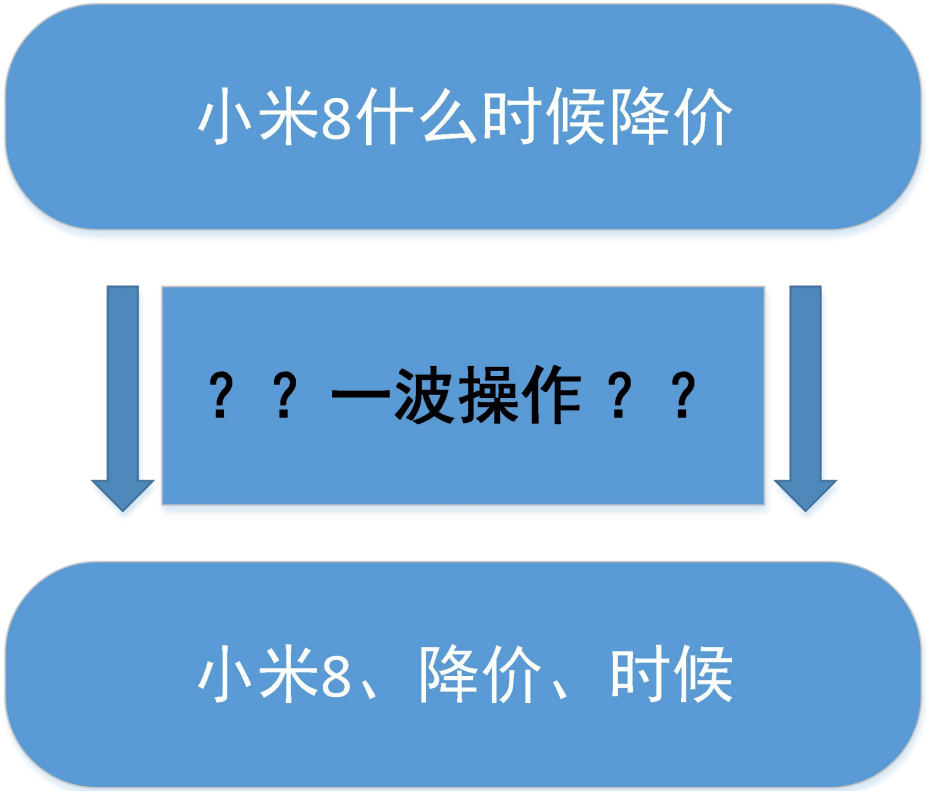
短文本关键词提取

TF-IDF&GPU-DMM

组长：高望　　**组员**：白春飞、赵海玮、刁永祥

第一部分　文本及其特征

需求简介



长文本 VS 短文本

- 长文本（网页新闻、文献等）
- 篇幅长，信息量大
- 可用的词语特征多

- 短文本（微博、题目摘要等）
