bm25_article_stop671_porter_anchors__k1.2_b0.75
MohammedWilliam_10_bm25_article_title_bdy_p_stop671_porter_k1.2_b0.9
MohammedWilliam_602_smart_ltc_article_stop671_porter_anchors_
MohammedWilliam_601_smart_ltn_article_stop671_porter_anchors_
MohammedWilliam_03_smart_lnu_article_title_bdy_p_stop671_porter_slope0.1
MohammedWilliam_603_smart_lnu_article_stop671_porter_anchors__slope0.1
P[0.1]
MAgP
0%
10%
20%
30%
40%
50%

Projet RI : practice 6, rendu n°10 du 9/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 1. InessAliMohamedFatiha : 0 run
  - Parsing XML ok ?
  - « L'utilisateur choisi le type de tags retournés » ?
  - Runs « éléments » ?

Projet RI : practice 6, rendu n°10 du 9/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 2. BengezzouIdrissMezianeGhilas : 100 runs, 100 résultats
  - 100 x 10500 lignes, 1 x 1, 2 x 29-35, 97 x 0 small irrelevant nodes
  - 21 runs éléments, 7349 éléments / run
  - Correction de bugs.
  - Calcul de corrélation entre les runs.
  - « calcul de notre df prenait en compte les xpath »
  - Amélioration du preprocessing
  - Granularité des éléments ?
  - Continuum BM25/BM25FR ?
  - Runs « deeplearning »
  - Description des runs ?

17_bm25_bdy_stop670_porter_k1_b0.5	0,2513	0,51
7_bm25_article_stop670_porter_k1_b0.5	0,2368	0,51
8_ltn_article_stop670_nostem	0,1699	0,39
92_bm25_bdy_header_stop670_porter_k1_b0.5	0,1685	0,37
61_bm25fw_article_stop670_porter_k1.2_b0.75_alpha1.0_be	0,0769	0,22
71_bm25fr_article_stop670_porter_k1.2_b0.75_alpha1.5_bet	0,0528	0,14
85_SBERT_msmarco_distilbert_base_tas_b_article_stop670_p	0,0209	0,04
87_DPR_dpr_ctx_encoder_multiset_base_dpr_ctx_encoder_n	0,0168	0,05
86_ANCE_msmarco_roberta_base_ance_firstp_article_stop67	0,0136	0,04
74_lnu_article_stop670_porter_slope0.1	0,0091	0,0

Projet RI : practice 6, rendu n°10 du 9/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 3. FlorianArthurJocelynJeoffrey : 14 runs, 14 résultats
  - 14 x 10500 lignes
  - 8 x 0 et 6 x ~9500 small irrelevant nodes
  - 6 runs éléments, 9473 éléments / run
  - Granularité des éléments ?

11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,2155	0,5739
12_bm25_documents_NoStop_Nostem_k1.2_b0.75	0,1894	0,5004
3_ltn_documents_Stop365_NoStem	0,1571	0,3622
4_ltn_documents_NoStop_Nostem	0,1527	0,3296
16_bm25_documents_Stop365_NoStem_k0.25_b0.0	0,0337	0,1022
17_bm25_documents_Stop365_NoStem_k1.75_b1.0	0,0330	0,1135
15_bm25_documents_Stop365_NoStem_k0.5_b0.25	0,0327	0,1041
14_bm25_documents_Stop365_NoStem_k1.0_b0.5	0,0327	0,1020
13_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,0321	0,1007
18_bm25_documents_Stop365_NoStem_k2.0_b0.75	0,0317	0,1019
1_ltn_documents_Stop365_Porter	0,0228	0,0674
2_ltn_documents_NoStop_Porter	0,0225	0,0668
9_bm25_documents_Stop365_Porter_k1.2_b0.75	0,0202	0,0485
10_bm25_documents_NoStop_Porter_k1.2_b0.75	0,0185	0,0454

Projet RI : practice 6, rendu n°10 du 9/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 4. MohammedWilliam : 100 runs, 100 résultats
  - 100 x 10500 lignes, 100 x 0 small irrelevant nodes
  - 19 runs éléments, 2737 éléments / run
  - Granularité des éléments ?
  - Exploitation des ancres, PageRank (RSV * PR).
  - Quel run baseline ? (continuum BM25)

