相关性：定义与分档

王树森

http://wangshusen.github.io/

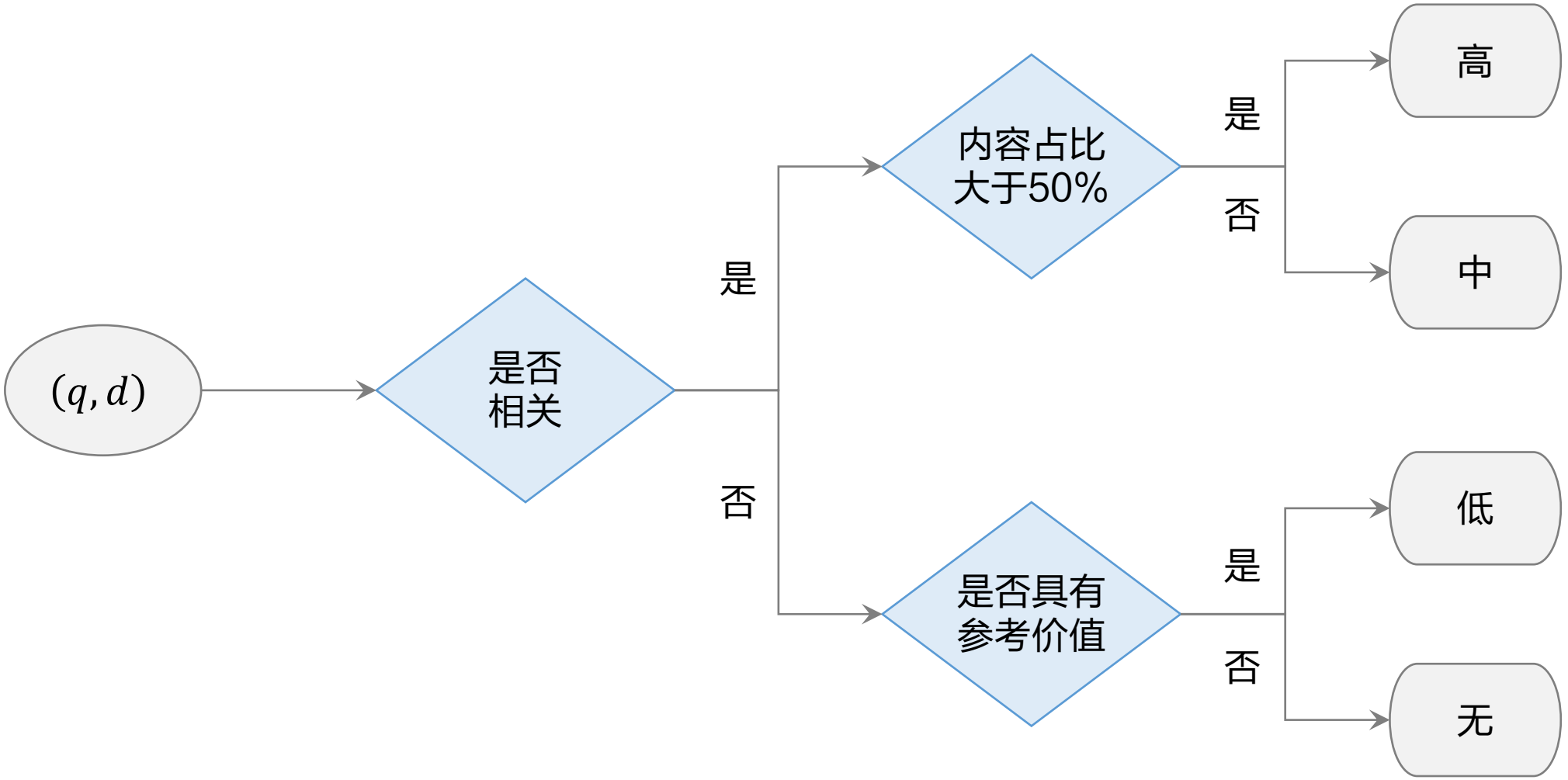
工业界是怎么做的？

- 制定标注规则 → 标注数据 → 训练模型 → 线上推理。

工业界是怎么做的？

- 制定标注规则 → 标注数据 → 训练模型 → 线上推理。
- 搜索产品和搜索算法团队定义相关性标注规则。
  - 人为将 (q, d) 的相关性划分为 4 个（或 5 个）档位。
  - 相关性分档规则非常重要！假如日后有大幅变动，需要重新标注数据，丢弃积累的数据。
- 产品和算法团队监督指导标注团队的工作，累积数十万、数百万条 (q, d) 样本。
- 算法团队用人工标注的数据训练相关性模型。

相关性档位划分



相关 vs 不相关

字面匹配 vs 需求匹配

- 相关性是指 d 能满足 q 的需求或回答 q 提出的问题。
- 哪怕 q 和 d 字面上完全不匹配，两者也可以被判定为相关。
  - q = 谁掌握芯片制造的尖端技术
  - d = 全球最先进的光刻机都由荷兰 ASML 公司制造……

字面匹配 vs 需求匹配

- 相关性是指 d 能满足 q 的需求或回答 q 提出的问题。
- 哪怕 q 和 d 字面上完全不匹配，两者也可以被判定为相关。
- 即便 q 和 d 字面匹配，两者也可能不相关。
  - q = 巴伦西亚旅游
  - d = 我去巴伦西亚旅游，吃到了最好最正宗的西班牙海鲜饭，回来研究了一番，这个视频给大家介绍西班牙海鲜饭的做法……

相关性标注只考虑相关性!

- 相关性标注只考虑相关性，不考虑内容质量、时效性等因素。
- 满足相关性，但是内容质量低：
  - q = 什么药物可以治愈新冠？
