Corso di laurea STS
Information Retrieval
2023/2024

Università degli Studi di Padova

Mini-progetto

Un esempio di Pseudo-RF basato sul clustering con k-NN

Quartuccio Andrea 2001321
Varotto Luca 2054021

2023/2024

Indice della presentazione

- Obiettivi
- Metodi utilizzati
- Esperimenti e discussione dei risultati
- Ottimizzazione dei parametri
- Esempi di applicazioni con alcune query
- Conclusioni

Obiettivi

- Studio di un metodo trovato in letteratura:

  “Lee, K. S., Croft, W. B., & Allan, J. (2008, July). A cluster-based resampling method for pseudo-relevance feedback.”
- Impatto dello pseudo-RF basato su un clustering k-NN sull’AP e sulla P@5
- Confronto con la configurazione standard di Elasticsearch con il BM25

Indicizzazione e reperimento

- Indicizzazione senza stop-word
- Reperimento basato sul BM-25, si reperiscono N documenti per ogni interrogazione
- Funzioni di post-reperimento:
  - Pseudo-RF basato sul clustering k-NN

k-NN

"Next, clusters are generated by k-nearest neighbors (k-NN) clustering method for the top-retrieved N documents to find dominant documents. (In experiments, N is set to 100.) Note that one document can belong to several clusters.
In k-NN clustering, each document plays a central role in making its own cluster with its k closest neighbors by similarity."

Documenti dominanti con il k-NN

"In overlapped clusters, a dominant document will appear in multiple highly-ranked clusters. [...] A document that deals with all subtopics will likely be in all subtopic clusters, so we call that document dominant. From such a dominant document, expansion terms that retrieve documents related to all subtopics can be selected."

Funzione post-reperimento I

- Creazione di una DTM con i TF-IDF
- Clustering con k-NN: la scelta di k=5
- Identificazione documenti dominanti: la scelta della soglia al 60%
- Se non ci sono documenti dominanti restituisce i risultati del BM25
- Aggregazione dei testi dei cluster contenenti documenti dominanti: i mega-testi

Funzione post-reperimento II

- Reperimento sui mega-testi
- Scelta del numero dei descrittori per la QE: *e*
- Scelta del peso dei descrittori nella nuova query: *λ*
- Secondo reperimento con la QE sulla collezione

Riassunto della funzione di post-reperimento

- Primo reperimento sulla collezione
- Formazione dei cluster
- Individuazione dei documenti dominanti
- Formazione dei mega-testi
- Reperimento sui mega-testi
- Q.E.
- Secondo reperimento sulla collezione

Indice degli esperimenti

- Collezione sperimentale
- Strumenti
- Descrizione dei risultati
- Esempi di query
- Significatività dei risultati
- Discussione

Collezione sperimentale

- Robust 2004
- Divisione in training e testing dei topics
  - primi 150 topics come training
  - ultimi 100 topics come test
- Stime dei parametri basate sul training
- Valutazione dei modelli basata sul test

Strumenti

- trec-eval
- Python 3.12 e le librerie:
  - elasticsearch
  - scikit-learn
  - ntlk
  - numpy
  - time
  - math
  - concurrent
- R e librerie:
  - ggplot

La scelta della soglia

- La scelta del 60%
- Eseguendo una serie di tentativi con il resto dei parametri fissi forniva i risultati migliori per map e P@5
- Lunghi tempi computazionali