- 篇幅短，信息量少
- 可用的词语特征少

词语的统计特征

词频（TF）

逆文档频率（IDF）

词语长度

词语位置信息

……

词语的语义特征

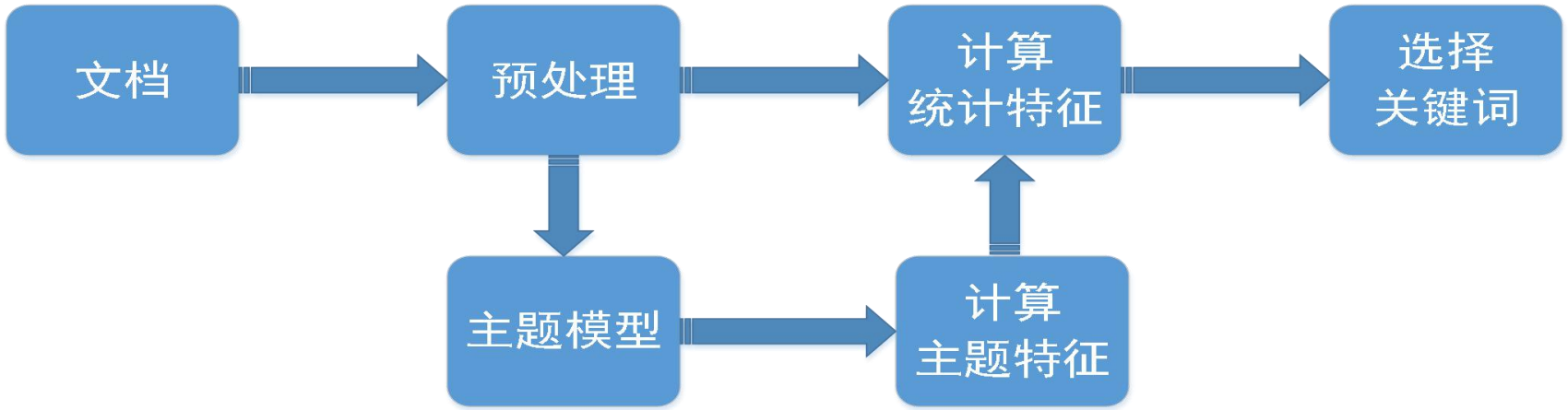
词性：词性标注工具直接获得

潜在语义：由 pLSA、LDA 等主题模型获得

近义词：利用同义词林、wiki 语料等方式扩展

……