  - d = 一则虚假广告，声称某种草药可以治愈新冠，并用阴阳调和原理解释该草药克制新冠病毒。
- 满足相关性，但是时效性低：
  - q = 上海落户政策
  - d = 一篇过时的文章，介绍2015年的上海落户政策。

多意图

- 查询词 q 可能有多种意图，文档 d 只需命中一种意图就算相关。
- 黑寡妇：黑寡妇蜘蛛、漫威电影黑寡妇角色、车臣黑寡妇组织。
- 用户搜 q =“黑寡妇”，不论用户的意图是什么，黑寡妇蜘蛛、黑寡妇角色、黑寡妇组织的文档都满足相关性。

上位词、下位词

- 搜上位词，出下位词，判定为相关。
  - 搜 q =“广东菜”，出 d =“潮汕美食”。
  - 搜 q =“红色口红”，出 d =“玫红色口红”。
- 搜下位词，出上位词，判定为不相关。
  - 搜 q =“潮汕美食”，出 d =“经典广东菜”。
  - 搜 q =“玫红色口红”，出 d =“红色口红”。

丢词的判定

- 丢失核心词，判定为不相关。
  - 搜 q =“情人节餐厅”，出 d =“情人节礼物”。
  - 搜 q =“黄晓明”，出 d =“杨颖拍过的电影”。

丢词的判定

- 丢失核心词，判定为不相关。
- 丢失重要限定词，判定为不相关。
  - 搜 q =“初二物理考点”，出 d =“初三物理考点”。
  - 搜 q =“黄石公园春季旅游”，出 d =“黄石公园秋季旅游”。

丢词的判定

- 丢失核心词，判定为不相关。
- 丢失重要限定词，判定为不相关。
- 丢失不重要限定词，判定为相关。
  - 搜 q =“精彩的好莱坞动作片”，出 d =“好莱坞动作片 top 10”。
  - 搜 q =“东南亚十大旅游景点”，出 d =“东南亚热门旅游景点”。