Grafico map~N

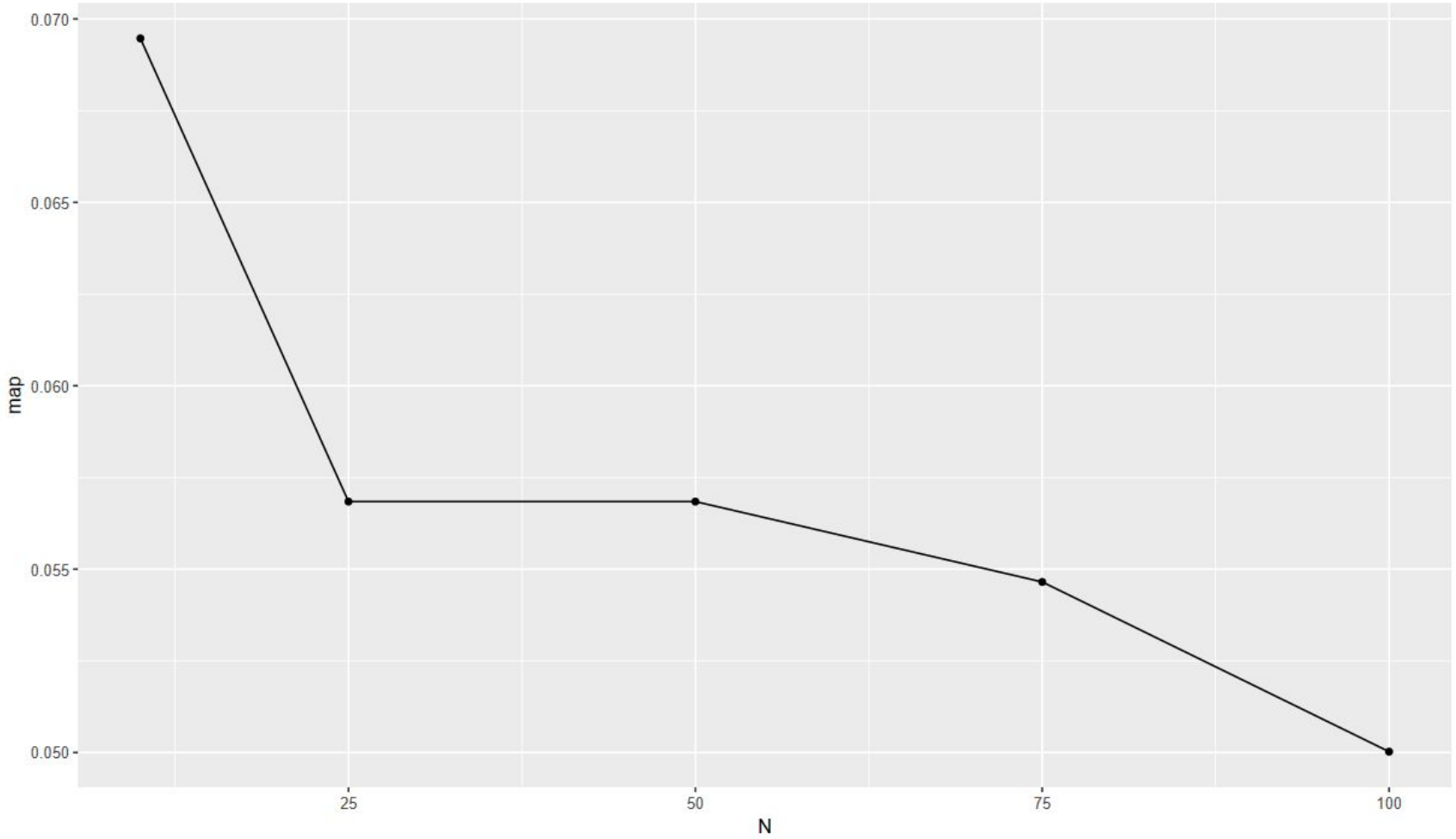


Grafico map~e