结合主题、统计特征的关键词抽取

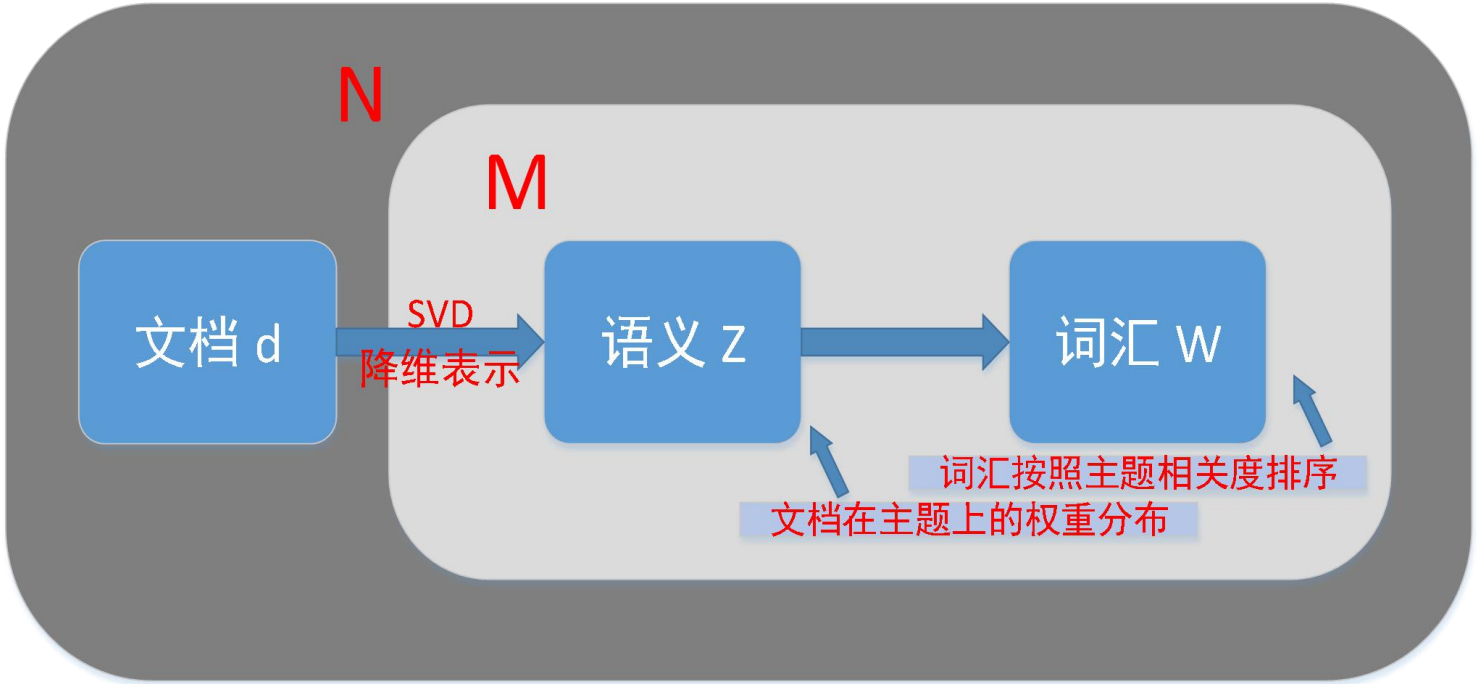


第二部分　模型介绍与选择

主题模型

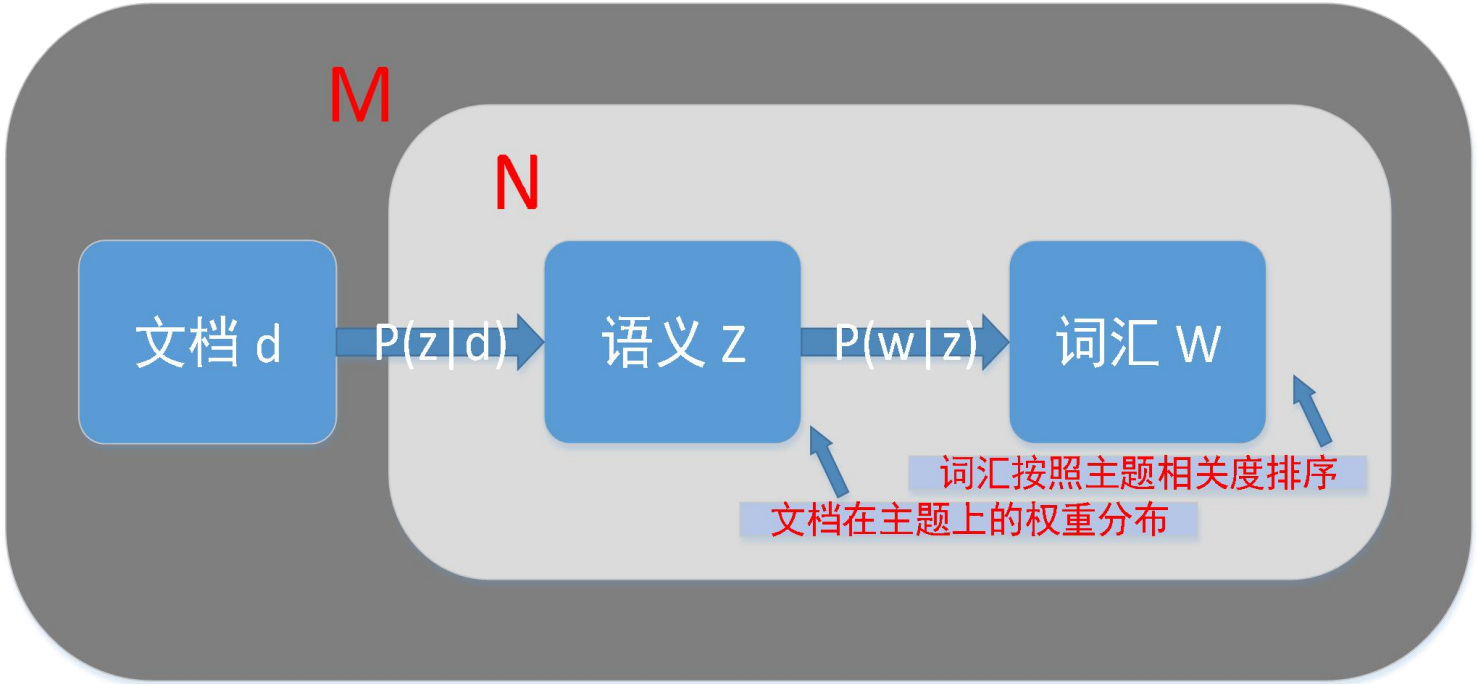
- 对文档隐含主题挖掘建模
- 主要思想：
  引入主题中间层
  主题是多个词语的条件概率分布
  实现文档到词汇的映射或表示

