ایندکس، امتیازدهی و جستجوی صفحات

مهدی حسین‌زاده

پروژه اول: اصلاحیه!

- کد زمان ارائه:
  - عدم درج نودهایی که هیچ لینک خروجی نداشتند!
  - تعداد نودها: ~۸۵۵,۰۰۰
  - زمان اجرا: حدود ۶۰ ثانیه
- بعد از اصلاح:
  - تعداد نودها: ~۹۴۶,۰۰۰
  - زمان اجرا: حدود ۸۰ ثانیه
  - Core و In و Out بدون تغییر

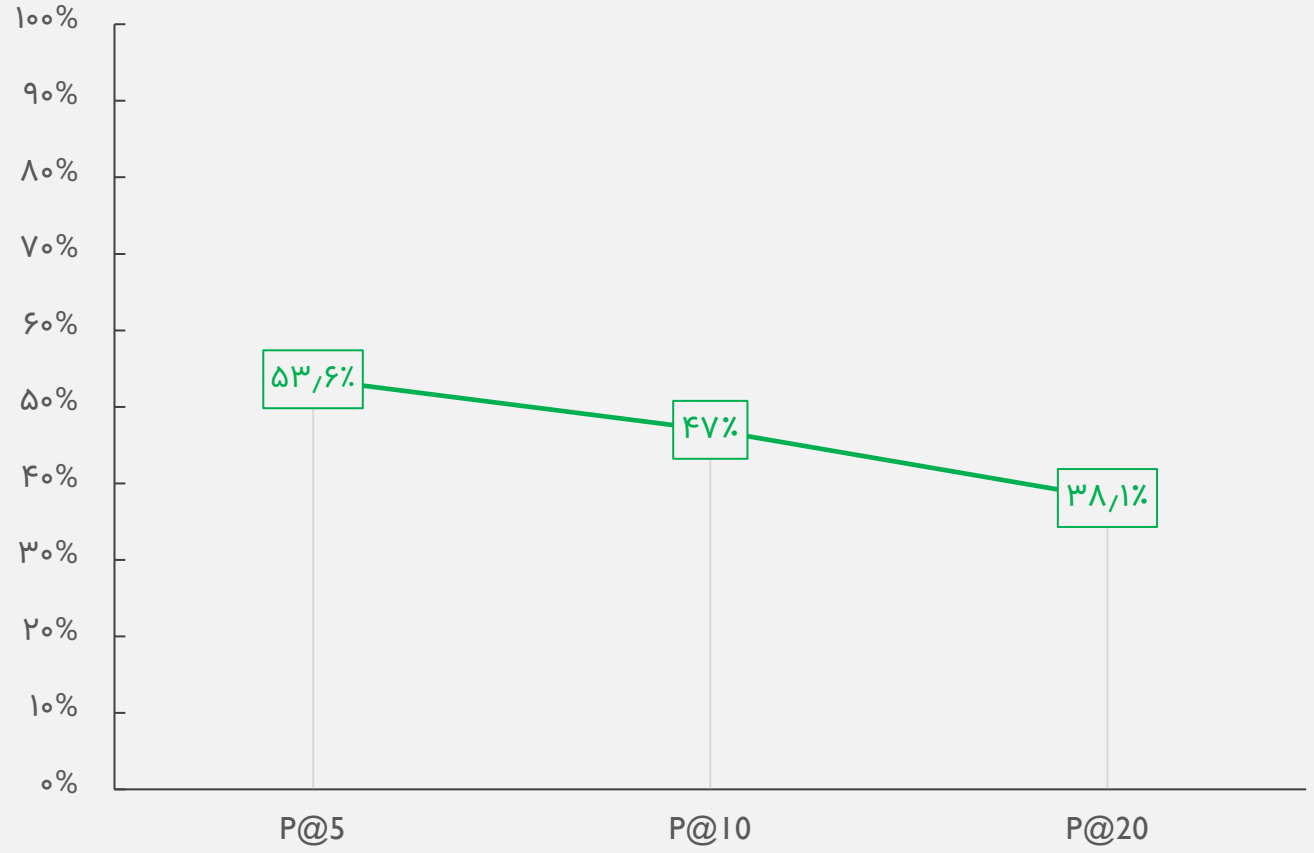
مرحله ۱: ایندکس

ایندکس صفحات

- اندازه ایندکس: ۱٫۶۲ گیگابایت
- زمان اجرا:
  - تک ریسمانی: ۳۰ دقیقه
  - چند ریسمانی: ۱۷ دقیقه
- خواندن داده مستقیما از فایل فشرده
- روش امتیازدهی به اسناد: BM25