丢词的判定

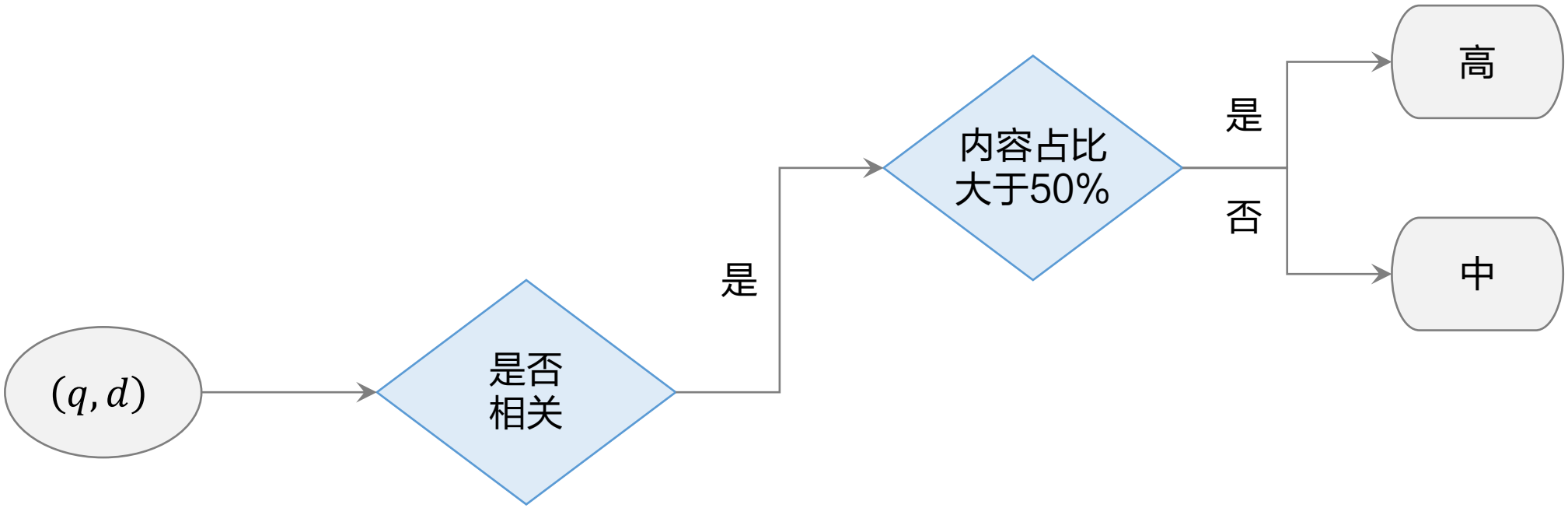
- 丢失核心词，判定为不相关。
- 丢失重要限定词，判定为不相关。
- 丢失不重要限定词，判定为相关。
- 具体要看 d 能否满足 q 的主要需求或回答 q 提出的问题。
  - 搜 q =“精彩的好莱坞动作片”，d =“好莱坞动作片 top 10”可以满足 q 的需求，所以相关。
  - 搜 q =“精彩的好莱坞动作片”，d =“精彩的宝莱坞动作片”无法满足 q 的需求，所以不相关。
  - 搜 q =“精彩的好莱坞动作片”，d =“精彩的好莱坞爱情动作片”无法满足 q 的需求，所以不相关。

相关性判定：小结

- 相关性是指 d 能满足 q 的需求或回答 q 提出的问题，而非字面上的匹配。
- 相关性标注只考虑相关性，不考虑内容质量、时效性。
- 如果 q 有多种意图，只要命中一种意图，就判定为相关。
- 搜上位词出下位词，判定为相关；搜下位词出上位词，通常判定为不相关。
- 丢核心词、重要限定词，判定为不相关；丢不重要的限定词，判定为相关。