LSA(Latent Semantic Analysis)



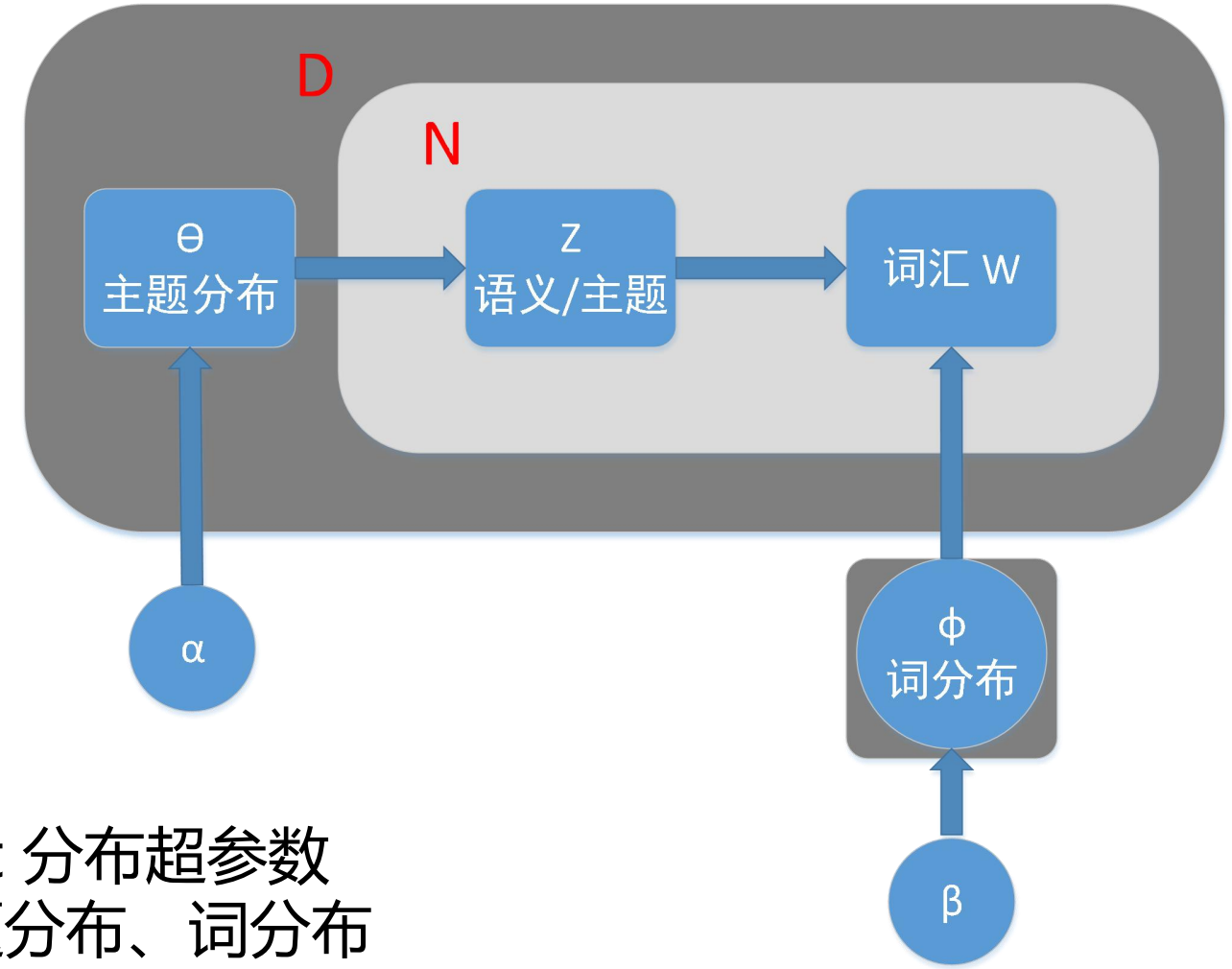
备注：不是一种概率生成模型

pLSA(probabilistic LSA)文档生成



缺陷：p(z|d)、p(w|z) 直接由数据估计得出