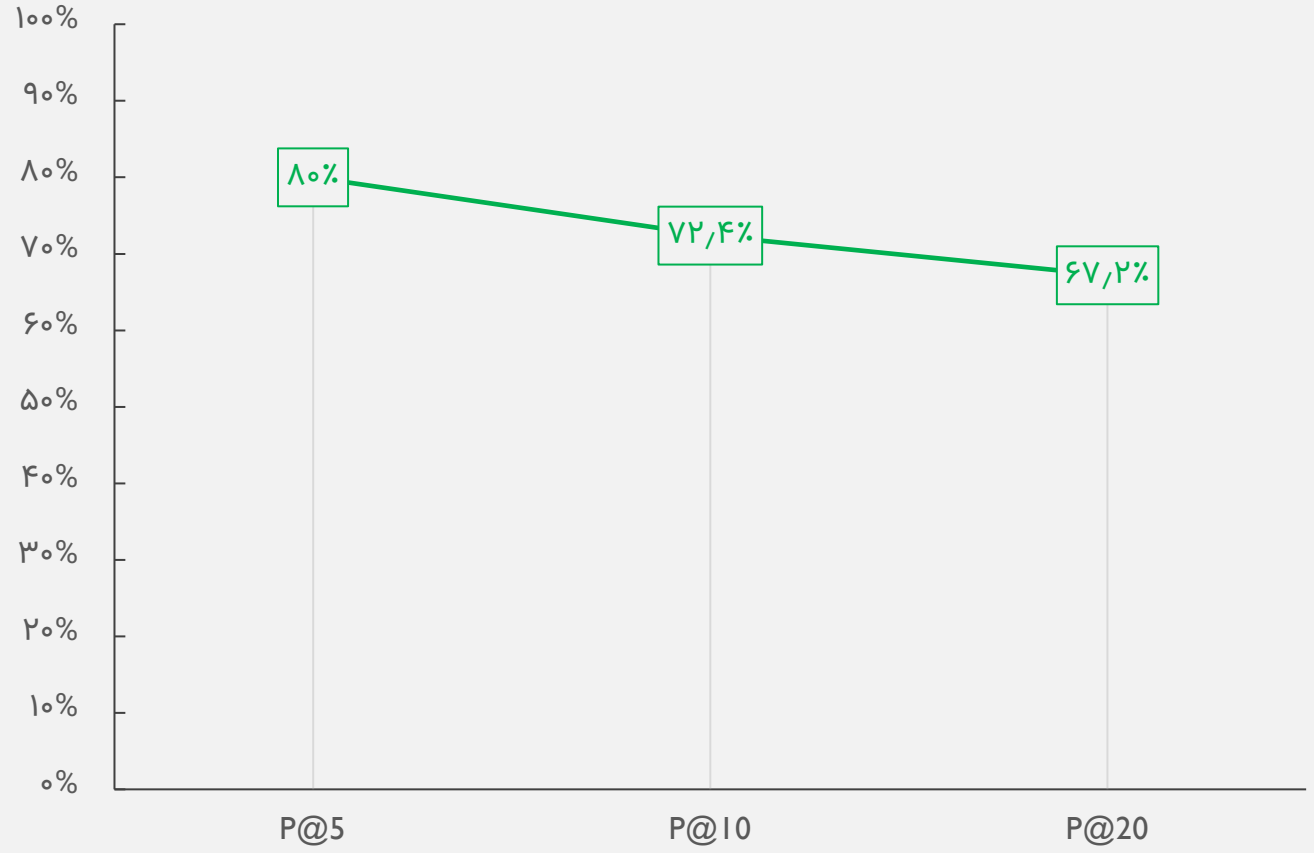
مرحله ۲: جستجو

جستجو با PhraseQuery



