掷骰子到视频主题模型

分享by：曹文龙

主要内容

- 为什么选择主题特征
- 从掷骰子看主题模型
- 应用主题特征
  - LDA在图文中的应用
  - LDA特征在视频中的问题
  - BTM视频主题特征

如何表示语料

- 图文语料长，词汇规模大(27w+)
- 使用关键词进行查询
  - 维度高
  - 词义鸿沟
    - 苹果秋季新品发布会将于北京时间9月13日凌晨01:00在乔布斯剧院正式召开。
    - 荣耀8玩王者荣耀很是流畅。
- 文档主题生成模型

文档主题生成文档案例

- 骰子生成：666
- 主题模型生成：星期六是假期。

Unigram model主题模型

- 一个普通的6面骰子掷3次，每次都是6
- 一个6面都是6的骰子投掷3次，每次都是6
- 哪个最有可能生成666？
- 星期六 是 假期。[星期六 工作日 假期 周五 是 不是]

Mixture of unigram主题模型

- K个特殊的6面骰子， 每个骰子都不同，从中选择1个骰子，投掷三次
- 如何出现666的概率最大？
  - 选择6出现概率最大的骰子A
  - 想尽一切办法选中A

- 星期六 是 假期。[星期六 工作日 假期 周五 是 不是]

应用主题模型

应用场景文本案例

诗词：寒蝉凄切，对长亭晚，骤雨初歇。都门帐饮无绪，留恋处，兰舟催发。执手相看泪眼，竟无语凝噎。念去去，千里烟波，暮霭沉沉楚天阔。
多情自古伤离别，更那堪、冷落清秋节。
今宵酒醒何处？杨柳岸晓风残月。此去经年，应是良辰好景虚设。便纵有千种风情，更与何人说！

送别
儿女情长
以景写情
虚实结合

新闻：从北京铁路局警方获悉，近日，河北滦县村民庞某，在当地青龙山高铁沿线附近试飞自制航模时，因操作不当导致航模失控，掉落在京秦线高铁线路上，“逼停”高速行驶的G2604次列车，造成列车晚点22分钟，危及了广大乘车旅客和列车的运行安全。目前，庞某已被依法行政拘留。