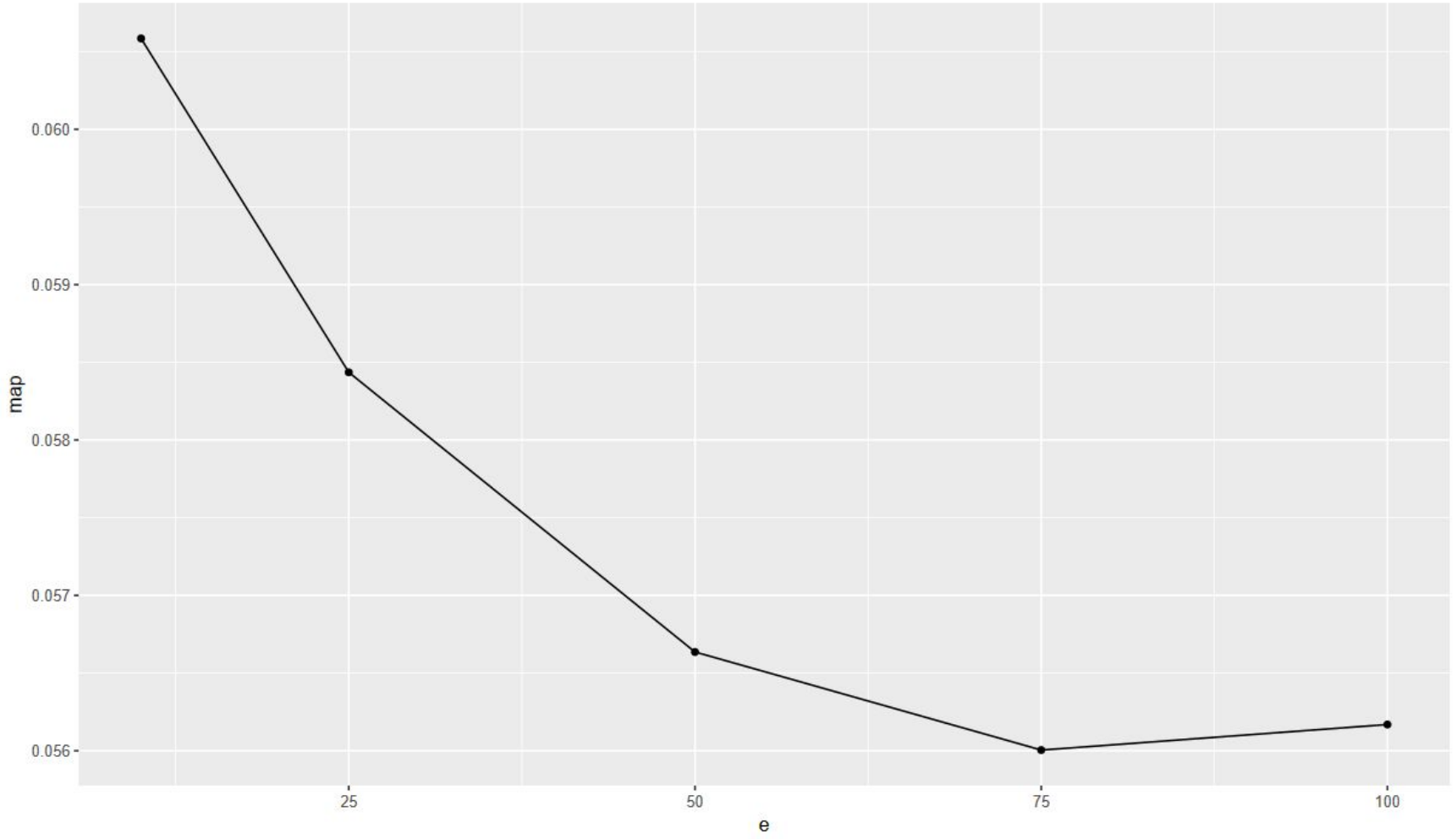


Grafico map~λ

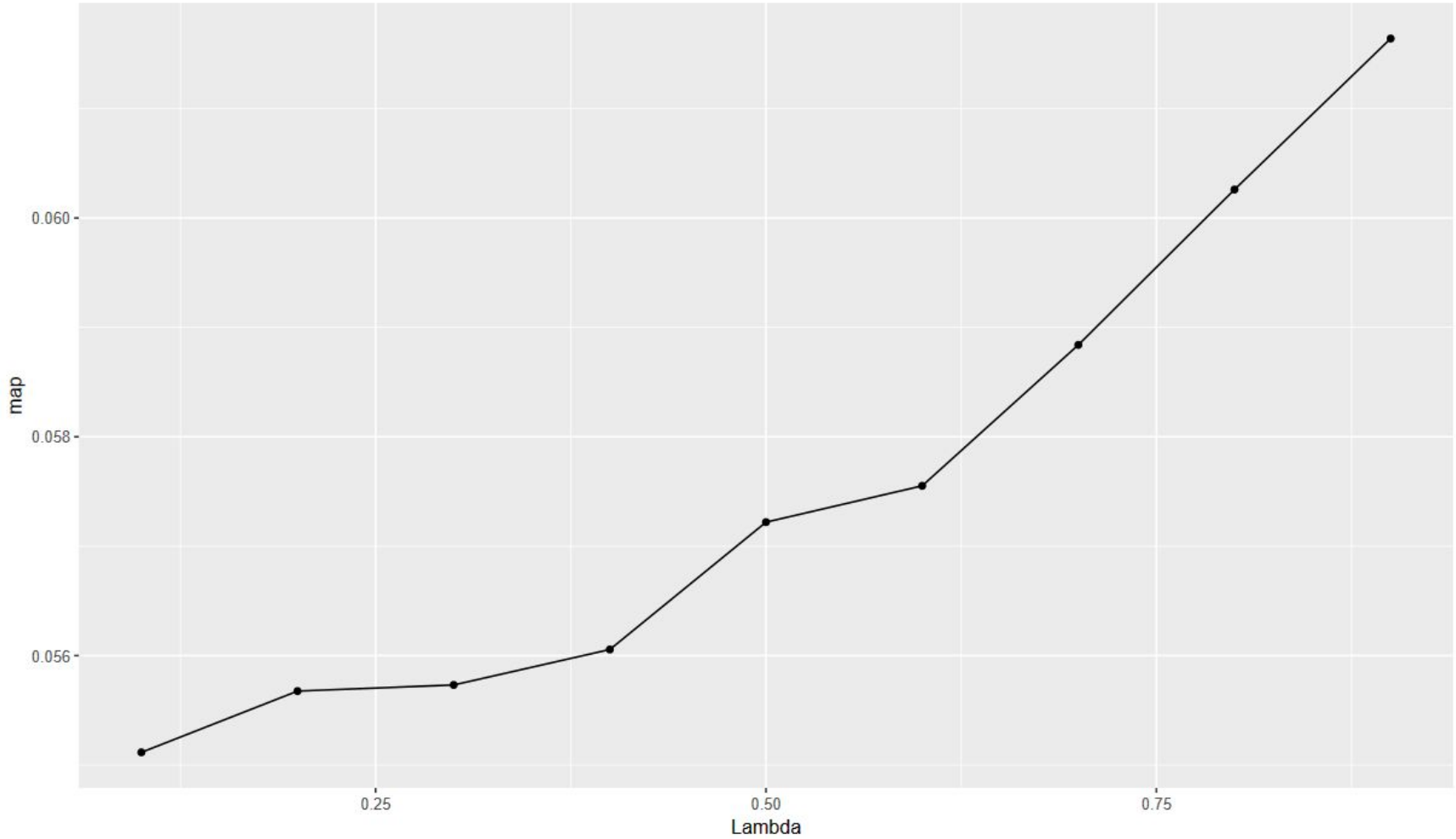