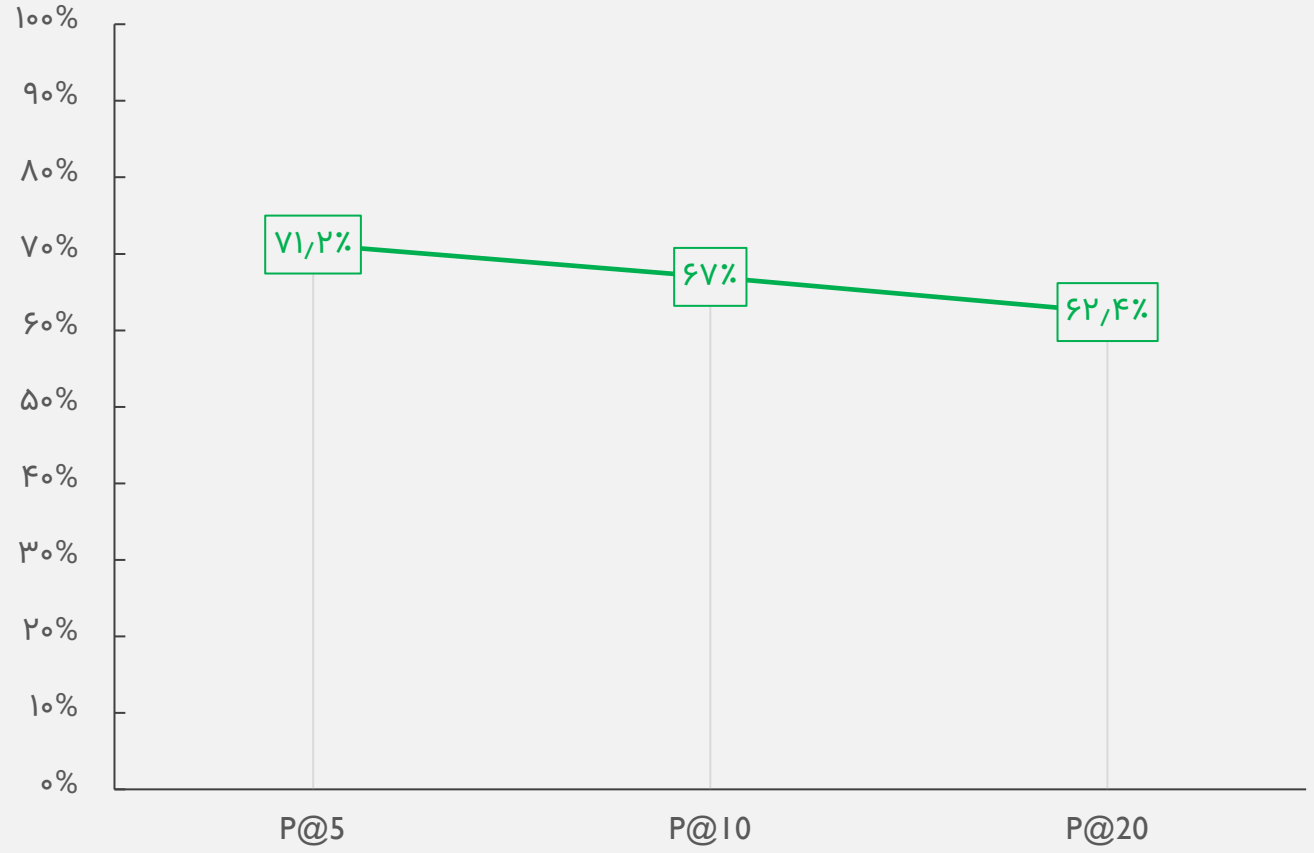
- مقدار slope:
  - ۱۱ برای عنوان
  - ۳۲ برای بدنه
- زمان اجرا: ۱٫۱ ثانیه

