به نام خدا

پروژه ۱ ذخیره و بازیابی اطلاعات

محمد حسین رهبان ۸۱۱۳۷۳۰۷

روش شناسی

برای مقایسه دو روش در بازیابی ابتدا مقادیر دقت و یادآوری را مقایسه می کنیم. در صورت برابر بودن آنها، در هر پرسش جواب های بازیابی شده را مرتب می کنیم و در هر نقطه (در هر سند) مقدار دقت و یادآوری را از ابتدای اسناد تا آن نقطه بدست می آوریم. سپس این مقادیر را به ترتیب در هم ضرب کرده و سپس با هم جمع می کنیم. به این معیار یک معیار ترکیبی دقت و یادآوری می گوییم. در صورت بالاتر بودن این مقدار آن روش بهتر است.

سوال ۱) روش boolean در این مثال هیچ سندی را بازیابی نمی کند.

سوال ۲) اسناد بازیابی شده برای مقادیر مثبت sim به صورت زیر است که در ستون اول شماره سند و در ستون دوم مقدار ranking آمده است

Query 1 Precision = 0.5 Recall = 0.75 36 0.19972074795327802 17 0.17893176333153005 25 0.1252138048441429 29 0.11476734306409303 15 0.09273908119112326 32 0.09173043375685096	Query 2 Precision = 1.0 Recall = 0.53 4 0.09936311221854804 21 0.08666587310866176 9 0.07169112880828901 38 0.06826173203739799 37 0.04788884176725734 36 0.0469108861010324 19 0.037756489941762734
Query 3 Precision = 0.53 Recall = 0.53 34 0.2302149202103528 22 0.2217339483735625 14 0.12065864035045117 35 0.10951887063520688 27 0.10203429905505638 32 0.10064584823802036 8 0.08286168932562035 20 0.0650699981506598	Query 4 Precision = 1.0 Recall = 0.2 7 0.07940031365460147 Query 5 Precision = 1.0 Recall = 0.5 18 0.23291942889672404 32 0.15450100726241803

33 0.0637225042072604 19 0.06020321141611322 24 0.05880445295407482 13 0.04482956251378173 2 0.04385707779346627	

b) از طریق مقادیر داده شده برای Precision و Recall می توان با اسناد داده شده توسط متخصصان مقایسه کرد.

c) روش Boolean سندی را بازیابی نمی کند.

سوال ۳)

Query 1 Precision =0.57 Recall = 1.0 17 5.574053367981417 36 5.574053367981417 5 4.467821994836868 15 4.467821994836868 25 4.467821994836868 29 4.467821994836868 32 4.467821994836868	Query 2 Precision = 1.0 Recall = 0.84 4 8.005923024411096 9 8.005923024411096 19 8.005923024411096 21 8.005923024411096 37 8.005923024411096 38 8.005923024411096 14 4.056604234239254 36 4.056604234239254 1 3.949318790171843 6 3.949318790171843 29 3.949318790171843
Query 3 Precision = 0.57 Recall = 0.61 22 9.885810806808742 14 4.877103901471733 19 4.877103901471733 32 4.877103901471733 8 4.467821994836868 13 4.467821994836868 24 4.467821994836868 27 4.467821994836868 35 4.467821994836868 2 4.311757438827326 11 4.311757438827326	Query 4 Precision = 0.8 Recall = 0.8 7 4.467821994836868 13 4.467821994836868 27 4.467821994836868 30 4.467821994836868 35 4.467821994836868 Precision = 1.0 Recall = 1.0 18 10.043792825541333 8 5.1666889240696 10 4.877103901471733

20 4.311757438827326 33 4.311757438827326 34 4.311757438827326	32 4.877103901471733

b) از طریق Precision و Recall ارائه شده است.

c) مقادیر Precision تقریبا برابر ولی Recall در روش احتمالی بیشتر از روش برداری است. در جدول های زیر دو روش توسط معیارهای دقت و یادآوری مقایسه شده اند.