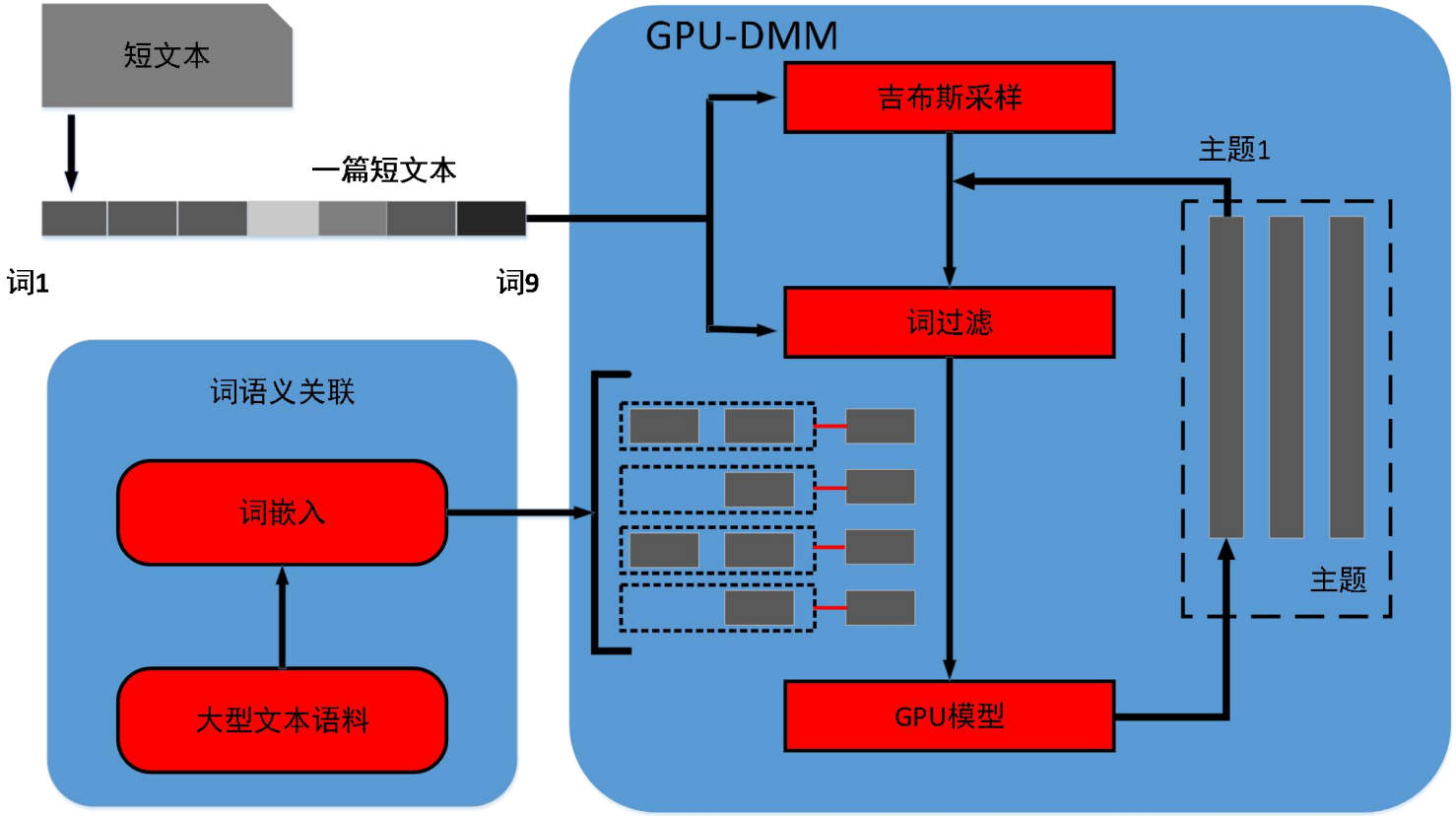
LDA(Latent Dirichlet Allocation)文档生成



α、β:

Dirichlet 分布超参数
产生主题分布、词分布

GPU-DMM结构