جستجو با FuzzyQuery



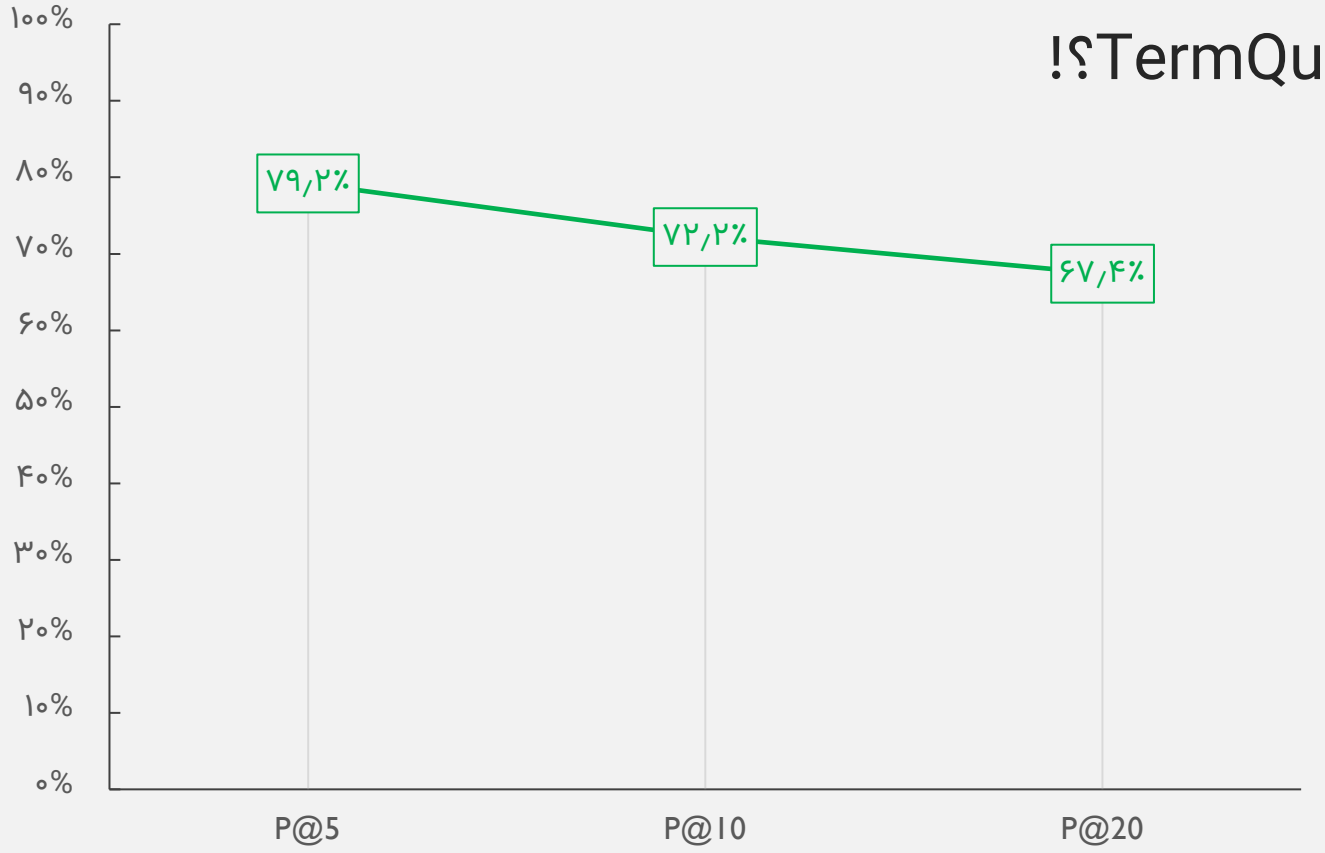
- مقدار boost:
  - ۱ برای عنوان
  - ۶٫۱ برای بدنه
- زمان اجرا: ۵٫۶ ثانیه