	Vector	Probabilistic
1	0.5	0.57
2	1.0	1.0
3	0.53	0.57
4	0.5	0.8
5	1.0	1.0

جدول ۱ دقت دو روش احتمالی و برداری

	Vector	Probabilistic
1	0.75	1.0
2	0.53	0.84
3	0.53	0.61
4	0.2	0.8
5	0.5	1.0

جدول ۲ یادآوری دو روش احتمالی و برداری

با استفاده از روش شناسی ارایه شده مقدار معیار برای دو روش به ازاء پرسش های متفاوت به صورت زیر است :

	Vector	Probabilistic
1	۱،۷۸	۳،۰۵
2	۲،۱۵	۵،۰۷
3	۱،۶	۱،۶۵
4	۰،۳	۱،۶۵
5	۰،۷۵	۲،۵

جدول ۳ معیار ترکیبی یادآوری و دقت برای دو روش احتمالی و برداری

ملاحظه می شود که روش احتمالی نتایج به مراتب بهتری نسبت به روش برداری بر روی ۵ پرسش ارایه می دهد.

سوال ۴)

Query 1 Precision = 0.57 Recall = 1.0 5 7.740229524763182 15 7.740229524763182 25 7.740229524763182 29 7.740229524763182 32 7.740229524763182 17 5.485435095605484 36 5.485435095605484	Query 2 Precision = 1.0 Recall = 0.84 4 16.110478678192298 9 16.110478678192298 19 16.110478678192298 21 16.110478678192298 37 16.110478678192298 38 16.110478678192298 1 8.279088274961868 6 8.279088274961868 29 8.279088274961868 14 7.83139040323043 36 7.83139040323043
Query 3 Precision = 0.57 Recall = 0.6۱ 22 11.411080404103021 2 6.67002049742228 11 6.67002049742228 20 6.67002049742228 33 6.67002049742228 34 6.67002049742228 8 6.39174077764889 13 6.39174077764889 24 6.39174077764889 27 6.39174077764889 35 6.39174077764889 14 5.748700417143124 19 5.748700417143124 32 5.748700417143124	Query 4 Precision = 0.8 Recall = 0.8 7 9.353487881194667 13 9.353487881194667 27 9.353487881194667 30 9.353487881194667 35 9.353487881194667 Query 5 Precision = 1.0 Recall = 1.0 18 14.762292602141105 10 7.804795231264155 32 7.804795231264155 8 6.957497370876951

b) برای دقت و یادآوری در حالت های تک تکرار و چند بار تکرار دیده می شود که به ازاء همه پرسش ها این مقادیر برابرند. به ازاء مقدار ترکیبی دقت و یادآوری داریم :

	Single	Multiple
1	3.05	3.45
2	5.07	۵،۰۷
3	1.65	5.10
4	1.65	۱،۶۵
5	2.5	۲،۵

دیده می شود بر اساس این معیار ترکیبی روش چند تکراره در مورد پرسش های ۱ و ۳ به میزان قابل توجهی بهتر می شود.

سوال ۵)

Query 1 Precision = 0,57 Recall = 1.0 5 6.19154415882364 15 6.19154415882364 25 6.19154415882364 29 6.19154415882364 32 6.19154415882364 36 6.110199510489747 17 3.659755742391968	Query 2 Precision = 1.0 Recall = 0.84 4 16.110478678192298 9 16.110478678192298 19 16.110478678192298 21 16.110478678192298 37 16.110478678192298 38 16.110478678192298 1 8.279088274961868 6 8.279088274961868 29 8.279088274961868 14 7.83139040323043 36 7.83139040323043
Query 3 Precision = 0.57 Recall = 0.61 22 13.24578644259488 2 7.905367900028208 11 7.905367900028208 20 7.905367900028208 33 7.905367900028208	Query 4 Precision = 0.8 Recall = 0.8 7 8.054204897064407 13 8.054204897064407 27 8.054204897064407 30 8.054204897064407 35 8.054204897064407

34 7.905367900028208 8 4.042665614146777 13 4.042665614146777 24 4.042665614146777 27 4.042665614146777 35 4.042665614146777 14 2.164963715117998 19 2.164963715117998 32 2.164963715117998	Query 5 Precision = 1.0 Recall = 1.0 18 14.762292602141105 10 7.804795231264155 32 7.804795231264155 8 6.957497370876951

b) در این مورد هم مقادیر دقت و یادآوری به ازاء تمام پرسش ها در دو روش یکسان هستند.