537_bm25fr_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle1_alph	0,1637	0,5625
530_bm25fw_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle1_alp	0,1637	0,5625
661_bm25fr_article_stop671_porter_anchors__k1.2_b0.75_alphaarticle1_alpha	0,1067	0,2861
660_bm25fw_article_stop671_porter_anchors__k1.2_b0.75_alphaarticle1_alph	0,1067	0,2861
44_bm25fw_article_title_bdy_p_stop671_porter_k1.2_b0.75_alphaarticle1_alph	0,0781	0,2912
608_bm25_article_stop671_porter_anchors__k1.2_b0.75	0,0046	0,0096
18_bm25_article_title_bdy_p_stop671_porter_k3.7_b0.75	0,0046	0,0102

Projet RI : practice 6, rendu n°10 du 9/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 5. AlexandreIanOpheliePierre : 24 runs.
  - AlexandreIanOpheliePierre : 24 runs, 24 resultats
  - 24 x 10500 lignes, 24 x 0 small irrelevant nodes
  - Rien de neuf ?

Projet RI : practice 6, rendu n°10 du 9/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 6. MahmoudMohammedEmrick : 92 runs, 30 résultats
  - 92 x 10500 lignes, 36 x 0 et 56 x 3500-8000 small irrelevant nodes
  - 62 runs éléments, 9489 éléments / run

  66_bm25_elements_nostop_porter_k1.2_b0.5.trec++) passage overlaps with previously retrieved passages in topic 2009074 article id: 4239838in line: 6431

  2009011 Q0 10323757 128 13733 MahmoudMohammedEmrick /article[1]/organization[1]/bdy[1]/sec[1]/p[3]
  2009011 Q0 10323757 129 13723 MahmoudMohammedEmrick /article[1]/organization[1]/bdy[1]/p[1]
  2009011 Q0 10323757 130 13713 MahmoudMohammedEmrick /article[1]/organization[1]/bdy[1]/sec[1]
  - Runs « élements », profondeur 5
  - Granularité des éléments ?
  - Runs BM25FR, profondeur 2... ???

90_bm25R_elements_stop671_nostem_k1.2_	0,1813	0,5137
22_ltn_articles_nostop_nostem	0,1337	0,3581
18_bm25L_articles_stop671_nostem_k1.2_b0	0,1035	0,2516
17_bm25_articles_stop671_nostem_k1.2_b0.!	0,0944	0,2287

Projet RI : practice 6, rendu n°10 du 9/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 7. SaadZakariaBadreddineIgnacio :
  - Rien de neuf ?

Projet RI : practice 6, rendu n°10 du 9/1

- **Meilleurs scores**
  - 3 années précédentes : MAgP : 0,2770 ; P[0.1] : 0,6247
  - Practice 5, rendu n°6 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,133	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	6_bm25_articles_stop671_nostem_k1.2_b0.75	0,2025	6_bm25_articles_stop671_nostem_k1.2_b0.75	0,5287
MahmoudMohammedEmrick	35_bm25_articles_stop671_nostem_k1.1_b0.75	0,1734	26_bm25_articles_stop671_nostem_k1.2_b1	0,3806
MohammedWilliam	83_bm25_article[1]_nostop_porter_k1.2_b0.5	0,0612	85_bm25_article[1]_stop211_porter_k1.2_b0.5	0,2208

  - Practice 5v2, rendu n°7 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,1330	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	15_bm25_bdy_stop671_porter	0,2258	15_bm25_bdy_stop671_porter	0,4984
FlorianArthurJocelynJeoffrey	5_ltc_documents_Stop635_Porter	0,0419	2_ltn_documents_Stop365_NoStemmer	0,1050
MohammedWilliam	40_bm25_article[1]_stop211_porter_k0.1_b0.5	0,1713	33_bm25_article[1]_stop211_porter_k1.2_b0.7	0,3916
SaadZakariaBadreddineIgnacio	10_ltn_element_stop671_nostem_k1.2_b0.75	0,2192	10_ltn_element_stop671_nostem_k1.2_b0.75	0,5200