جستجو با PhraseQuery + FuzzyQuery



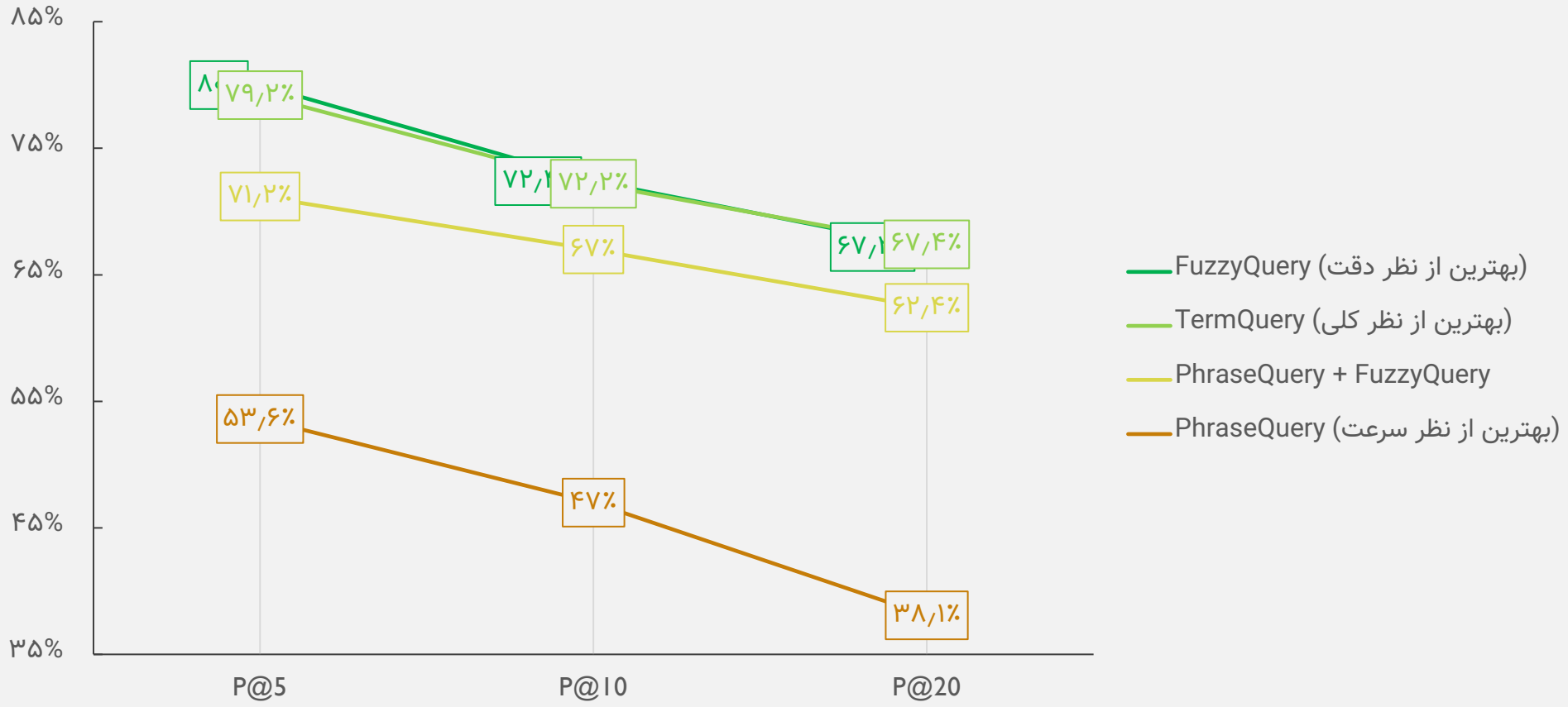
- مقدار boost:
  - ۱ برای Fuzzy
  - ۰٫۳ برای Phrase
- زمان اجرا: ۹٫۵ ثانیه