GPU-DMM提高了单词和其语义相关单词在该主题下共现概率

对信息稀疏做了补充

词过滤策略：指导模型的整个推导过程

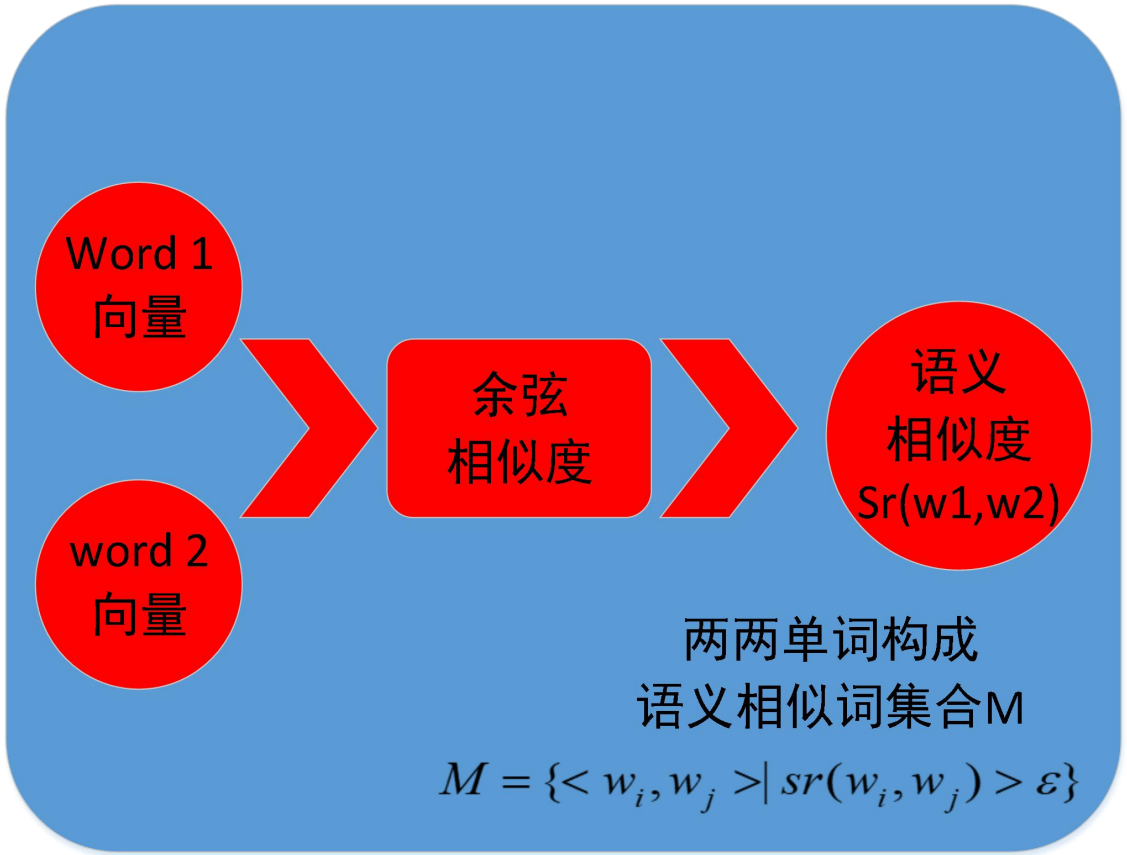
选出某主题下具有代表性的单词引入 GPU 策略，若所有单词都加强单词相似性关联，易产生噪音