  - Practice 5v3, rendu n°8 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,1330	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	1_bm25_article_stop671_porter_k1_b0.5	0,2016	1_bm25_article_stop671_porter_k1_b0.5	0,4330
FlorianArthurJocelynJeoffrey	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,2155	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,5739
InessAliMohammedFatiha	BM25_article_stop792682_porter_k1.2_b0.75	0,1692	BM25_article_stop792682_porter_k1.2_b0.75	0,4625
MahmoudMohammedEmrick	4_ltn_articles_nostop_nostem	0,1857	4_ltn_articles_nostop_nostem	0,4854

  - Practice 5v4, rendu n°9 :

Equipe	MAgP		P[0.1]	
BengezzouIdrissMezianeGhilas	36_bm25_article_stop671_porter_k1.2_b0.5	0,1989	13_bm25_article_stop671_porter_k0.6_b0.75	0,5001
FlorianArthurJocelynJeoffrey	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,2155	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,5738
MahmoudMohammedEmrick	4_ltn_elements_nostop_nostem	0,1857	4_ltn_elements_nostop_nostem	0,4853
MohammedWilliam	220_smart_lnu_element_stop671_nostem_slope0.1	0,1974	221_smart_lnu_element_stop671_nostem_slope0.2	0,5261

  - Practice 6, rendu n°10 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,1330	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	17_bm25_bdy_stop670_porter_k1_b0.5	0,2512	30_bm25_article_stop670_porter_k1.0_b0.75	0,5691
FlorianArthurJocelynJeoffrey	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,2155	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,5739
MahmoudMohammedEmrick	90_bm25R_elements_stop671_nostem_k1.2_b0.5	0,1813	90_bm25R_elements_stop671_nostem_k1.2_b0.5	0,5137
MohammedWilliam	508_bm25fw_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle0.5_alphatitle2_alphabdy1.75_alphap0.15	0,1636	508_bm25fw_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle0.5_alphatitle2_alphabdy1.75_alphap0.15	0,5624

Avant-dernier rendu (n°11)

- Practice6v2 : idem.
- Date limite : 16/1 puis 23/1 (dernier rendu).
  - 24 runs « baseline ltn / ltc / bm25 »
  - Run(s) « articles » optimisés (BM25, LNU, BM25L, etc.).
  - Practice 5 :
    - Run(s) « éléments » optimisés.
    - Run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 :
    - Runs « liens » : SNA metrics, popularity, PageRank, HITS, etc.
    - Runs « anchor text ».
  - Jusqu'à 200 runs.