جستجو با TermQuery



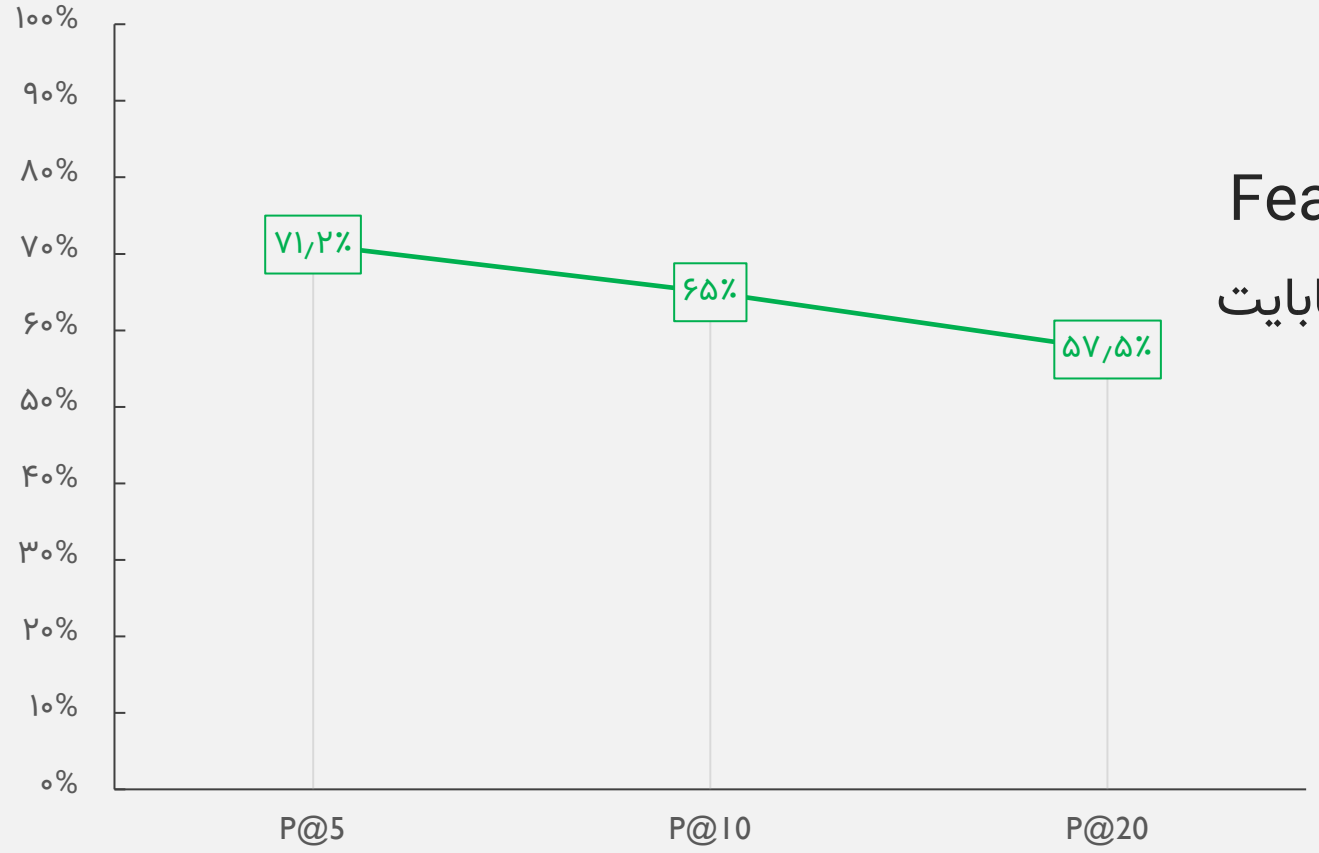
- TermQuery == FuzzyQuery؟!
- مقدار boost:
  - ۱ برای عنوان
  - ۸٫۲ برای بدنه
- زمان اجرا: ۲٫۲ ثانیه

جستجو (همه در یک نگاه)



