https://docs.google.com/a/morebits.org/File?id=dhcqmhqx_4x76fvrfm_b

**姓名：杨波**

**题目：基于改进语言模型的网页排序问题研究**

**评价结果：（进一步修改后）通过**

**有关意见：**

**[1] 论文的题目过于一般化，从题目本身不知道我们工作重点何在。**

***由于论文本身在传统语言模型的三各方面都作出了改进，“改进语言模型”可能是对本文做法的最好的称呼，所以暂时对题目未作改动。***

**[2] 加强调研分析。**

* **本文是对语言模型进行研究，那么语言模型的适用范围，在信息检索中与其它模型相比为什么是值得研究的？这些问题应进行认真阐述。**
* **对论文中提出的问题需要进一步花力气进行调研分析，加以界定。所提p(q|d)的估计、文档中未见词的p(term not in d| d)的估计和p(d)的估计，均为LM中无法回避的基础问题，需将其发展和最新动态讲清。通过调研分析，一方面让自己对这三个概率的估计有更深的认识，另一方面通过同别人工作的比较可以让你更好的突显你的工作.**

***报告修改版，加大了对语言模型，尤其是n-gram方法的调研力度。增加了关于语言模型，这种基于统计学的方法优势的论述，对其使用范围进行了调研并且在报告中增加了相应说明。***

***对于语言模型三个概率估计的问题，报告修改版中，也加入了对三个概率估计的详细情况论述。尤其，以p(q|d)的估计，报告简单阐述了，基本n-gram模型和其主要变种，以及在信息检索中应用的语言模型的情况。报告增加了数篇新的引用论文，对这三个概率的估计方法主要代表性的方法和近些年来的进展都做了阐述。***

**[3] 论文的核心工作应给出具体的形式化描述，而不是仅仅说“提出了XX方法”。例如，文中称提出了“变长多元组”，那么是如何实现变长的？引入变长之后，相应的概率估计会有哪些变化，出现哪些新的问题？变长多元组是否能真正解决数据稀疏问题？这些都应该给与严谨的考虑和论述。**

***报告修改版，分别对三种提出的算法进行了形式化描述，同时对数据稀疏问题，做了简单说明。随着论文工作的深入，配合更多的实验结果，我会对数据稀疏问题进行更深层次的论述。***

**[4]  应对提出的模型有全面深入的认识：**

* **文中所提的“变长多元组”方法实际上是不同长度多元组的线性组合，虽然比原来的模型复杂，但是长度仍然是固定的，所以不宜称作“变长”。**
* **N-gram语言模型中之所以n通常取较小的值，是因为n的值越大，数据越稀疏，难于建模。数据的稀疏性是语言模型考虑的一个重要方面。那么本文方法所提的不同长度多元组方法的数据稀疏性也应给出一定分析。**

***变长多元组这个名称确实有所不妥，报告修改版将其改为“多维多元组”，即我们的试验中采用3个不同长度的多元组，可以认为是，3个维度的数据进行加权得到最终结果。***

**[5] 请尽快开展实验设计工作。**

* **把试验平台的目标、功能、结构描述清楚，把技术路线确定。**
* **实验目的阐述不清，例如：所给出的不同n-gram下返回结果的相关文档比例的统计结果 距离你的报告目标——“使用统计语言模型提升信息检索排序性能”过于细节，设计出如下实验会离你要验证的目标会更近些，如，统计语言模型同其它模型的信息检索的性能比较；统计语言模型中各方法间的性能比较；变长多元组统计语言模型中，不同参数设置下的性能比较。**

***报告修改版，对实验平台lemur的情况给出了详细描述，并且将将来要进行的实验的技术路线做了一定的说明。***

***目前，报告中出现的实验，都是论证性的实验，目的在于指出问题所在，下一步的实验正是，多种模型性能的对比等等，下一步实验的详细步骤，已经在报告修改版中给出。由于，实验处理数据量巨大，所以进展不是很快，在本学期的讨论班中，我会给出提出算法的实验结果。***