第三部分　GPU-DMM深入介绍

词语义关联

- 词向量将单词映射到高维空间，在语义、词法上接近的单词在高维空间中更加接近，可以理解为，高维空间中，单词间的距离关系能带来更多语义知识
- 词嵌入技术
  - 向量空间中，每个词向量对应一个点，借助向量空间中度量距离的算法，直接得到单词间的相似性

如何从词嵌入得到语义近似词？？



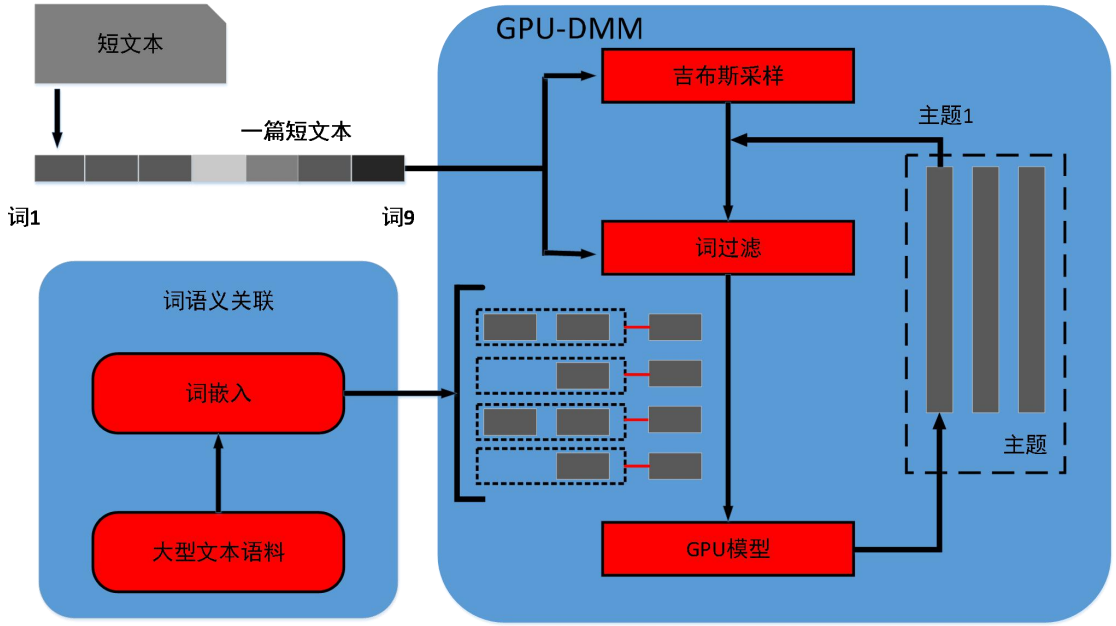
拿多少个相似词对到主题内？？

固定相似单词对的GPU促进量为一个常数
作者的策略：构建GPU促进量矩阵 A

$$\mathbb{A}_{w,w'} = \begin{cases} 1, & w = w' \\ \mu, & w' \in \mathbb{M}_w \; and \; w' \neq w \\ 0, & otherwise \end{cases}$$

利用GPU模型引入词嵌入

- GPU-DMM利用由词向量提供额外的语义关联信息，并借助GPU（一般化波利亚罐子模型）提高主题推导的效果
- 标准波利亚罐子模型 (罐子与彩色球的故事)
- 一般化波利亚罐子模型（Generalized Polya urn Uodel）
- 罐子 -> 主题，彩色球 -> 单词