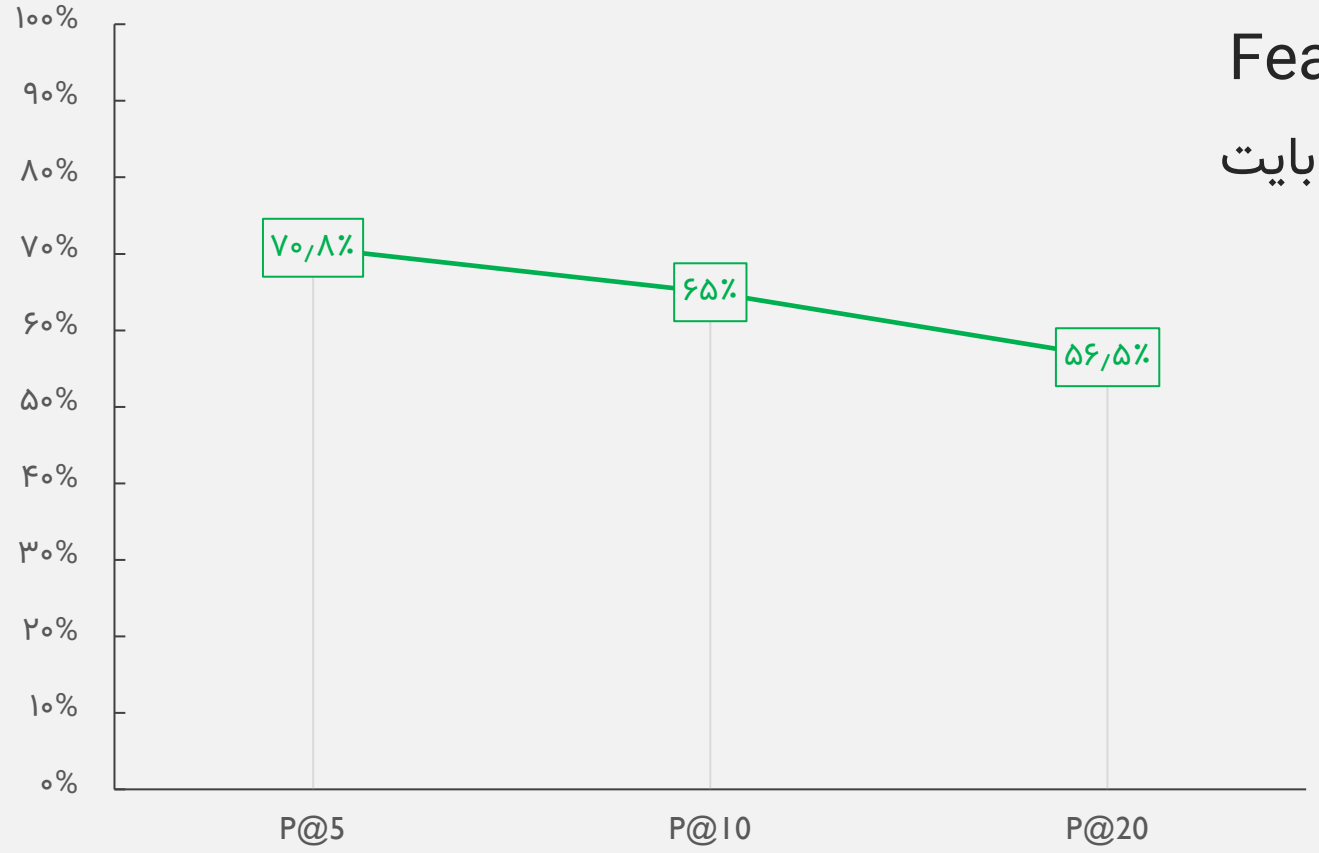
روش‌های دیگر امتیازدهی و رتبه‌بندی

امتیازدهی با PageRank و Lucene



- $\varepsilon = 10^{-10}$
- استفاده از FeatureField
  - حجم ایندکس: ۷۵۲ مگابایت
  - تعداد اسناد: ~۳۷۰,۰۰۰
- زمان اجرا:
  - PageRank: ۱۱ دقیقه
  - جستجو: ۱٫۹ ثانیه

امتیازدهی با HostRank و Lucene



- استفاده از FeatureField
  - حجم ایندکس: ۷۵۰ مگابایت
  - تعداد اسناد: ۳۷۰,۰۰۰~
- زمان اجرا:
  - HostRank: ۴ ساعت
  - جستجو: مثل قبل

