信息检索实验四 VSM & BM25

实验硬件 & 软件

i5-6200U + python3.7 + bs4

实验内容

本次实验要求在实验3的基础上对查询结果进行相关度打分排序。实现VSM & BM25.

实验步骤

- 首先看了看能不能复用 hw3 的代码，后来放弃了
- 然后看了看需不需要 nlp 之类的 tokenization，好像也不太需要 正则就够了
- 然后，给的查询文本是 query171-225.txt 是xml的，转换成json并转小写 trans_query.py -> query171-225_cleaned.txt
- get_solve.py

读取tweets，建立posting_list，两个相关函数计算

```
def VSM_F(c_w_q,c_w_d,tweet_l,df_w,b=0.5,avdl=20,M=len(tweets)):
    return c_w_q*(log(1+log(1+c_w_d))/(1-b+b*(tweet_l/avdl)))*(log((M+1)/df_w))
def BM25(c_w_q,c_w_d,tweet_l,df_w,k=8,b=0.5,avdl=20,M=len(tweets)):
    return c_w_q*((k+1)*c_w_d/(c_w_d+k*(1-b+b*(tweet_l/avdl))))*(log((M+1)/df_w))
```

因为同一个查询与同一篇文章的 `f(q,d)` 是多个单词的累加，所以上述函数是：

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

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

可能会遇到，查询的是单词的一部分，这时候有两种情况：

一、用户所输入的单词是目标单词的一部分，且我们的posting—list没有用户输入的单词

二、用户所输入的单词是目标单词的一部分，但是我们的posting—list有这个单词

我们假设用户都是"聪明"人，不会有类似于 输入git ，找github 的情况

只是当用户输入的word不在posting—list中时才查找所有的posting—list

avdl 第一次没有用“doc长度的平均值”，而是直接用的一个20常数。

后来尝试算平均值再比较。发现使用 avdl=20 算出来效果还比“doc长度平均值还行”，详见 BM25_envalue_result.txt。 大概比真正的avdl高出 2% 左右

