

学号：201600130077

姓名：于海洋

班级：16 人工智能

实验题目：Pivoted Length Normalization VSM and BM25

实验内容：

1、根据实验三搜索结果，对于返回的结果应用 BM25 和 VSM 的方法进行打分，然后按照打分结果从高到低输出，

• Pivoted Length Normalization VSM [Singhal et al 96]

$$f(q, d) = \sum_{w \in q \cap d} c(w, q) \frac{\ln[1 + \ln[1 + c(w, d)]]}{1 - b + b \frac{|d|}{avdl}} \log \frac{M + 1}{df(w)}$$

• BM25/Okapi [Robertson & Walker 94]

$$b \in [0, 1]$$

$$k_1, k_3 \in [0, +\infty)$$

$$f(q, d) = \sum_{w \in q \cap d} c(w, q) \frac{(k + 1)c(w, d)}{c(w, d) + k(1 - b + b \frac{|d|}{avdl})} \log \frac{M + 1}{df(w)}$$

```
ans += cwq[q_word]*(math.log(1+math.log(1+cwd[q_word]))/(1-b+b*(tweet_l/avdl)))*(math.log((M+1)/DF[q_word]))
# vsm_f
#_ans += cwq[q_word]*((k+1)*cwd[q_word]/(cwd[q_word]+k*(1-b+b*(tweet_l/avdl)))*(math.log((M+1)/DF[q_word]))
# bm
```

2、使用 sorted 对结果排序

```
f[q_id][rst_list[i]] = ans
sf[q_id] = sorted(f[q_id].items(), key=lambda d: d[1], reverse=True)
```

结论分析与体会：

avdl 和 K 的值会对结果有一定影响，自己调整 avdl 要比直接计算出来的结果要好很多

VSM 结果 (avdl)