Practice 6v2, rendu n°11

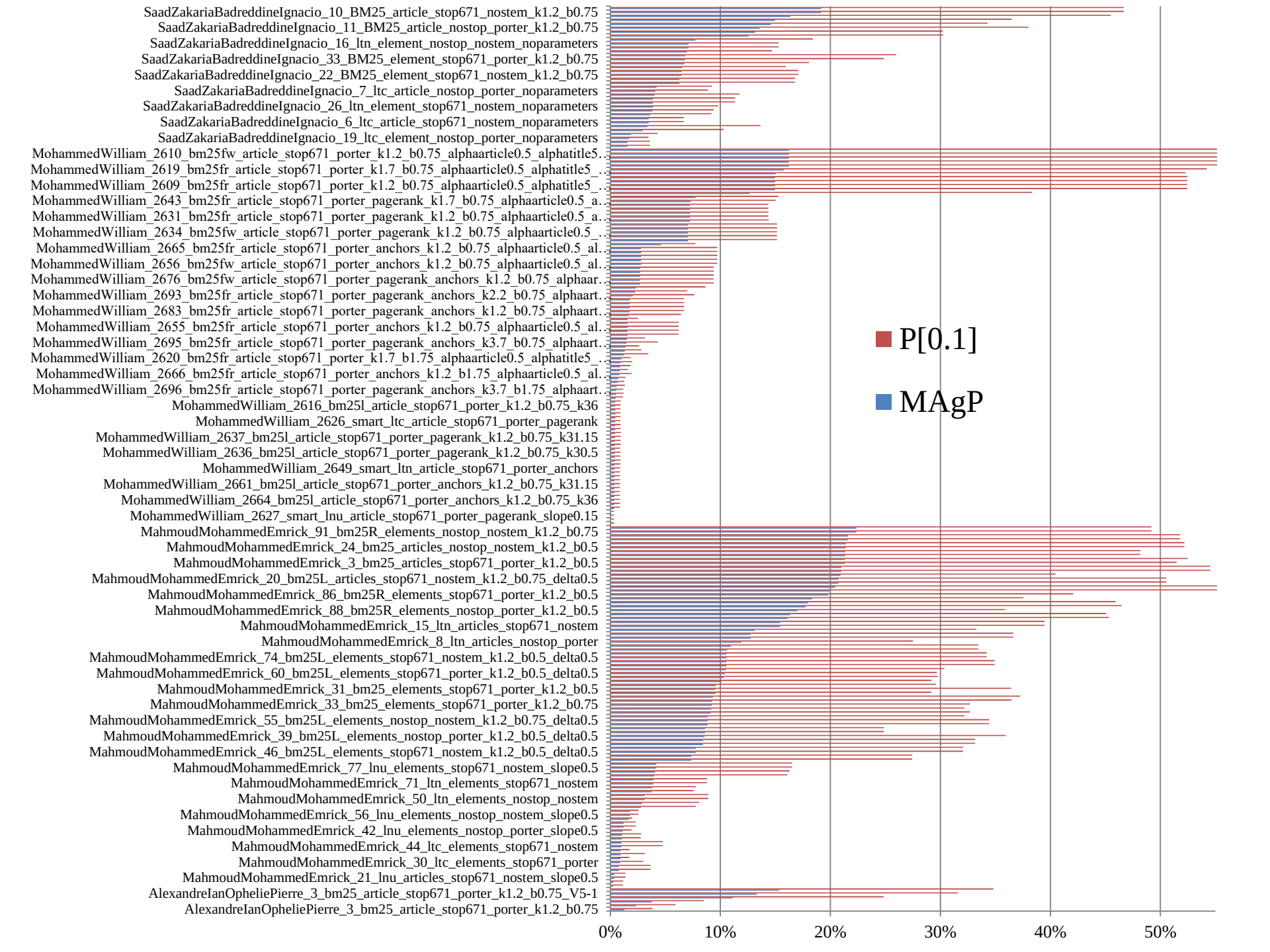
Projet RI : practice 6v2, rendu n°11 du 16/1

- Practice6v2
- Date limite : 16/1 puis 23/1 (dernier rendu).
  - 24 runs « baseline ltn / ltc / bm25 »
  - Run(s) « articles » optimisés (BM25, LNU, BM25L, etc.).
  - Practice 5 :
    - Run(s) « éléments » optimisés.
    - Run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 :
    - Runs « liens » : SNA metrics, popularity, PageRank, HITS, etc.
    - Runs « anchor text ».
  - Jusqu'à 200 runs.

Projet RI : practice 6v2, rendu n°11 du 16/1

- **Remarques :**
  - 230 runs pour 4 équipes (sur 7*200 = 1400 runs possibles).
  - 80 runs « éléments » (2 équipes)
  - 230 résultats
  - Pour plus d'infos : cf. les sorties d'*inex_eval* sur cours en ligne :
    - resultats_runs7.tar

- **Rapide débrief de chaque équipe :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».