两个方法的评估结果分别存在：

```
VSM_envalue_result.txt
BM25_envalue_result.txt
```

贴一下BM25的成绩：

```
avdl = 20

query: 171 ,AP: 0.9841991063308005
query: 172 ,AP: 0.33340222599888975
query: 173 ,AP: 0.31567543366767686
query: 174 ,AP: 0.7837623025123026
query: 175 ,AP: 0.38910505836575876
query: 176 ,AP: 0.9009262938455572
query: 177 ,AP: 0.5071983940925266
query: 178 ,AP: 0.4349258382604033
query: 179 ,AP: 0.5362415852757823
query: 180 ,AP: 0.17093211555831836
query: 181 ,AP: 0.8784449891067538
query: 182 ,AP: 0.19305019305019305
query: 183 ,AP: 0.40261139546974
query: 184 ,AP: 0.4913751187211148
query: 185 ,AP: 0.7874829424643556
query: 186 ,AP: 0.865780741317506
query: 187 ,AP: 0.9368170982382031
query: 188 ,AP: 0.36961361516230273
query: 189 ,AP: 0.12419942509228223
query: 190 ,AP: 0.6069542727923912
query: 191 ,AP: 0.6612860715742275
query: 192 ,AP: 0.6412566209219471
query: 193 ,AP: 0.2927910231029232
query: 194 ,AP: 0.9027589886964886
query: 195 ,AP: 0.2561251639653753
query: 196 ,AP: 0.670590934782289
```

```
query: 197 ,AP: 0.8018388687068974
query: 198 ,AP: 0.5157256700559849
query: 199 ,AP: 0.2375296912114014
query: 200 ,AP: 0.38990957499276363
query: 201 ,AP: 0.37037037037037035
query: 202 ,AP: 0.6333502742108185
query: 203 ,AP: 0.04635527117796621
query: 204 ,AP: 0.8955971196910478
query: 205 ,AP: 0.5996269251361193
query: 206 ,AP: 0.7393691497709068
query: 207 ,AP: 0.7576644085514801
query: 208 ,AP: 0.30027491623197033
query: 209 ,AP: 0.16447368421052633
query: 210 ,AP: 0.6170820643072044
query: 211 ,AP: 0.8562161245017732
query: 212 ,AP: 0.6036681920460846
query: 213 ,AP: 0.37611673135240004
query: 214 ,AP: 0.6936091048965641
query: 215 ,AP: 0.30120481927710846
query: 216 ,AP: 0.5939181835210928
query: 217 ,AP: 0.44497251408872024
query: 218 ,AP: 0.24556753773279044
query: 219 ,AP: 0.26899219217558346
query: 220 ,AP: 0.38486226965736386
query: 221 ,AP: 0.1988071570576541
query: 222 ,AP: 0.3322528087616565
query: 223 ,AP: 0.7796197504641966
query: 224 ,AP: 0.775448717948718
query: 225 ,AP: 0.957056732521723
MAP = 0.5330725049635453
query 171 , NDCG:  0.9692912068993811
query 172 , NDCG:  0.9140975137410852
query 173 , NDCG:  0.4106873089767272
query 174 , NDCG:  0.8923867042591163
query 175 , NDCG:  0.7425509592426285
query 176 , NDCG:  0.7979067036416867
query 177 , NDCG:  0.6994188191743311
query 178 , NDCG:  0.7290081554224127
query 179 , NDCG:  0.6419612794808927
query 180 , NDCG:  0.47919859278247656
query 181 , NDCG:  0.7973880585790277
query 182 , NDCG:  0.48317036882681597
query 183 , NDCG:  0.9334739440870602
query 184 , NDCG:  0.657289608120126
query 185 , NDCG:  0.8492859942635851
query 186 , NDCG:  0.7982625507126624
query 187 , NDCG:  0.7336287368557511
query 188 , NDCG:  0.5394499939898107
query 189 , NDCG:  0.2698193939025438
```

```
query 190 , NDCG:  0.6812540167735646
query 191 , NDCG:  0.7420263390078213
query 192 , NDCG:  0.6995672265327972
query 193 , NDCG:  0.40049332775517527
query 194 , NDCG:  0.9053416289485289
query 195 , NDCG:  0.5761890524919981
query 196 , NDCG:  0.7506605908220092
query 197 , NDCG:  0.8583372627804899
query 198 , NDCG:  0.5650616742951504
query 199 , NDCG:  0.9188764054550188
query 200 , NDCG:  0.8012972829910111
query 201 , NDCG:  0.7696056311597761
query 202 , NDCG:  0.7933641616410014
query 203 , NDCG:  0.13226228479746177
query 204 , NDCG:  0.8351374707012883
query 205 , NDCG:  0.9185127202534908
query 206 , NDCG:  0.7118901528197593
query 207 , NDCG:  0.8192573443191097
query 208 , NDCG:  0.637957527964691
query 209 , NDCG:  0.7335235281932502
query 210 , NDCG:  0.7645241615821637
query 211 , NDCG:  0.890258583255237
query 212 , NDCG:  0.8765327598652619
query 213 , NDCG:  0.9671106642947606
query 214 , NDCG:  0.8287772410706995
query 215 , NDCG:  0.5527757806583375
query 216 , NDCG:  0.8158158813584936
query 217 , NDCG:  0.5921866927671411
query 218 , NDCG:  0.4962399912576918
query 219 , NDCG:  0.3877165162211433
query 220 , NDCG:  0.46820467017107176
query 221 , NDCG:  0.7203354382864378
query 222 , NDCG:  0.5516217222796262
query 223 , NDCG:  0.7355380872692105
query 224 , NDCG:  0.8533895605520561
query 225 , NDCG:  0.7700888332494576