档位细分

根据内容占划分高、中档位



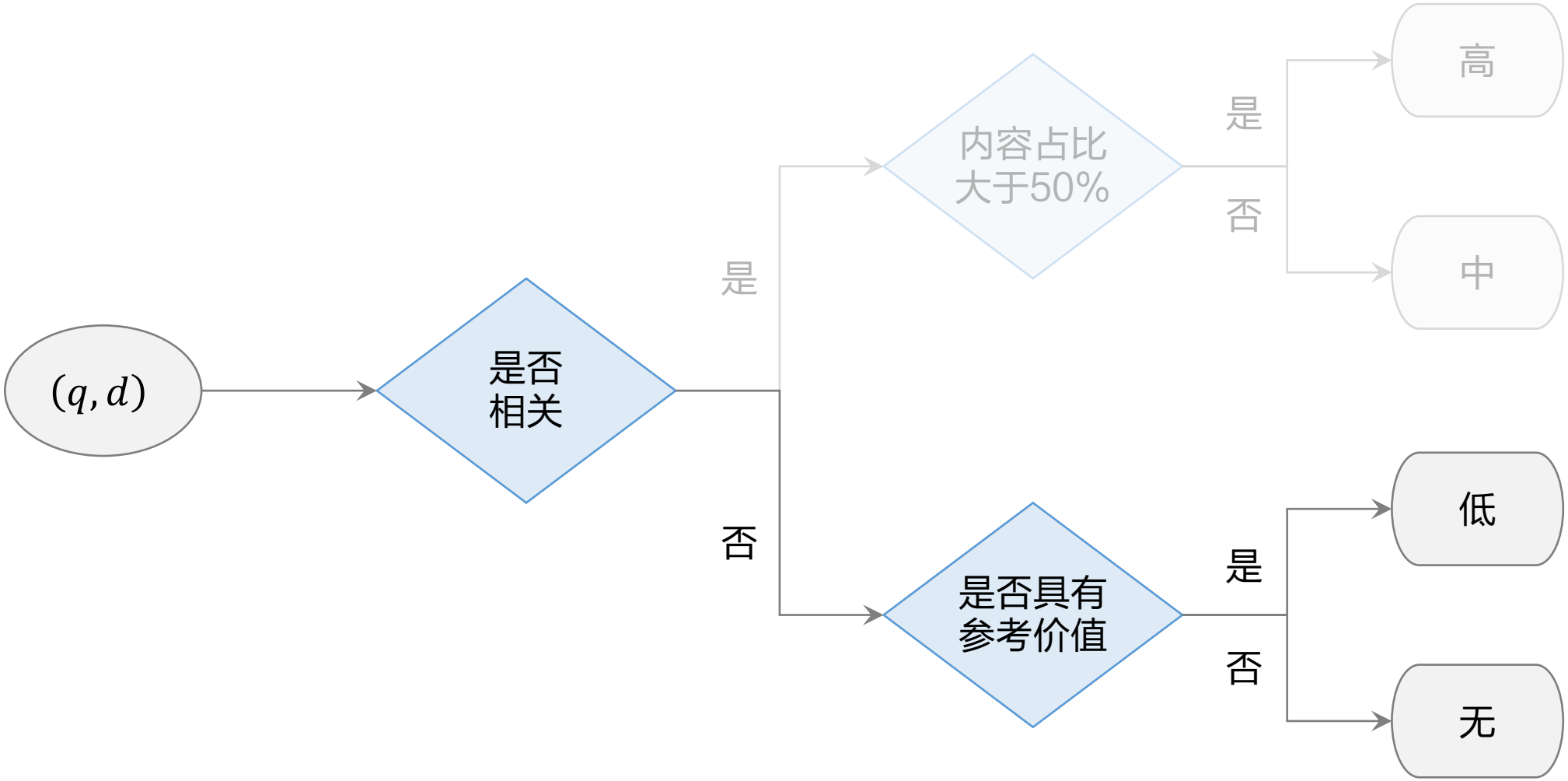
根据内容占划分高、中档位

- 如果 (q,d) 相关，则进一步划分为高、中两档。细分规则：满足需求的内容的篇幅占比是否超过 50%。
- 例 1：
  - 搜索 q =“泰坦尼克号”，出 d = 演员莱昂纳多关于他的代表作的访谈，其中重点谈了《泰坦尼克号》电影。
  - 文档 d 满足查询词 q 的需求，判定为相关。
  - 如果访谈内容中《泰坦尼克号》篇幅占比大于 50%，判定为高档位，否则判定为中档位。