query: 171 ,AP: 0.9096374618100043

query: 172 ,AP: 0.3260430853069655

query: 173 ,AP: 0.3287159659985747

query: 174 ,AP: 0.8940493369064798

query: 175 ,AP: 0.377313185608745

query: 176 ,AP: 0.7611535789844033

query: 177 ,AP: 0.7182678101717503

query: 178 ,AP: 0.45376948475084283

query: 179 ,AP: 0.5427909576475564

query: 180 ,AP: 0.16560911685278426

query: 181 ,AP: 0.8221891534391534

query: 182 ,AP: 0.19305019305019305

query: 183 ,AP: 0.33897496313836917

query: 184 ,AP: 0.481049627613706

query: 185 ,AP: 0.5238095238095238

query: 186 not found a true value

query: 187 ,AP: 0.9384193584193584

query: 188 ,AP: 0.16471204959256427

query: 189 ,AP: 0.07165096807953951

query: 190 ,AP: 0.1806121067234472

query: 191 ,AP: 0.5876782705149423

query: 192 ,AP: 0.6892346509671994

query: 193 ,AP: 0.288113458898076

query: 194 ,AP: 0.9001716678574264

query: 195 ,AP: 0.21273775354794625

query: 196 ,AP: 0.6603302130348397

query: 197 ,AP: 0.8154572619783833

query: 198 ,AP: 0.5577994941116361

query: 199 ,AP: 0.22028491439799652

query: 200 ,AP: 0.3838807603748176

query: 201 ,AP: 0.37037037037037035

query: 202 ,AP: 0.658106059430807

query: 203 ,AP: 0.0738669043914009

query: 204 ,AP: 0.29728835978835977

query: 205 ,AP: 0.6060606060606061

query: 206 ,AP: 0.7596718757397245

query: 207 ,AP: 0.48776983495361226

query: 208 ,AP: 0.303951367781155

query: 209 ,AP: 0.11371677980717465

query: 210 ,AP: 0.7035339092303

query: 211 ,AP: 0.33154247547451343

query: 212 ,AP: 0.5794708869203072

query: 213 ,AP: 0.385561219540435

query: 214 ,AP: 0.704225352112676

query: 215 ,AP: 0.30120481927710846

query: 216 ,AP: 0.4294331321386912

query: 217 ,AP: 0.36946403339739936

query: 218 ,AP: 0.23507798452776443

query: 219 ,AP: 0.21349164526094372

query: 220 ,AP: 0.3131469610636277

query: 221 ,AP: 0.1988071570576541

query: 222 ,AP: 0.29277326427120615

query: 223 ,AP: 0.6126446841487477

query: 224 ,AP: 0.6231379731379731

query: 225 ,AP: 0.6312502040912813

MAP = 0.456419531519292

query 171 , NDCG: 0.8941743522307255

query 172 , NDCG: 0.935786845136712

query 173 , NDCG: 0.4777347408965493

query 174 , NDCG: 0.8734287539144301

query 175 , NDCG: 0.8109575757340682

query 176 , NDCG: 0.7442990130516465

query 177 , NDCG: 0.8323707384302289

query 178 , NDCG: 0.7857269759642064

query 179 , NDCG: 0.48571219288636047

query 180 , NDCG: 0.5632983133925671

query 181 , NDCG: 0.8645285146205596

query 182 , NDCG: 0.7343196189892779

query 183 , NDCG: 0.842218110904324

query 184 , NDCG: 0.6036015078685814

query 185 , NDCG: 0.7249654012174153

query 186 , NDCG: 0.0

query 187 , NDCG: 0.8050882707005312

query 188 , NDCG: 0.36365849995336663

query 189 , NDCG: 0.18449638839084828

query 190 , NDCG: 0.36897162758231616

query 191 , NDCG: 0.7120351225179544

query 192 , NDCG: 0.7807532715932586

query 193 , NDCG: 0.31711587460375007

query 194 , NDCG: 0.8737558294046541

query 195 , NDCG: 0.5027984697956249

query 196 , NDCG: 0.7385015778686099

query 197 , NDCG: 0.8246804065350793

query 198 , NDCG: 0.5521197003404978

query 199 , NDCG: 0.8431785818414863

query 200 , NDCG: 0.8071629800313015

query 201 , NDCG: 0.8131579214456046

query 202 , NDCG: 0.8701752187419105

query 203 , NDCG: 0.15617313593718807

query 204 , NDCG: 0.5019280144358281

query 205 , NDCG: 0.8807404097725423

query 206 , NDCG: 0.8099199455508068

query 207 , NDCG: 0.6274973260606325

query 208 , NDCG: 0.6967867854018366

query 209 , NDCG: 0.563440974745184

query 210 , NDCG: 0.79017794138468

query 211 , NDCG: 0.5230881910444493

query 212 , NDCG: 0.8734020961891296

query 213 , NDCG: 0.9717534384725915

query 214 , NDCG: 0.8375036211749237

query 215 , NDCG: 0.5792156253407705

query 216 , NDCG: 0.6567225210888656

query 217 , NDCG: 0.5635198698215564

query 218 , NDCG: 0.4528887728438863

query 219 , NDCG: 0.34892721208527516

query 220 , NDCG: 0.2946313133687763

query 221 , NDCG: 0.8870788860187027

query 222 , NDCG: 0.4298470879149592

query 223 , NDCG: 0.6431026151016132

query 224 , NDCG: 0.6637622952800909

query 225 , NDCG: 0.4674228391459435

NDCG = 0.6500055148139032

BM25 结果