NDCG = 0.7065456019418237

avdl = 108

query: 171 ,AP: 0.9828862203770016
query: 172 ,AP: 0.33371554092104055
query: 173 ,AP: 0.271637823841046
query: 174 ,AP: 0.768371212121212
query: 175 ,AP: 0.38910505836575876
query: 176 ,AP: 0.8843577596814124
query: 177 ,AP: 0.4779189737624244
query: 178 ,AP: 0.4215943233002161
query: 179 ,AP: 0.530731015044368
```

```
query: 180 ,AP: 0.17271157167530224
query: 181 ,AP: 0.8784449891067538
query: 182 ,AP: 0.19305019305019305
query: 183 ,AP: 0.41039053313624946
query: 184 ,AP: 0.5082934095020252
query: 185 ,AP: 0.7814225695846989
query: 186 ,AP: 0.865780741317506
query: 187 ,AP: 0.9270619386988926
query: 188 ,AP: 0.3465729843165554
query: 189 ,AP: 0.10869640491855657
query: 190 ,AP: 0.582987800017266
query: 191 ,AP: 0.657217616165317
query: 192 ,AP: 0.6492407058025836
query: 193 ,AP: 0.2739667012875358
query: 194 ,AP: 0.8657839227343538
query: 195 ,AP: 0.2521173804076733
query: 196 ,AP: 0.6671136235366413
query: 197 ,AP: 0.7906307447363234
query: 198 ,AP: 0.4721423625469731
query: 199 ,AP: 0.2375296912114014
query: 200 ,AP: 0.383118860973156
query: 201 ,AP: 0.37037037037037035
query: 202 ,AP: 0.5999480968902542
query: 203 ,AP: 0.046269598357734844
query: 204 ,AP: 0.8899396952032215
query: 205 ,AP: 0.5872620913918946
query: 206 ,AP: 0.7322095385320941
query: 207 ,AP: 0.7443828296767915
query: 208 ,AP: 0.2962246579742382
query: 209 ,AP: 0.16447368421052633
query: 210 ,AP: 0.5508421449387835
query: 211 ,AP: 0.8213349502148265
query: 212 ,AP: 0.5996251967340038
query: 213 ,AP: 0.3722719388743559
query: 214 ,AP: 0.6917557023122992
query: 215 ,AP: 0.29791018703674066
query: 216 ,AP: 0.5922083589642355
query: 217 ,AP: 0.4388608093251552
query: 218 ,AP: 0.26069645292226823
query: 219 ,AP: 0.2639030582542098
query: 220 ,AP: 0.38404460637275273
query: 221 ,AP: 0.1988071570576541
query: 222 ,AP: 0.33035430134625515
query: 223 ,AP: 0.7378516165880517
query: 224 ,AP: 0.7649725274725274
query: 225 ,AP: 0.9270805324469833
MAP = 0.5226944141019757
query 171 , NDCG:  0.9499677964240851
query 172 , NDCG:  0.9210872421855857
```

```
query 173 , NDCG:  0.3755025241595464
query 174 , NDCG:  0.8886063162305364
query 175 , NDCG:  0.7285627092736963
query 176 , NDCG:  0.7600345622646916
query 177 , NDCG:  0.6896547661727742
query 178 , NDCG:  0.7318799844903986
query 179 , NDCG:  0.6235211466261734
query 180 , NDCG:  0.4184550713595052
query 181 , NDCG:  0.850655635736636
query 182 , NDCG:  0.48269128420711765
query 183 , NDCG:  0.9474262933690134
query 184 , NDCG:  0.6482148866647203
query 185 , NDCG:  0.8477643748123319
query 186 , NDCG:  0.7982625507126624
query 187 , NDCG:  0.7321622516552911
query 188 , NDCG:  0.4957284139799904
query 189 , NDCG:  0.2671687523696299
query 190 , NDCG:  0.6747079520595292
query 191 , NDCG:  0.7402554183536143
query 192 , NDCG:  0.7313093951118823
query 193 , NDCG:  0.39279604449349015
query 194 , NDCG:  0.8893994003187208
query 195 , NDCG:  0.572803200161778
query 196 , NDCG:  0.7502610763133895
query 197 , NDCG:  0.8533580147792659
query 198 , NDCG:  0.4981509382519009
query 199 , NDCG:  0.9050287652044412
query 200 , NDCG:  0.7808024968646106
query 201 , NDCG:  0.7565319321774889
query 202 , NDCG:  0.7635434409003884
query 203 , NDCG:  0.13208095687992583
query 204 , NDCG:  0.8340516583205824
query 205 , NDCG:  0.8957787048289949
query 206 , NDCG:  0.6998930016949738
query 207 , NDCG:  0.7989096581190824
query 208 , NDCG:  0.6272402644398822
query 209 , NDCG:  0.7273240323186971
query 210 , NDCG:  0.7190837403619529
query 211 , NDCG:  0.8845615561364529
query 212 , NDCG:  0.8662703465049315
query 213 , NDCG:  0.9634814418579712
query 214 , NDCG:  0.8250826602076867
query 215 , NDCG:  0.542982762907532
query 216 , NDCG:  0.8178073141528954
query 217 , NDCG:  0.5943050587585746
query 218 , NDCG:  0.5111497930347155
query 219 , NDCG:  0.3804350118218232
query 220 , NDCG:  0.43855449380317263
query 221 , NDCG:  0.6933375600987131
```

```
query 222 , NDCG:  0.5064829875764547
query 223 , NDCG:  0.7093506221588194
query 224 , NDCG:  0.8475837971095077
query 225 , NDCG:  0.7350066462731943
NDCG = 0.6948554310373035
```

