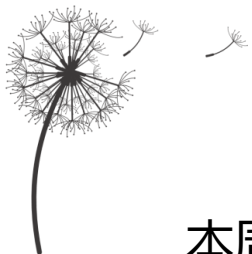


Weekly report


霍沛



本周完成：

- 1.修改之前提出的模型，并开始训练
- 2.整理评测指标的code，准备测试
- 3.学习 Biterm Topic Model （BTM短文本主题模型），并在数据集上进行测试





A Biterm Topic Model for Short Texts


Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng
Institute of Computing Technology, CAS
Beijing, China 100190

yanxiaohui@software.ict.ac.cn, {guojiafeng, lanyanyan, cxq}@ict.ac.cn

相比于LDA模型:

LDA是document-level, BTM是corpus-level

LDA 计算的是topic-document-word 的概率分布, BTM计算的是topic-Biterm的概率分布





假定语料库中有 N 个主题

1. 计算出每篇文档的主题概率分布，排序选出概率最大的前 k 个主题
2. 通过模型得到每个主题下词语的概率分布，排序选出符号当前主题下概率最大的前 p 个词语，作为候选关键词
3. 如果候选关键词在原文中，则判定该词语为关键词

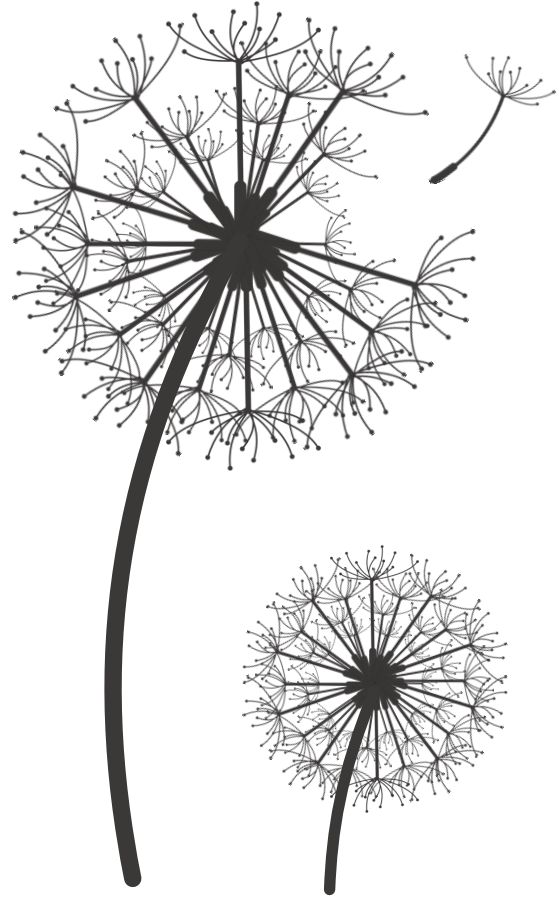




下周计划:

1. 查阅论文研究利用VAE学习主题信息，加载到模型中 **
2. 测试模型效果





THANKS