Grafico P5~N

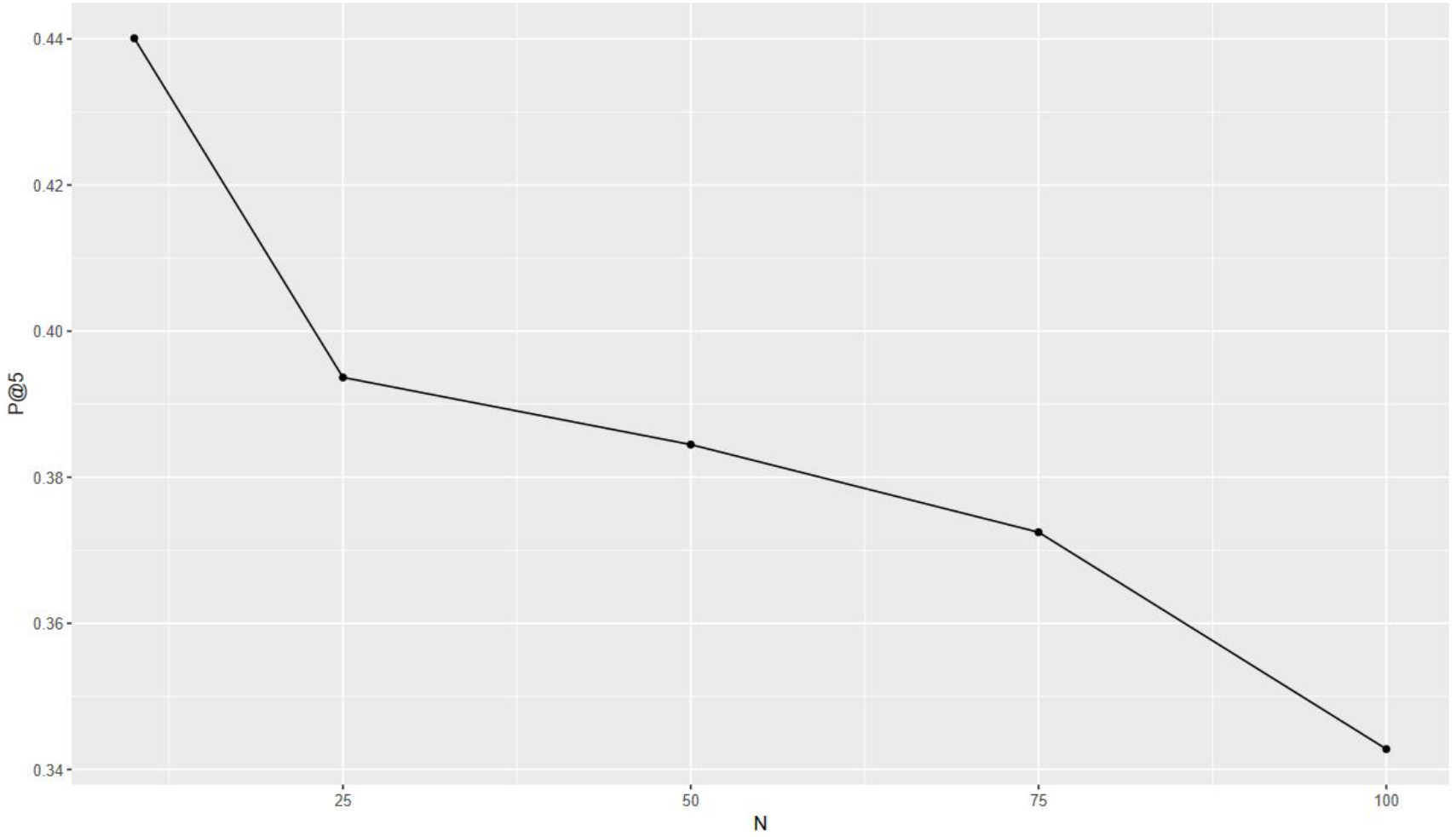


Grafico P5~e