SaadZakariaBadreddineIgnacio_10_BM25_article_stop671_nostem_k1.2_b0.75
SaadZakariaBadreddineIgnacio_11_BM25_article_nostop_porter_k1.2_b0.75
SaadZakariaBadreddineIgnacio_16_ltn_element_nostop_nostem_noparameters
SaadZakariaBadreddineIgnacio_33_BM25_element_stop671_porter_k1.2_b0.75
SaadZakariaBadreddineIgnacio_22_BM25_element_stop671_nostem_k1.2_b0.75
SaadZakariaBadreddineIgnacio_7_ltc_article_nostop_porter_noparameters
SaadZakariaBadreddineIgnacio_26_ltn_element_stop671_nostem_noparameters
SaadZakariaBadreddineIgnacio_6_ltc_article_stop671_nostem_noparameters
SaadZakariaBadreddineIgnacio_19_ltc_element_nostop_porter_noparameters
MohammedWilliam_2610_bm25fw_article_stop671_porter_k1.2_b0.75_alphaarticle0.5_alphatitle5…
MohammedWilliam_2619_bm25fr_article_stop671_porter_k1.7_b0.75_alphaarticle0.5_alphatitle5_…
MohammedWilliam_2609_bm25fr_article_stop671_porter_k1.2_b0.75_alphaarticle0.5_alphatitle5_…
MohammedWilliam_2643_bm25fr_article_stop671_porter_pagerank_k1.7_b0.75_alphaarticle0.5_a…
MohammedWilliam_2631_bm25fr_article_stop671_porter_pagerank_k1.2_b0.75_alphaarticle0.5_a…
MohammedWilliam_2634_bm25fw_article_stop671_porter_pagerank_k1.2_b0.75_alphaarticle0.5_…
MohammedWilliam_2665_bm25fr_article_stop671_porter_anchors_k1.2_b0.75_alphaarticle0.5_al…
MohammedWilliam_2656_bm25fw_article_stop671_porter_anchors_k1.2_b0.75_alphaarticle0.5_al…
MohammedWilliam_2676_bm25fw_article_stop671_porter_pagerank_anchors_k1.2_b0.75_alphaar…
MohammedWilliam_2693_bm25fr_article_stop671_porter_pagerank_anchors_k2.2_b0.75_alphaart…
MohammedWilliam_2683_bm25fr_article_stop671_porter_pagerank_anchors_k1.2_b0.75_alphaart…
MohammedWilliam_2655_bm25fr_article_stop671_porter_anchors_k1.2_b0.75_alphaarticle0.5_al…
MohammedWilliam_2695_bm25fr_article_stop671_porter_pagerank_anchors_k3.7_b0.75_alphaart…
MohammedWilliam_2620_bm25fr_article_stop671_porter_k1.7_b1.75_alphaarticle0.5_alphatitle5_…
MohammedWilliam_2666_bm25fr_article_stop671_porter_anchors_k1.2_b1.75_alphaarticle0.5_al…
MohammedWilliam_2696_bm25fr_article_stop671_porter_pagerank_anchors_k3.7_b1.75_alphaart…
MohammedWilliam_2616_bm25l_article_stop671_porter_k1.2_b0.75_k36
MohammedWilliam_2626_smart_ltc_article_stop671_porter_pagerank
MohammedWilliam_2637_bm25l_article_stop671_porter_pagerank_k1.2_b0.75_k31.15
MohammedWilliam_2636_bm25l_article_stop671_porter_pagerank_k1.2_b0.75_k30.5
MohammedWilliam_2649_smart_ltn_article_stop671_porter_anchors
MohammedWilliam_2661_bm25l_article_stop671_porter_anchors_k1.2_b0.75_k31.15
MohammedWilliam_2664_bm25l_article_stop671_porter_anchors_k1.2_b0.75_k36
MohammedWilliam_2627_smart_lnu_article_stop671_porter_pagerank_slope0.15
MahmoudMohammedEmrick_91_bm25R_elements_nostop_nostem_k1.2_b0.75
MahmoudMohammedEmrick_24_bm25_articles_nostop_nostem_k1.2_b0.5
MahmoudMohammedEmrick_3_bm25_articles_stop671_porter_k1.2_b0.5
MahmoudMohammedEmrick_20_bm25L_articles_stop671_nostem_k1.2_b0.75_delta0.5
MahmoudMohammedEmrick_86_bm25R_elements_stop671_porter_k1.2_b0.5
MahmoudMohammedEmrick_88_bm25R_elements_nostop_porter_k1.2_b0.5
MahmoudMohammedEmrick_15_ltn_articles_stop671_nostem
MahmoudMohammedEmrick_8_ltn_articles_nostop_porter
MahmoudMohammedEmrick_74_bm25L_elements_stop671_nostem_k1.2_b0.5_delta0.5
MahmoudMohammedEmrick_60_bm25L_elements_stop671_porter_k1.2_b0.5_delta0.5
MahmoudMohammedEmrick_31_bm25_elements_stop671_porter_k1.2_b0.5
MahmoudMohammedEmrick_33_bm25_elements_stop671_porter_k1.2_b0.75
MahmoudMohammedEmrick_55_bm25L_elements_nostop_nostem_k1.2_b0.75_delta0.5
MahmoudMohammedEmrick_39_bm25L_elements_nostop_porter_k1.2_b0.5_delta0.5
MahmoudMohammedEmrick_46_bm25L_elements_stop671_nostem_k1.2_b0.5_delta0.5
MahmoudMohammedEmrick_77_lnu_elements_stop671_nostem_slope0.5
MahmoudMohammedEmrick_71_ltn_elements_stop671_nostem
MahmoudMohammedEmrick_50_ltn_elements_nostop_nostem
MahmoudMohammedEmrick_56_lnu_elements_nostop_nostem_slope0.5
MahmoudMohammedEmrick_42_lnu_elements_nostop_porter_slope0.5
MahmoudMohammedEmrick_44_ltc_elements_stop671_nostem
MahmoudMohammedEmrick_30_ltc_elements_stop671_porter
MahmoudMohammedEmrick_21_lnu_articles_stop671_nostem_slope0.5
AlexandreIanOpheliePierre_3_bm25_article_stop671_porter_k1.2_b0.75_V5-1
AlexandreIanOpheliePierre_3_bm25_article_stop671_porter_k1.2_b0.75
P[0.1]
MAgP
0%
10%
20%
30%
40%
50%