query: 171 ,AP: 0.9084992932562858

query: 172 ,AP: 0.31801848327324184

query: 173 ,AP: 0.32686411414672284

query: 174 ,AP: 0.6947029912443446

query: 175 ,AP: 0.3706984117660653

query: 176 ,AP: 0.7600540699678563

query: 177 ,AP: 0.7247465952004165

query: 178 ,AP: 0.4538913171095173

query: 179 ,AP: 0.5272314018753559

query: 180 ,AP: 0.15436778680140975

query: 181 ,AP: 0.8221891534391534

query: 182 ,AP: 0.19305019305019305

query: 183 ,AP: 0.34149600233210253

query: 184 ,AP: 0.48357218388719647

query: 185 ,AP: 0.49559082892416223

query: 186 not found a true value

query: 187 ,AP: 0.8221763438324082

query: 188 ,AP: 0.1419449474136974

query: 189 ,AP: 0.05379382522239665

query: 190 ,AP: 0.1687786768567067

query: 191 ,AP: 0.5898959800992962

query: 192 ,AP: 0.6892346509671994

query: 193 ,AP: 0.288113458898076

query: 194 ,AP: 0.8832473340213279

query: 195 ,AP: 0.2128152696545958

query: 196 ,AP: 0.7035851054906109

query: 197 ,AP: 0.8154572619783833

query: 198 ,AP: 0.5573429765897979

query: 199 ,AP: 0.21961552643076696

query: 200 ,AP: 0.3838807603748176

query: 201 ,AP: 0.37037037037037035

query: 202 ,AP: 0.6607540916706761

query: 203 ,AP: 0.0738669043914009

query: 204 ,AP: 0.30411986161986165

query: 205 ,AP: 0.6060606060606061

query: 206 ,AP: 0.7647414907925407

query: 207 ,AP: 0.4868567633508855

query: 208 ,AP: 0.303951367781155

query: 209 ,AP: 0.15746350832731196

query: 210 ,AP: 0.7083934907187595

query: 211 ,AP: 0.3347632609151264

query: 212 ,AP: 0.5854285355656277

query: 213 ,AP: 0.3763631455727735

query: 214 ,AP: 0.704225352112676

query: 215 ,AP: 0.29292823796687095

query: 216 ,AP: 0.4294331321386912

query: 217 ,AP: 0.36927930433336

query: 218 ,AP: 0.24107526178704933

query: 219 ,AP: 0.12269833895039606

query: 220 ,AP: 0.20554114095780762

query: 221 ,AP: 0.1988071570576541

query: 222 ,AP: 0.2587536810009911

query: 223 ,AP: 0.5683937801080045

query: 224 ,AP: 0.6231379731379731

query: 225 ,AP: 0.6068000750956976

MAP = 0.4447102141070613

query 171 , NDCG: 0.8938922261898875

query 172 , NDCG: 0.9234044232997829

query 173 , NDCG: 0.47629889875338116

query 174 , NDCG: 0.6991606078747173

query 175 , NDCG: 0.8055113497178705

query 176 , NDCG: 0.7438266334232754

query 177 , NDCG: 0.8397050229422184

query 178 , NDCG: 0.7820582091440708

query 179 , NDCG: 0.48133355981363224

query 180 , NDCG: 0.5152327288210554

query 181 , NDCG: 0.8645285146205596

query 182 , NDCG: 0.7373687821891383

query 183 , NDCG: 0.843335874966871

query 184 , NDCG: 0.600892240920363

query 185 , NDCG: 0.6830319872732481

query 186 , NDCG: 0.0

query 187 , NDCG: 0.7731322821097363

query 188 , NDCG: 0.3484444100583365

query 189 , NDCG: 0.16234712260733572

query 190 , NDCG: 0.3623977964697284

query 191 , NDCG: 0.7124528651772082

query 192 , NDCG: 0.8189835112988078

query 193 , NDCG: 0.31711587460375007

query 194 , NDCG: 0.8307408281217493

query 195 , NDCG: 0.5111270605333144

query 196 , NDCG: 0.7816845199978009

query 197 , NDCG: 0.8254904463482857

query 198 , NDCG: 0.5645453058495407

query 199 , NDCG: 0.8648861834867744

query 200 , NDCG: 0.8156184444854139

query 201 , NDCG: 0.8164835706659821

query 202 , NDCG: 0.8720408971582932

query 203 , NDCG: 0.15617313593718807

query 204 , NDCG: 0.5048430720281382

query 205 , NDCG: 0.879457870934935

query 206 , NDCG: 0.8108849520430411

query 207 , NDCG: 0.6270613655051587

query 208 , NDCG: 0.689524421607916

query 209 , NDCG: 0.7061973105278798

query 210 , NDCG: 0.7923718802495426

query 211 , NDCG: 0.5253671257759734

query 212 , NDCG: 0.8758784279370232

query 213 , NDCG: 0.957328754688952

query 214 , NDCG: 0.8362573244827222

query 215 , NDCG: 0.5932702568462348

query 216 , NDCG: 0.6616048781592565

query 217 , NDCG: 0.567332210032747

query 218 , NDCG: 0.46382023613453205

query 219 , NDCG: 0.24235203849824408

query 220 , NDCG: 0.26020414204132447

query 221 , NDCG: 0.8977149890466922

query 222 , NDCG: 0.47308664275608014

query 223 , NDCG: 0.6256051972473456

query 224 , NDCG: 0.6637622952800909

query 225 , NDCG: 0.4534313489123576

NDCG = 0.6460109464653729