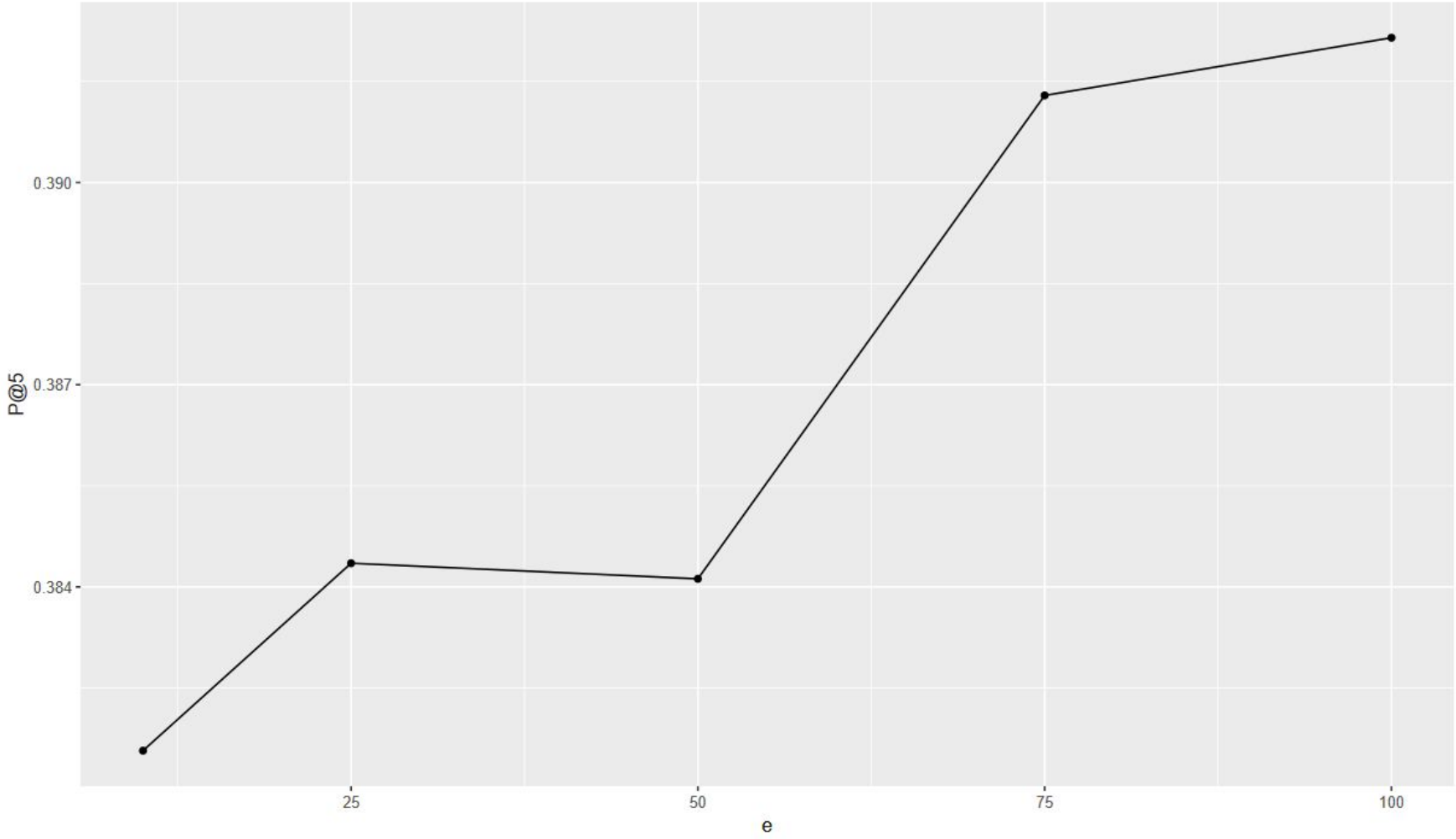
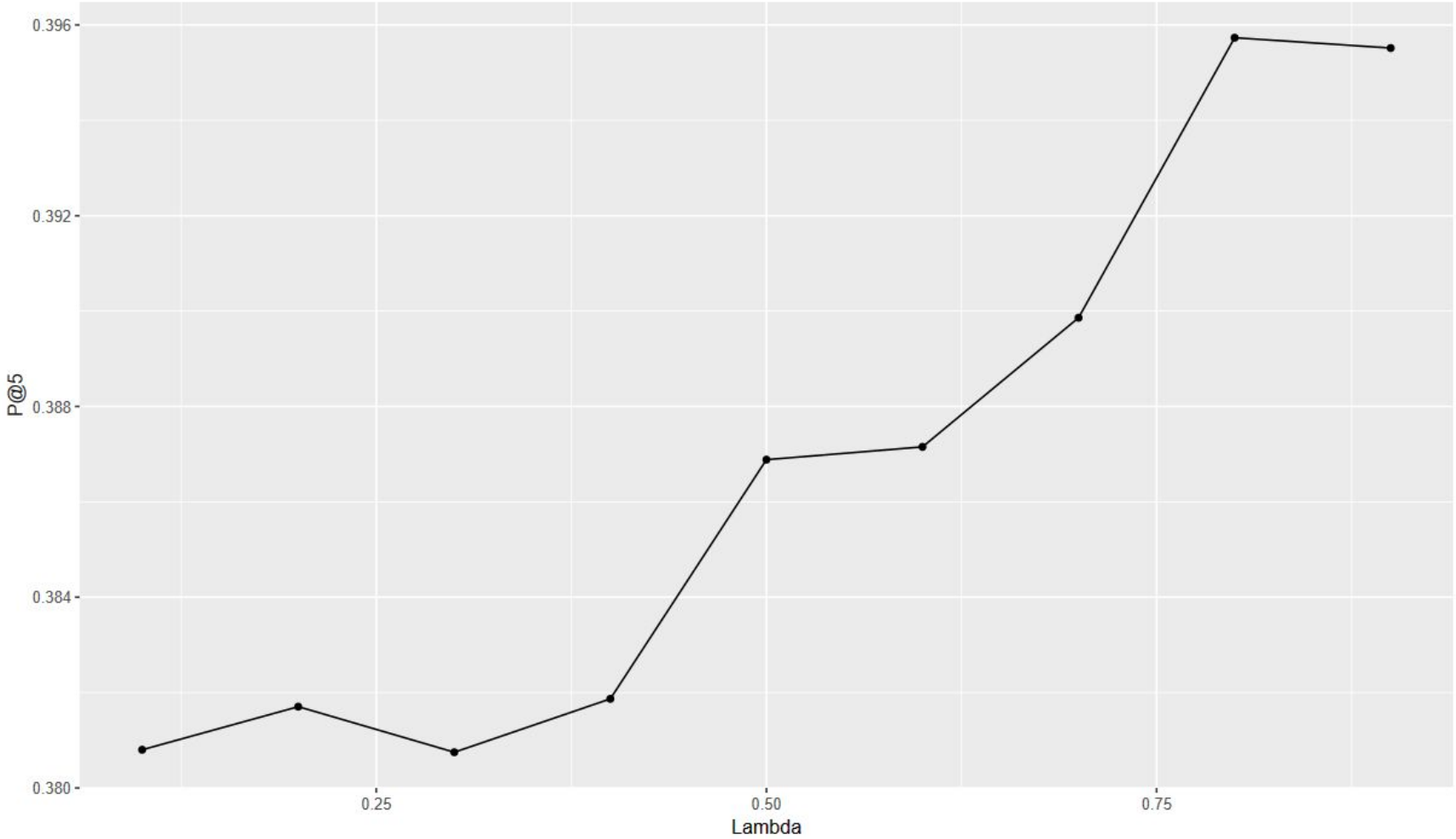