مقدار ترکیبی یادآوری و دقت به صورت زیر است :

	Multiple	Selective
1	۳،۴۵	3.75
2	5.07	۵،۰۷
3	۵،۱۰	5.10
4	۱،۶۵	۱،۶۵
5	2.5	۲،۵

دیده می شود که در مورد پرسش اول مقدار معیار برای روش جدید بهتر می شود

.

سوال ۶ و ۷) این کار از طریق اضافه کردن tf * idf به جای w_{ij} در فرمول مربوط به sim امکانپذیر است. با این وصف مقادیر زیر بدست می آیند

Query 1 Precision = 0.57 Recall = 1.0 17 28.52426249714852 36 28.52426249714852 5 22.446665621813224 15 22.446665621813224 25 22.446665621813224 29 22.446665621813224 32 22.446665621813224	Query 2 Precision = 1.0 Recall = 0.84 4 34.615144264526876 9 34.615144264526876 19 34.615144264526876 21 34.615144264526876 37 34.615144264526876 38 34.615144264526876 1 17.386085377419924

	6 17.386085377419924 29 17.386085377419924 14 17.229058887106948 36 17.229058887106948
Query 3 Precision = 0.57 Recall = 0.61 22 42.66256685778001 14 21.27019154342956 19 21.27019154342956 32 21.27019154342956 8 18.53604825518178 13 18.53604825518178 24 18.53604825518178 27 18.53604825518178 35 18.53604825518178 2 18.009055343040156 11 18.009055343040156 20 18.009055343040156 33 18.009055343040156 34 18.009055343040156	Query 4 Precision = 0.8 Recall = 0.8 7 27.125114855464535 13 27.125114855464535 27 27.125114855464535 30 27.125114855464535 35 27.125114855464535 Query 5 Precision = 1.0 Recall = 1.0 18 58.09923131336057 8 29.221488957683196 10 28.877742355677373 32 28.877742355677373

b) مقادیر هیچ کدام از معیارها تغییر نمی کنند.

سوال ۹ و ۱۰) این کار از طریق وزن دهی به پاسخ های دو روش بدست می آید. در این حالت وزن متناسب با دقت (میانگین Precision و Recall) هر کدام از روش ها روی تعداد نمونه ای از Query ها است. بر این اساس وزن مربوط به روش برداری ۰،۴۵ و روش احتمالی ۰،۵۵ است. نتایج در زیر آمده است :

Query 1 36 1.000446996 17 0.953601512 25 0.723321635 29 0.699781786 15 0.65014374 32 0.647870874 5 0.441167342	Query 2 4 1.001713928 21 0.943999205 9 0.875932186 38 0.860344018 37 0.767739972 19 0.721683827 36 0.491948643 1 0.271346076

	6 0.271346076 29 0.271346076 14 0.278717343
Query 3 22 0.984150766 34 0.690447479 14 0.507569952 32 0.468414489 35 0.462990838 27 0.448347111 8 0.410835484 19 0.389287591 20 0.367337849 33 0.364701448 24 0.363766978 13 0.336424801 2 0.325834309 11 0.240026983	Query 4 7 1.000964596 13 0.550964596 27 0.550964596 30 0.550964596 35 0.550964596 Query 5 18 1.0059197 32 0.569456606 8 0.283035748 10 0.267172026

b) در مورد یادآوری و دقت، هر دو به حد دقت و یادآوری روش احتمالی می رسند. در مورد معیار ترکیبی داریم :

	Single	Combination
1	3.05	2.35
2	5.07	۵،۰۷
3	1.65	1.85
4	1.65	۱،۶۵
5	2.5	۲،۵

دیده می شود که در مورد معیار ترکیبی روش احتمالی بهتر از ترکیب دو روش احتمالی و برداری به شکل بالا است. دلیل آن هم این است که روش احتمالی اکیدا بهتر از روش برداری است.

c) یکی هیچ یک از دو روش اکیدا بهتر از دیگری نباشد، ریسک جواب اشتباه کمتر می شود و به صورت متوسط میزان معیارها افزایش می یابد.