با تشکر

Hit 1 from Query 1:
94.57546 = sum of:
 94.487816 = sum of:
  2.9159868 = weight(TITLE: تلفنin 15513) [BM25Similarity], result of:
   2.9159868 = score(freq=1.0), computed as boost * idf * tf from:
    6.8644643 = idf, computed as log(1 + (N - n + 0.5) / (n + 0.5)) from:
     386 = n, number of documents containing term
     370124 = N, total number of documents with field
    0.4247945 = tf, computed as freq / (freq + k1 * (1 - b + b * dl / avgdl)) from:
     1.0 = freq, occurrences of term within document
     1.2 = k1, term saturation parameter
     0.75 = b, length normalization parameter
     8.0 = dl, length of field
     6.8306055 = avgdl, average length of field
  2.7476118 = weight(TITLE: همراهin 15513) [BM25Similarity], result of:
   2.7476118 = score(freq=1.0), computed as boost * idf * tf from:
    6.468096 = idf, computed as log(1 + (N - n + 0.5) / (n + 0.5)) from:
     574 = n, number of documents containing term
     370124 = N, total number of documents with field
    0.4247945 = tf, computed as freq / (freq + k1 * (1 - b + b * dl / avgdl)) from:
     1.0 = freq, occurrences of term within document
     1.2 = k1, term saturation parameter
     0.75 = b, length normalization parameter
     8.0 = dl, length of field
     6.8306055 = avgdl, average length of field
  88.82422 = sum of:
   34.295235 = weight(BODY: قاچاقin 15513) [BM25Similarity], result of:
    34.295235 = score(freq=6.0), computed as boost * idf * tf from:
     8.2 = boost
     4.835779 = idf, computed as log(1 + (N - n + 0.5) / (n + 0.5)) from:
      2937 = n, number of documents containing term
      369938 = N, total number of documents with field
     0.8648752 = tf, computed as freq / (freq + k1 * (1 - b + b * dl / avgdl)) from:
      6.0 = freq, occurrences of term within document
      1.2 = k1, term saturation parameter
      0.75 = b, length normalization parameter
      728.0 = dl, length of field (approximate)
      1027.8987 = avgdl, average length of field
   28.822557 = weight(BODY: گوشيin 15513) [BM25Similarity], result of:
    28.822557 = score(freq=9.0), computed as boost * idf * tf from:
     8.2 = boost
     3.8810537 = idf, computed as log(1 + (N - n + 0.5) / (n + 0.5)) from:
      7631 = n, number of documents containing term
      369938 = N, total number of documents with field
     0.90566796 = tf, computed as freq / (freq + k1 * (1 - b + b * dl / avgdl)) from:
      9.0 = freq, occurrences of term within document
      1.2 = k1, term saturation parameter
      0.75 = b, length normalization parameter
      728.0 = dl, length of field (approximate)
      1027.8987 = avgdl, average length of field
   13.594613 = weight(BODY: تلفنin 15513) [BM25Similarity], result of:
    13.594613 = score(freq=13.0), computed as boost * idf * tf from:
     8.2 = boost
     1.7774278 = idf, computed as log(1 + (N - n + 0.5) / (n + 0.5)) from:
      62546 = n, number of documents containing term
      369938 = N, total number of documents with field
     0.932741 = tf, computed as freq / (freq + k1 * (1 - b + b * dl / avgdl)) from:
      13.0 = freq, occurrences of term within document
      1.2 = k1, term saturation parameter
      0.75 = b, length normalization parameter
      728.0 = dl, length of field (approximate)
      1027.8987 = avgdl, average length of field
   12.111814 = weight(BODY: همراهin 15513) [BM25Similarity], result of:
    12.111814 = score(freq=13.0), computed as boost * idf * tf from:
     8.2 = boost
     1.5835592 = idf, computed as log(1 + (N - n + 0.5) / (n + 0.5)) from:
      75927 = n, number of documents containing term
      369938 = N, total number of documents with field
     0.932741 = tf, computed as freq / (freq + k1 * (1 - b + b * dl / avgdl)) from:
      13.0 = freq, occurrences of term within document
      1.2 = k1, term saturation parameter
      0.75 = b, length normalization parameter
      728.0 = dl, length of field (approximate)
      1027.8987 = avgdl, average length of field
 0.08764976 = Saturation function on the Features field for the PageRank feature, computed as w * S / (S + k) from:
  3.0 = w, weight of this function
  5.5 = k, pivot feature value that would give a score contribution equal to w/2
  0.16552734 = S, feature value