Grafico P5~λ



I modelli

- Abbiamo scelto il modello che ha massimizzato la map nel training set, ossia quello con parametri N=10, e=25 e λ=0.7. Aumenta la map da 0.13 a 0.15.
- Inoltre abbiamo scelto il modello che ha massimizzato la P@5 nel training set, ossia quello con parametri N=10, e=100 e λ=0.9. Aumenta la P@5 da 0.48 a 0.52.

Esempi di applicazione per alcune query I

- Per il modello che massimizza il **map**:
  - la query che migliora di più è la 607
  - la query che peggiora di più è la 626

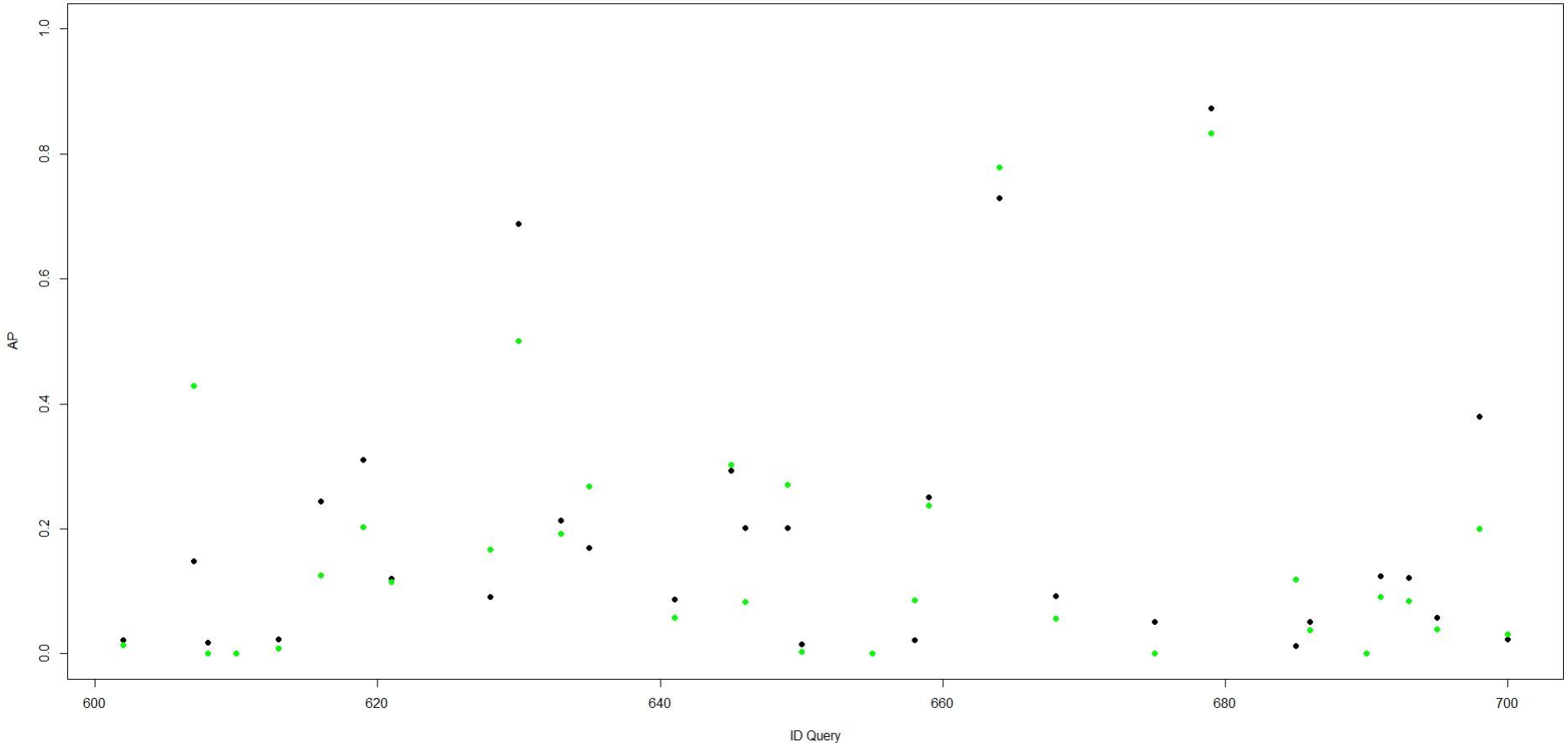
query 607: ‘Human genetic code’