根据内容占划分高、中档位

- 如果 (q,d) 相关，则进一步划分为高、中两档。细分规则：满足需求的内容的篇幅占比是否超过 50%。
- 例 2：
  - 搜索 q =“小米手机测评”，出 d = 几款安卓手机的测评，其中包括几款小米手机。
  - 文档 d 满足查询词 q 的需求，判定为相关。
  - 如果文档中小米手机篇幅占比大于 50%，判定为高档位，否则判定为中档位。

根据参考价值划分低、无档位



根据参考价值划分低、无档位

- 如果 (q,d) 不相关，则进一步划分为低、无两档。细分规则：文档是否具有参考价值。
- 例 1：
  - 搜索 q =“初二下册物理考点”，出 d =“中考物理考点”。
  - 丢失重要限定词，导致文档 d 无法满足查询词 q 的需求，判定为不相关。
  - “中考物理考点”有一定参考价值，档位为“低”。

根据参考价值划分低、无档位

- 如果 (q,d) 不相关，则进一步划分为低、无两档。细分规则：文档是否具有参考价值。
- 例 2：
  - 搜索 q = “初二下册物理考点”，出 d = “初一数学考点”。
  - 丢失重要限定词，导致文档 d 无法满足查询词 q 的需求，判定为不相关。
  - “初一数学考点”没有参考价值，档位为“无”。

总结

总结

- 相关性是指 d 能满足 q 的需求或回答 q 提出的问题。
- 先判断 q 与 d 是否相关，划分为两大档位。
  - 判断是否相关，只考虑相关性本身，不要考虑内容质量、时效性、个性化等其他因素。
  - q 可能有多种意图，只要 d 命中其中一种意图，就算相关。
  - 搜上位词出下位词，判定为相关；反之，通常判定为不相关。
  - 如果 d 丢弃了 q 中的词，需要判断 d 能否满足 q 的需求，从而判断是否相关。

总结

- 相关性是指 d 能满足 q 的需求或回答 q 提出的问题。
- 先判断 q 与 d 是否相关，划分为两大档位。
- 将大档位细分为 4 个小档位。
  - 根据所占篇幅，将“相关”细分为高、中 2 个小档位。
  - 根据文档是否有参考价值，将“不相关”细分为低、无 2 个小档位。
  - 相关性细分为高、中、低、无 4 个小档位。
  - 有的公司将“相关”细分为 3 个小档位，“不相关”细分为 2 个小档位，一共 5 个小档位。

标注的流程

- 由算法团队抽取待标注样本。
  - 从搜索日志中随机抽取 n 条查询词。既有高频查询词，也有中、低频查询词。
  - 给定 q，从搜索结果中抽取 k 篇文档，组成二元组 $(q, d_1), \cdots, (q, d_k)$。4 个相关性档位的样本数量尽可能平衡。
  - 不能直接取搜索结果页排名 top k 的文档，否则高档位文档过多，低档位文档过少。

标注的流程

- 由算法团队抽取待标注样本。
- 由产品团队和算法团队监督标注过程和验收结果。
  - 遇到难以界定档位的 (q, d)，由产品和算法团队做界定和解释。
  - 一条样本由至少两人标注，两人标注的结果需要有一致性。
  - 一致率大于某个阈值（例如 80%）才会被接受。
  - 产品团队抽查标注结果，要求准确率高于某个阈值。
  - 可以事先往数据中“埋雷”（产品团队自己标注的样本），考察埋雷样本的标注准确率。

Thank You!

http://wangshusen.github.io/