Projet RI : practice 6v2, rendu n°11 du 16/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 1. InessAliMohamedFatiha : 0 run.
  - ?

Projet RI : practice 6v2, rendu n°11 du 16/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 2. BengezzouIdrissMezianeGhilas : 0 run.
  - ?

Projet RI : practice 6v2, rendu n°11 du 16/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 3. FlorianArthurJocelynJeoffrey : 0 run.
  - ?

Projet RI : practice 6v2, rendu n°11 du 16/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 4. MohammedWilliam : 96 runs, 96 résultats
  - 96 x 10500 lignes, 96 x 0 small irrelevant nodes
  - Correction de bugs.
  - ltn, ltc, lnu, bm25*8, bm25fw*4, bm25fr*4, bm25l*5
  - 0 runs éléments
  - Exploitation des ancres + PageRank :
    - » Baseline, PR, Ancres, PR & Ancres.
    - » Intégration ?

2604_bm25fw_article_stop671_porter_k1.2_b0.75_alphaarticle0.5_alphatitle2_alphabdy1.75_alphap2.15	0,1622	0,5621
2617_bm25fr_article_stop671_porter_k1.2_b0.75_alphaarticle0.5_alphatitle5_alphabdy2.86_alphap2.15	0,1622	0,5621
2647_bm25fr_article_stop671_porter_pagerank_k3.7_b0.75_alphaarticle0.5_alphatitle5_alphabdy2.86_alphap2.15	0,0774	0,1527
2665_bm25fr_article_stop671_porter_anchors_k1.2_b0.75_alphaarticle0.5_alphatitle5_alphabdy2.86_alphap2.15	0,0279	0,0972
2682_bm25fw_article_stop671_porter_pagerank_anchors_k1.2_b0.75_alphaarticle0.5_alphatitle5_alphabdy2.86_alphap2.15	0,027	0,094
2602_smart_ltc_article_stop671_porter_	0,0045	0,0093
2640_bm25l_article_stop671_porter_pagerank_k1.2_b0.75_k36	0,0045	0,0093
2625_smart_ltn_article_stop671_porter_pagerank	0,0043	0,0096
2675_smart_lnu_article_stop671_porter_pagerank_anchors_slope0.15	0,0008	0,0034

Projet RI : practice 6v2, rendu n°11 du 16/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 5. AlexandreIanOpheliePierre : 6 runs, 6 résultats.
  - 6 x 10500 lignes, 6 x 0 small irrelevant nodes
  - AlexandreIanOpheliePierre : 0 runs éléments, 63000 documents
  - Crash PC
  - Scikit-learn
  - 0 runs éléments