- Primo reperimento:
  - '_id': 'FT923-7735', '_score': 21.507097, NR, rank 1
  - '_id': 'LA062589-0077', '_score': 20.654396, R, rank 2
  - '_id': 'FT934-1290', '_score': 18.350971, R, rank 10
- Aggiunta dei descrittori:
  (**'genetic'**)(**'human'**)(**'gene'**)(**'genome'**)(**'s'**)(**'research'**)(**'genes'**)(**'map'**)(**'dr'**)(**'dna'**)(**'genetics'**) (**'information'**)(**'researchers'**)(**'project'**)(**'scientists'**)
- Reperimento query espansa:
  - '_id': 'FT934-1290', '_score': 62.01763, R, rank 1
  - '_id': 'FT934-2039', '_score': 58.450726, R, rank 2
  - '_id': 'LA062589-0077', '_score': 54.14889, R, rank 3
  - '_id': 'FT923-7735', '_score': 46.06618, NR, rank 6
- Vengono aggiunti descrittori inerenti ai documenti rilevanti. Vengono trovati nuovi documenti rilevanti (FT934-2039).

query 626: ‘Human Stampede’

- Primo reperimento:
  - '_id': 'LA021390-0144', '_score': 14.368825 NR, rank 1
  - '_id': 'FT944-7609', '_score': 14.168889 MR, rank 2
  - '_id': 'LA080389-0083', '_score': 13.022946 NR, rank 8
- Aggiunta dei descrittori: (**'said'**) (**'calgary’**) (**'stampede'**) (**"n't"**) (**'s'**) (**'city'**) (**'years'**)(**'stampeders'**)(**'strandquist'**)(**'wagons'**)(**'one'**)(**'year'**) (**'olympics'**)(**'two'**)(**'rodeo’**)
- Reperimento query espansa:
  - '_id': LA080389-0083', '_score': 46.885597 NR, rank 1
  - '_id': 'FT944-7609', '_score': 15.399456 MR rank 10
  - '_id': 'LA021390-0144' NR rank n.d.
- Vengono aggiunti descrittori non inerenti ai documenti rilevanti. I documenti rilevanti (FT944-7609) vengono spostati alla fine del rank.

Grafico per map ~ query



Esempi di applicazione per alcune query II

- Per il modello che massimizza la **P@5**:
  - la query che migliora di più è la 622
  - la query che peggiora di più è la 626

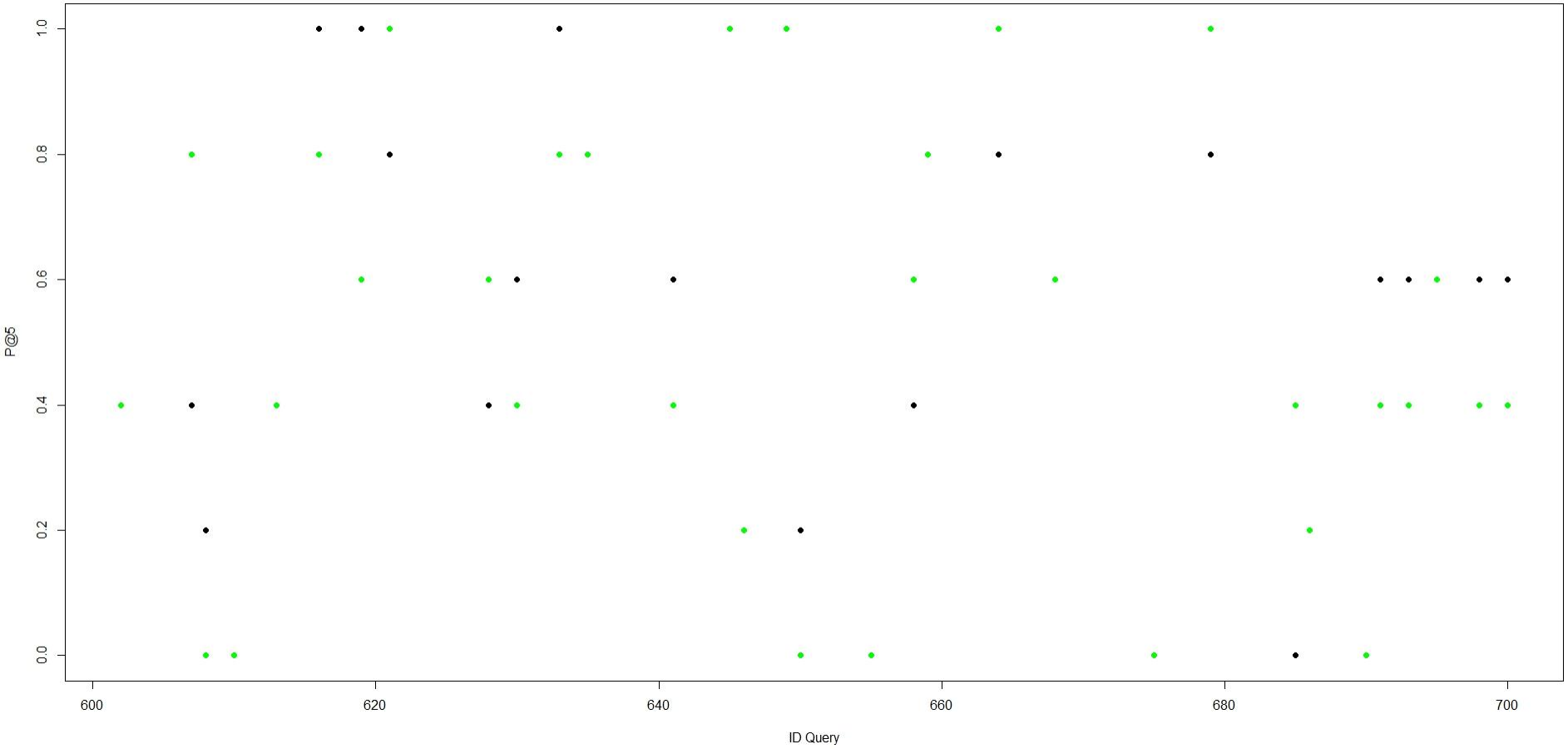
query 622: ‘Price fixing’