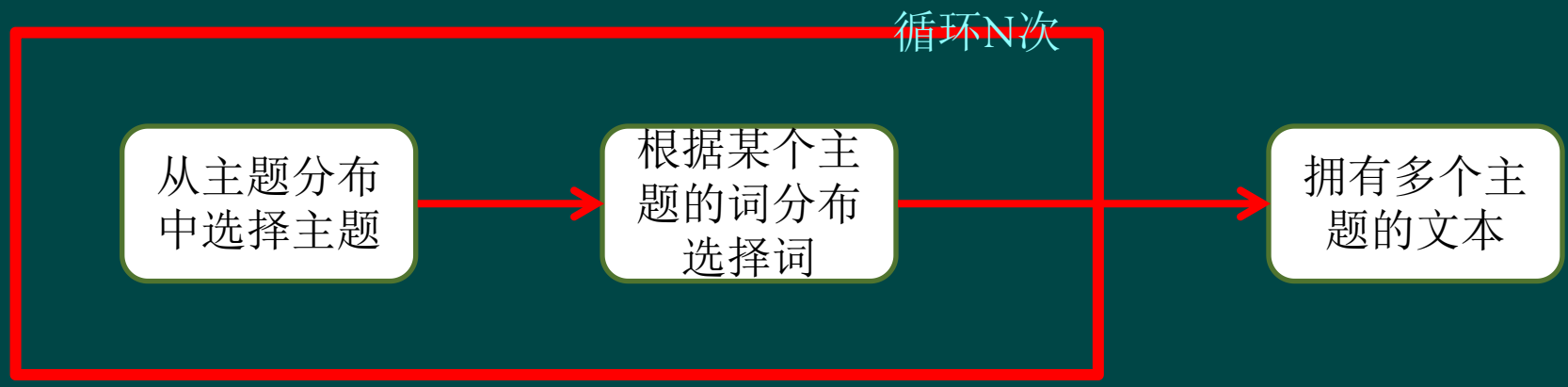
警方执法
航模监管
高铁
手工DIY

应用场景文本特点

- 前面介绍的两个主题模型特点：
  - 单一主题生成
- 现实中文本：混合多主题

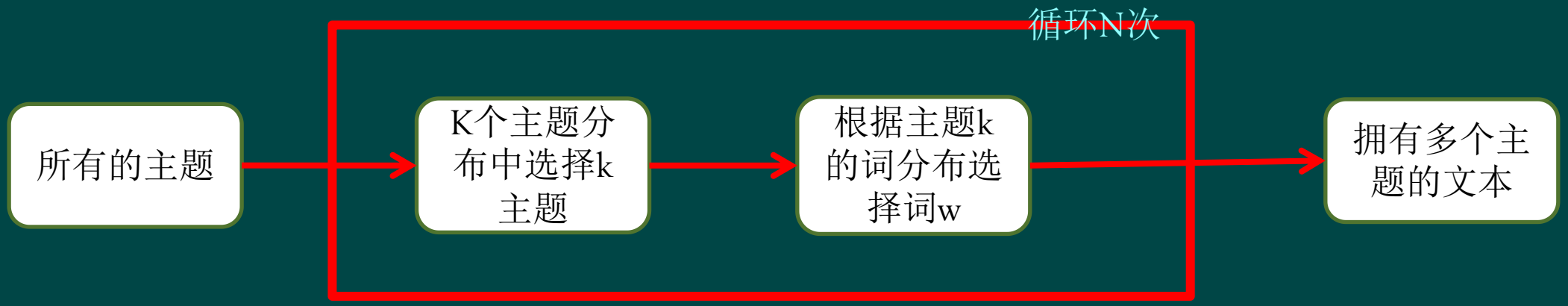
概率潜在语义分析(pLSA)

- K个特殊的6面骰子， 每个骰子都不同，[从中选择1个骰子，投掷一次]，重复三次操作
- N个词的文本



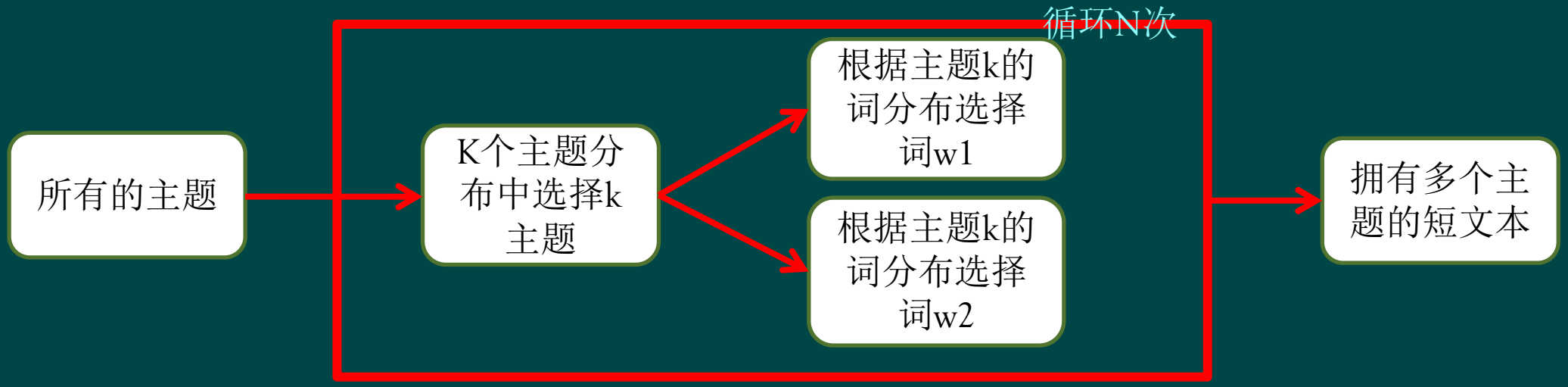
潜在狄利克雷分布模型(LDA)

- 一个罐子里盛放着N骰子（每个骰子都不同），从中抓取K(K < N)个骰子，[从K个骰子中选择1个骰子，投掷1次]，重复3次
- N个词的文本



双词主题模型

- 人眼分辨短文本的过程，并不是孤立地看每个词是否出现，而是要关注某些词是否一起出现
- N个双词的文本(文本456 : [45, 46, 56])



参考文献

1. Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short
2. texts. In Proceedings of the 22nd international conference on World Wide Web (WWW '13). ACM,
3. New York, NY, USA, 1445-1456. DOI: https://doi.org/10.1145/2488388.2488514
4. [BTM 代码](https://github.com/xiaohuiyan/BTM)
5. D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models.In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 262–272. Association for Computational Linguistics, 2011.
6. 欢迎留言：http://km.oa.com/group/1397/articles/show/314103
7. http://blog.csdn.net/pipisorry/article/details/42560693
8. GloVe: Global Vectors forWord Representation

谢谢！