1_ltn_article_stop671_porter	0,1535	0,3481
3_bm25_article_stop671_porter_k1.2_b0.75_V5-1	0,133	0,3158
1_ltn_article_stop671_porter_V5-1	0,1112	0,2485
2_ltc_article_stop671_porter_V5-1	0,0377	0,085
2_ltc_article_stop671_porter	0,0235	0,0593
3_bm25_article_stop671_porter_k1.2_b0.75	0,0128	0,0383

Projet RI : practice 6v2, rendu n°11 du 16/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».
- 6. MahmoudMohammedEmrick : 92 runs, 92 résultats
  - 90 x 10500 lignes, 2 x 10454 lignes
  - 36 x 0 + 56 x 1284-6468 small irrelevant nodes
  - 56 runs éléments, 10498 éléments / run
  - Correction overlap.
  - Params BM25FR ? BM25FW ? Liens ?

89_bm25R_elements_stop671_nostem_k1.2_b0.75	0,2236	0,4918
18_bm25L_articles_stop671_nostem_k1.2_b0.5_delta0.5	0,2162	0,5179
17_bm25_articles_stop671_nostem_k1.2_b0.5	0,2142	0,5219
22_ltn_articles_nostop_nostem	0,1544	0,3947
84_lnu_elements_nostop_nostem_slope0.5	0,0418	0,1651
16_ltc_articles_stop671_nostem	0,012	0,0232

Projet RI : practice 6v2, rendu n°11 du 16/1

- **Débrief :**
  - Qu'avez-vous fait ? Temps de calcul ? Tests ? Difficultés ? Questions ?
  - 24 runs « baseline ltn / ltc / bm25 »
  - Practice 5 : indexer XML, run(s) « éléments » optimisés, run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 : runs « liens », runs « anchor text ».

- 7. SaadZakariaBadreddineIgnacio : 36 runs, 36 résultats.
  - 32 x 10500 lignes, 4 x 2147-3367 lignes
  - 12 x 0 + 24 x 1772-8689 small irrelevant nodes
  - 24 runs éléments, 9201 éléments / run
  - practice_report ?

12_BM25_article_nostop_nostem_k1.2_b0.75	0,1915	0,4666
3_ltn_article_stop671_nostem_noparameters	0,1496	0,3647
13_ltn_element_stop671_porter_noparameters	0,0772	0,184
35_BM25_element_nostop_porter_k1.2_b0.75	0,0683	0,2597
5_ltc_article_stop671_porter_noparameters	0,042	0,0922
29_ltc_element_stop671_porter_noparameters	0,0401	0,1175

Projet RI : practice 6v2, rendu n°11 du 16/1

- **Meilleurs scores**
  - 3 années précédentes : MAgP : 0,2770 ; P[0.1] : 0,6247
  - Practice 5, rendu n°6 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,133	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	6_bm25_articles_stop671_nostem_k1.2_b0.75	0,2025	6_bm25_articles_stop671_nostem_k1.2_b0.75	0,5287
MahmoudMohammedEmrick	35_bm25_articles_stop671_nostem_k1.1_b0.75	0,1734	26_bm25_articles_stop671_nostem_k1.2_b1	0,3806
MohammedWilliam	83_bm25_article[1]_nostop_porter_k1.2_b0.5	0,0612	85_bm25_article[1]_stop211_porter_k1.2_b0.5	0,2208

  - Practice 5v2, rendu n°7 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,1330	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	15_bm25_bdy_stop671_porter	0,2258	15_bm25_bdy_stop671_porter	0,4984
FlorianArthurJocelynJeoffrey	5_ltc_documents_Stop635_Porter	0,0419	2_ltn_documents_Stop365_NoStemmer	0,1050
MohammedWilliam	40_bm25_article[1]_stop211_porter_k0.1_b0.5	0,1713	33_bm25_article[1]_stop211_porter_k1.2_b0.7	0,3916
SaadZakariaBadreddineIgnacio	10_ltn_element_stop671_nostem_k1.2_b0.75	0,2192	10_ltn_element_stop671_nostem_k1.2_b0.75	0,5200