- Primo reperimento:
  - '_id': 'FT922-1544', '_score': 16.20115 NR, rank 1
  - '_id': 'FT944-1131', '_score': 15.271421 R, rank 6
  - '_id': 'FT944-12691', '_score': 15.22755 R, rank 7
- Aggiunta dei descrittori: (**'fixing'**)(**'price'**)(**'market'**)(**'companies'**)(**'gold'**) (**'s'**) (**'said'**)(**'shipping'**)(**'london'**)(**'commission'**)(**'eu'**)
- Reperimento query espansa:
  - '_id': 'FT922-1544', '_score': 34.709557 NR, rank 1
  - '_id': 'FT944-1131', '_score': 32.87567 R, rank 2
  - '_id': 'FT944-12691', '_score': 32.056187 R, rank 3
- Vengono aggiunti descrittori inerenti ai documenti rilevanti. I documenti rilevanti vengono spostati all’inizio. Nella prima query il primo documento rilevante era a rank 6 seguito a rank 7.

query 626: ‘Human Stampede’

- Primo reperimento:
  - '_id': 'LA021390-0144', '_score': 14.368825 NR, rank 1
  - '_id': 'FT944-7609', '_score': 14.168889 MR, rank 2
  - '_id': 'LA080389-0083', '_score': 13.022946 NR, rank 8
  - '_id': 'LA072989-0053', '_score': 12.567555 NR, rank 10
- Aggiunta dei descrittori: (**'said'**) (**'calgary’**) (**'stampede'**) (**"n't"**) (**'s'**) (**'city'**) (**'years'**)(**'stampeders'**)(**'strandquist'**)(**'wagons'**)(**'one'**)(**'year'**) (**'olympics'**)(**'two'**)(**'rodeo’**)
- Reperimento query espansa:
  - '_id': 'LA072989-0053', '_score': 45.176605, NR, rank 1
  - '_id': 'LA080389-0083', '_score': 42.75491 NR, rank 2
  - '_id': 'LA021390-0144', '_score': 16.56825 NR, rank 9
  - '_id': 'FT944-7609 MR, rank n.d.
- Simile a quanto accade con la configurazione precedente vengono aggiunti descrittori non inerenti a documenti rilevanti, che vengono spostati alla fine del rank o per nulla reperiti.

Grafico per P@5 ~ query



Paired t-test e Wilcoxon signed-rank test

- Confrontando la configurazione che massimizza il map con il BM25 abbiamo trovato che il miglioramento a livello di AP è significativo (p-value < 0.05)
- Confrontando la configurazione che massimizza il P@5 con il BM25 abbiamo trovato che il miglioramento a livello di P@5 non è significativo (p-value > 0.05)
- In entrambi i casi abbiamo usato un'alternativa unilaterale

Conclusioni

- Cosa si è imparato: un nuovo metodo di clustering, ricercare materiale specifico usando i motori di ricerca.
- Difficoltà emerse: tempi computazionali
- Sviluppi futuri: ricampionamento, language model, scelta della soglia di dominanza e tau di kendall
- Bibliografia:
  - Lee, K. S., Croft, W. B., & Allan, J. (2008, July). *A cluster-based resampling method for pseudo-relevance feedback*.
  - Smucker, M. D., Allan, J., & Carterette, B. (2007, November). *A comparison of statistical significance tests for information retrieval evaluation.*
  - *Massimo Melucci, Information Retrieval Macchine e Motori di ricerca.*