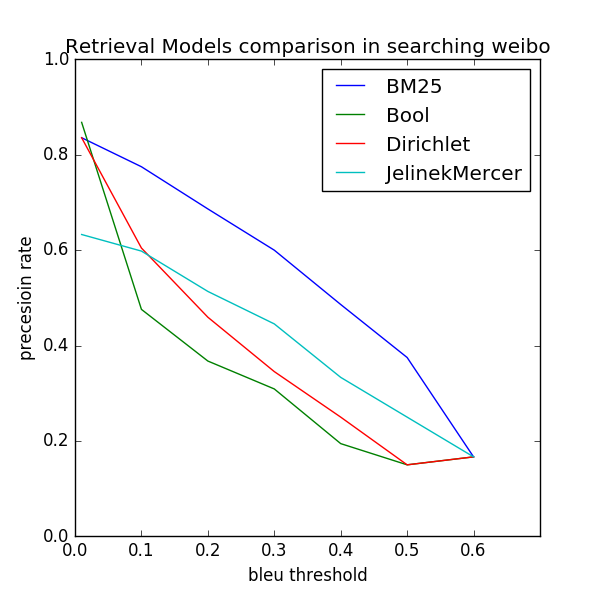
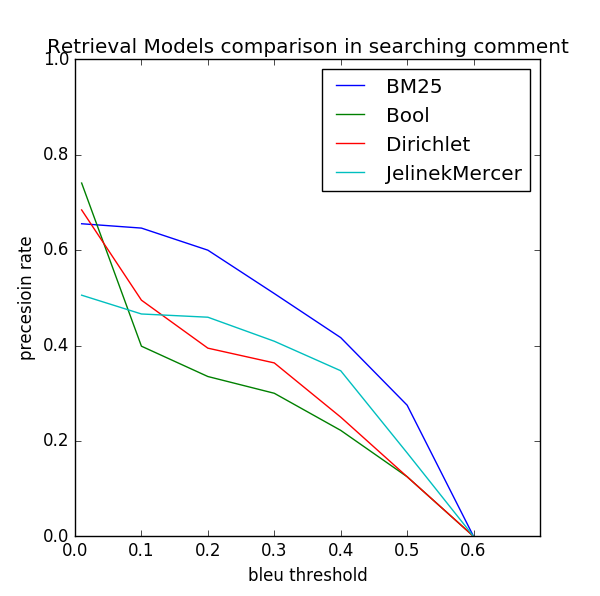
本周主要任务：

相关模型的评价问题

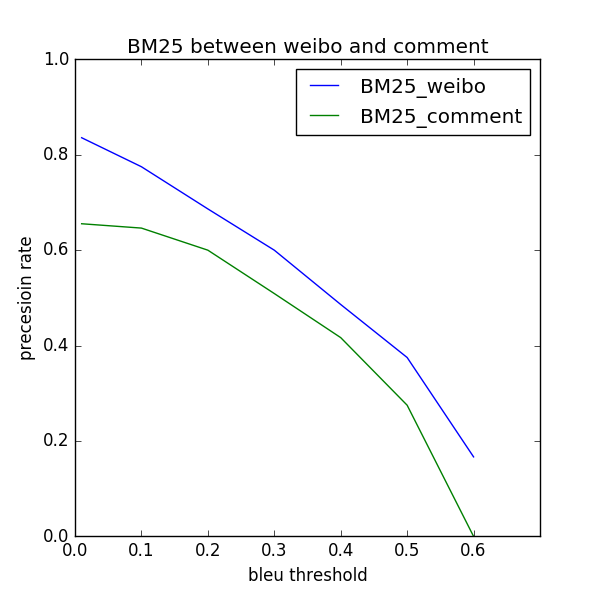
1、评价数据来自后期爬取的2752条微博数据及对应的评论数据

2、评价方法：通过设置bleu阈值，进行筛选对话（微博与对应的一条评论）；检索微博计算检索结果与评论的bleu值，计算达到阈值的数目，作为准确率。检索评论计算检索结果与微博的bleu值，计算达到阈值的数目。





由结果可见，BM25的效果是最好的。 对单独的模型，也对比了检索微博与检索评论的结果性能，以BM25为例，如图



**评价数据不靠谱**，原因一是，再次爬取的数据中，有少量的微博数据和作为数据源的数据有重复，影响评价的结果；二是再次爬取的数据，大部分电影都不是作为数据源的那部分电影，因此相关性较差。

**方案：**数据源微博分为5类，首先对微博进行了去重，然后选择一类作为评价数据（1787条微博），剩下4类作为检索的数据源；保证去除上述不靠谱因素。如下图，BM25仍有相对较好的性能。