  - Practice 5v3, rendu n°8 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,1330	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	1_bm25_article_stop671_porter_k1_b0.5	0,2016	1_bm25_article_stop671_porter_k1_b0.5	0,4330
FlorianArthurJocelynJeoffrey	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,2155	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,5739
InessAliMohammedFatiha	BM25_article_stop792682_porter_k1.2_b0.75	0,1692	BM25_article_stop792682_porter_k1.2_b0.75	0,4625
MahmoudMohammedEmrick	4_ltn_articles_nostop_nostem	0,1857	4_ltn_articles_nostop_nostem	0,4854

  - Practice 5v4, rendu n°9 :

Equipe	MAgP		P[0.1]	
BengezzouIdrissMezianeGhilas	36_bm25_article_stop671_porter_k1.2_b0.5	0,1989	13_bm25_article_stop671_porter_k0.6_b0.75	0,5001
FlorianArthurJocelynJeoffrey	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,2155	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,5738
MahmoudMohammedEmrick	4_ltn_elements_nostop_nostem	0,1857	4_ltn_elements_nostop_nostem	0,4853
MohammedWilliam	220_smart_lnu_element_stop671_nostem_slope0.1	0,1974	221_smart_lnu_element_stop671_nostem_slope0.2	0,5261

  - Practice 6, rendu n°10 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	3_bm25_article_stop671_porter_k1.2_b0.75	0,1330	12_bm25_article_nostop_porter_k1.2_b0.75	0,3329
BengezzouIdrissMezianeGhilas	17_bm25_bdy_stop670_porter_k1_b0.5	0,2512	30_bm25_article_stop670_porter_k1.0_b0.75	0,5691
FlorianArthurJocelynJeoffrey	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,2155	11_bm25_documents_Stop365_NoStem_k1.2_b0.75	0,5739
MahmoudMohammedEmrick	90_bm25R_elements_stop671_nostem_k1.2_b0.5	0,1813	90_bm25R_elements_stop671_nostem_k1.2_b0.5	0,5137
MohammedWilliam	508_bm25fw_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle0.5_alphatitle2_alphabdy1.75_alphap0.15	0,1636	508_bm25fw_article_stop671_porter_pagerank__k1.2_b0.75_alphaarticle0.5_alphatitle2_alphabdy1.75_alphap0.15	0,5624

  - Practice 6v2, rendu n°11 :

Equipe	MAgP		P[0.1]	
AlexandreIanOpheliePierre	1_ltn_article_stop671_porter	0,1535	1_ltn_article_stop671_porter	0,3481
MahmoudMohammedEmrick	89_bm25R_elements_stop671_nostem_k1.2_b0.75	0,2236	6_bm25L_articles_stop671_porter_k1.2_b0.75_delta0.5	0,5656
MohammedWilliam	2604_bm25fw_article_stop671_porter_k1.2_b0.75_alphaarticle0.5_alphatitle2_alphabdy1.75_alphap2.15	0,1622	2604_bm25fw_article_stop671_porter_k1.2_b0.75_alphaarticle0.5_alphatitle2_alphabdy1.75_alphap2.15	0,5621
SaadZakariaBadreddineIgnacio	10_BM25_article_stop671_nostem_k1.2_b0.75	0,1915	10_BM25_article_stop671_nostem_k1.2_b0.75	0,4666

Université Jean Monnet Saint-Étienne

Dernier rendu (n°12)

- Practice6v3 : idem.
- Date limite : 23/1 (dernier rendu).
  - 24 runs « baseline ltn / ltc / bm25 »
  - Run(s) « articles » optimisés (BM25, LNU, BM25L, etc.).
  - Practice 5 :
    - Run(s) « éléments » optimisés.
    - Run(s) « articles + structure » (BM25F, BM25E, etc.).
  - Practice 6 :
    - Runs « liens » : SNA metrics, popularity, PageRank, HITS, etc.
    - Runs « anchor text ».
  - Jusqu'à 200 runs.
- Rendu final du projet : 4 février.