VSM的:

```
query: 171 ,AP: 0.9833369307940968
query: 172 ,AP: 0.3371687625949428
query: 173 ,AP: 0.32737934622685255
query: 174 ,AP: 0.7972507297938333
query: 175 ,AP: 0.38910505836575876
query: 176 ,AP: 0.886089959996881
query: 177 ,AP: 0.4653401945465848
query: 178 ,AP: 0.41051038273771767
query: 179 ,AP: 0.5836911622005546
query: 180 ,AP: 0.17271157167530224
query: 181 ,AP: 0.8784449891067538
query: 182 ,AP: 0.19305019305019305
query: 183 ,AP: 0.3965950580099138
query: 184 ,AP: 0.5043235179075601
query: 185 ,AP: 0.7892380748970754
query: 186 ,AP: 0.8607807413175059
query: 187 ,AP: 0.9934769857984143
query: 188 ,AP: 0.36292224770014686
query: 189 ,AP: 0.12507851891702806
query: 190 ,AP: 0.5720286889903101
query: 191 ,AP: 0.6592541659028109
query: 192 ,AP: 0.6950599712838033
query: 193 ,AP: 0.2533950023356128
query: 194 ,AP: 0.9068275851563894
query: 195 ,AP: 0.25585335565864303
query: 196 ,AP: 0.6414139879108638
query: 197 ,AP: 0.8262907299039878
query: 198 ,AP: 0.49109815656159245
query: 199 ,AP: 0.2375296912114014
query: 200 ,AP: 0.379504372471809
query: 201 ,AP: 0.37037037037037035
query: 202 ,AP: 0.6205546549950702
query: 203 ,AP: 0.047595150517625445
query: 204 ,AP: 0.8994610222940808
query: 205 ,AP: 0.6060606060606061
query: 206 ,AP: 0.7143276422965896
query: 207 ,AP: 0.7816878789477705
query: 208 ,AP: 0.30027491623197033
query: 209 ,AP: 0.162484288553155
query: 210 ,AP: 0.6013186892025746
```

```
query: 211 ,AP: 0.8637698595264385
query: 212 ,AP: 0.5903055076103568
query: 213 ,AP: 0.3710861856270742
query: 214 ,AP: 0.6931274073038283
query: 215 ,AP: 0.2981014906875492
query: 216 ,AP: 0.5894443884334954
query: 217 ,AP: 0.41966041120430286
query: 218 ,AP: 0.24123532440626003
query: 219 ,AP: 0.35460001986563483
query: 220 ,AP: 0.4151585635799762
query: 221 ,AP: 0.1988071570576541
query: 222 ,AP: 0.35003469489911565
query: 223 ,AP: 0.7684761168793831
query: 224 ,AP: 0.7611630036630037
query: 225 ,AP: 0.9872289176045429
MAP = 0.5342015345607777
query 171 , NDCG:  0.9527938859239503
query 172 , NDCG:  0.9149601580660057
query 173 , NDCG:  0.4262081046108922
query 174 , NDCG:  0.890747779114999
query 175 , NDCG:  0.7060592618965732
query 176 , NDCG:  0.8137843598609691
query 177 , NDCG:  0.6756140728746463
query 178 , NDCG:  0.786379290477364
query 179 , NDCG:  0.6358818972625863
query 180 , NDCG:  0.45904400448574273
query 181 , NDCG:  0.7973880585790277
query 182 , NDCG:  0.47353585971677636
query 183 , NDCG:  0.9406209916763422
query 184 , NDCG:  0.6442791811433259
query 185 , NDCG:  0.8492859942635851
query 186 , NDCG:  0.7964857835761114
query 187 , NDCG:  0.8351688757193857
query 188 , NDCG:  0.5198918166174461
query 189 , NDCG:  0.2780737137518055
query 190 , NDCG:  0.6943120714052511
query 191 , NDCG:  0.734176892781448
query 192 , NDCG:  0.7612733787138043
query 193 , NDCG:  0.3684567741795951
query 194 , NDCG:  0.9015955006741198
query 195 , NDCG:  0.6005675132999104
query 196 , NDCG:  0.726555441916076
query 197 , NDCG:  0.8472255025000217
query 198 , NDCG:  0.5614520368850593
query 199 , NDCG:  0.9046137126908289
query 200 , NDCG:  0.7636826980254062
query 201 , NDCG:  0.7624959738931167
query 202 , NDCG:  0.7857949668211217
query 203 , NDCG:  0.13349095404863798
```

```
query 204 , NDCG:  0.8362307551045658
query 205 , NDCG:  0.9147207818479473
query 206 , NDCG:  0.7060543448044819
query 207 , NDCG:  0.8252627976490451
query 208 , NDCG:  0.644436391097654
query 209 , NDCG:  0.7026156184378821
query 210 , NDCG:  0.753645878438511
query 211 , NDCG:  0.890258583255237
query 212 , NDCG:  0.862834809647083
query 213 , NDCG:  0.9622232235551959
query 214 , NDCG:  0.8230806505455427
query 215 , NDCG:  0.5468916400717928
query 216 , NDCG:  0.8099176315206646
query 217 , NDCG:  0.6015702634542226
query 218 , NDCG:  0.4899997550011071
query 219 , NDCG:  0.44808937550676664
query 220 , NDCG:  0.4876446362007501
query 221 , NDCG:  0.6963287210347119
query 222 , NDCG:  0.48771134077702927
query 223 , NDCG:  0.621132869770906
query 224 , NDCG:  0.8401683339336118
query 225 , NDCG:  0.8178012760169435
NDCG = 0.7038275670022469
```