

山东大学 计算机科学与技术 学院

信息检索与数据挖掘 课程实验报告

学号：201600130032	姓名： 陆禹承	班级： 智能 16
实验题目：BM25		
实验内容： 实现 BM25 模型		
<p>实验过程中遇到和解决的问题：</p> <p>使用和上次相似的代码，主要更改的是增加 rating 功能</p> <p>算法过程：</p> <p>首先根据用户查询筛选出文档，然后对每个文档进行 rating（这些文档数据已经被分过词）</p> <p>rating 的式子大致如下：</p> <ul style="list-style-type: none">• Pivoted Length Normalization VSM [Singhal et al 96] $f(q, d) = \sum_{w \in q \cap d} c(w, q) \frac{\ln[1 + \ln[1 + c(w, d)]]}{1 - b + b \frac{ d }{avdl}} \log \frac{M + 1}{df(w)}$ <p>其中常数值根据多次尝试自行选择了一个合适的数值，这样排名出来的效果会好很多</p> <p>此外还对运算过程进行了优化，计算速度都有大幅度的提升。多个 AND/OR 运算可以同时运算（使用一个小根堆进行排序），对于 a&b&c 这种情况效率有小幅提升。</p> <p>为了方便 rating 的计算，</p> <p>做 inverted index list 时统计了词汇频率（tf），并统计所有文档长度</p> <p>结果出来以后，再对每个文档进行评价，最后根据得分排序。</p> <p>评析时对每个查询词计算结果求和</p> <p>文档处理要花费几分钟的时间，为了方便调试，将处理过的文档信息暂存下来（cache），方便之后读入使用（只需要几秒的时间）</p> <p>查询花费时间较小，简单查询基本<1s</p> <p>运行效果：</p>		

```

=====
input command:
I want to
---expression tree-----
||Tokens.ExpAnd3
|
|to
|want
|
|
=====result=====
Total: [ 49]
print all tweets? Y/n
Y
6.55783750916231 score I want my dad to come home quicker because I want to buy some clothes online
5.5010111772184365 score When I grow up I want to be Tony Mendez #argo @pencilgHH
5.5010111772184365 score I don't want to leave Manhattan, even when I'm gone. -Ed Koch"
5.5010111772184365 score Ughhh I want mcdonalds breakfast but I can't physically walk to mcdonalds...
5.231006492265586 score I totally want to cruise in this storm!
5.231006492265586 score I just want my snow blower to work
5.075100802020371 score I want to go see the Oz with Mila Kunis. ☐☐
4.992891455726123 score Why is the storm called Nemo? I mean really? How am I suppose to be worried about a storm with that name? I still want to find him.
4.947018336071816 score I really want to go to the cherry blossom festival in DC this weekend but idk
4.947018336071816 score I think the snow blower people just did not want to go to work today.
#thanks
4.930254024054658 score I really want to know who named this snowstorm nemo. i wonder if the next storm will be named dori?
4.930254024054658 score I really want to be patriotic and say Andy Murray for Aussie Open, but damn I just love Novak better
4.928219418907919 score All the clothing I want to buy online isn't in stock. Hmph.
4.917810818817871 score im so ready for ariana to be on knock knock live i want her to come to my house
4.8802612865053785 score @Paris_Plaza I want to know if there is a site online to buy clothes brand of paris hilton
4.8579214533801725 score I Want to Fall in Love with @OfficialBWFCAgain by @volsey28 #bwfc #f172 http://t.co/MzeQHRFSsp http://t.co/2KFYcTVwUv
4.8579214533801725 score yuummy mcdonald fries today . woow i don't want to catch i ha k chickens .
4.79188226283218 score Ive never retweeted myself but wanted to pass on to @atu2 RT @tommymcgregor: I Want Bono To Sing At My Funeral! http://bit.ly/iOkdEn
4.789600854077074 score I want to see the new Oz movie strictly because Mila Kunis is in it!☐☐
4.770636331457698 score I just want to find a review of a bathing suit I want to buy online and can't find one and am not buying it. #lavienrose #swimsuit
4.745611272084742 score @poooop I want a record player Then Imma order all the bands I think need to be on vinyl. Like the white stripes :)
4.723175262931498 score @SincerelyTumblr: perfect world: being able to buy all the concert tickets and clothes i want"
4.723175262931498 score Next I want to make homemade hush puppies and a variation using black eyed peas...
4.723175262931498 score Omw to mcdonalds who want some dats if u round whea i at! Lol aha
4.714580281079062 score I'm going to watch keeping up with the Kardashians at 10. I think it'll cheer me up. Really want to see the new Kim & Khloe in New York one.
4.690379557343701 score @knockknockFOX #im watchibg knock knock live # i want to win to buy my boyfriend of 18 years a harley motor cycle
4.687099149351 score Wow, fuck you #downton abbey. I have NEVER cried that hard over a show before. I didn't want to watch again after Sybill.And now Matthew ='(
4.65566943256257 score @todayshow: New owners hope to have twinkles in stores soon. Story: http://t.co/8l1J7R5cIB" Now I want one! :)
4.630327135882837 score Kim Kardashian: I Want to Raise My Children Not Just to See Color? While promoting her new film 'Temptation on ... http://t.co/mdjXs0J0RY

```

结论分析与体会：

对模型优化有所了解