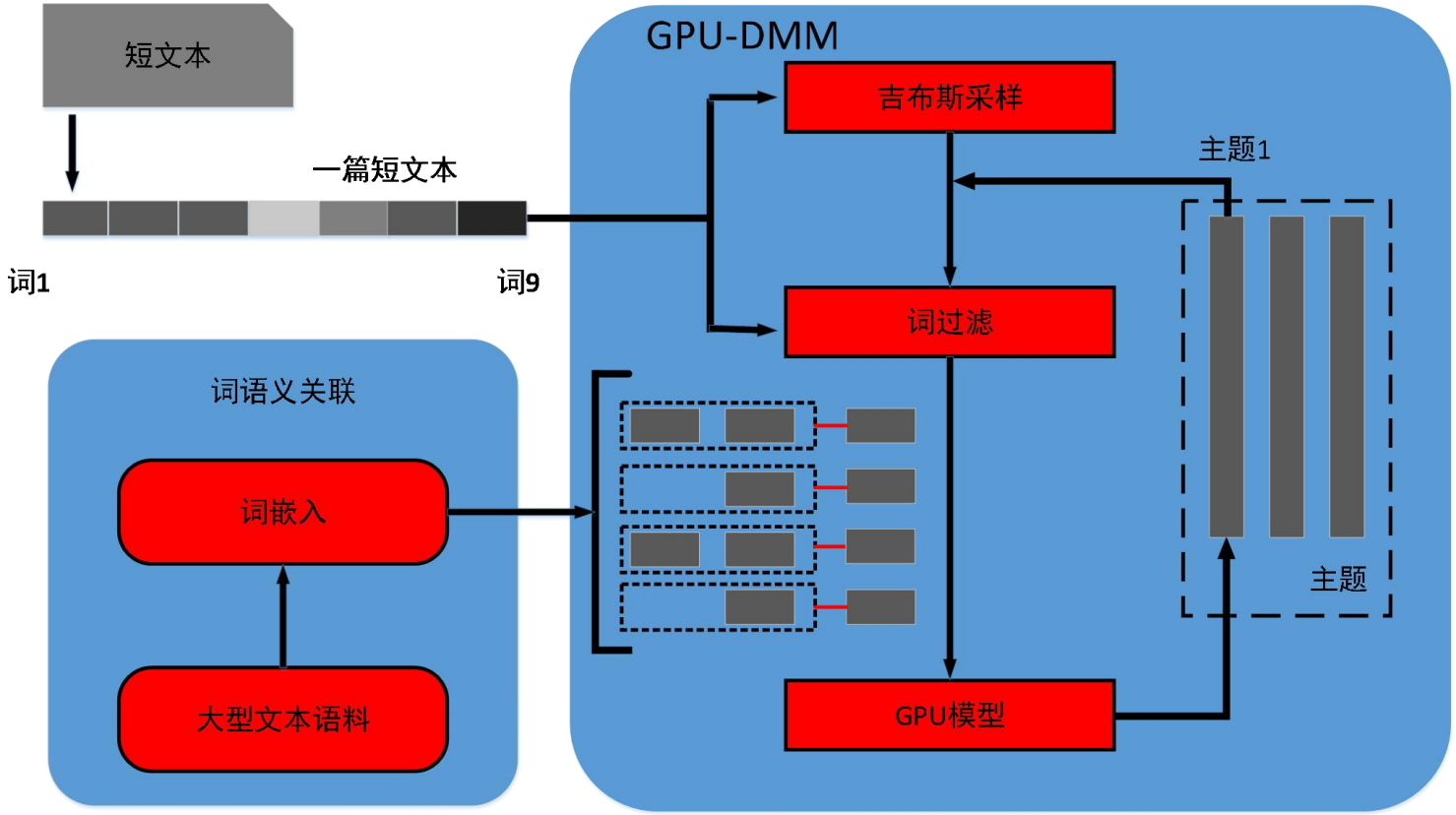
单词过滤

相似词集合M中，为避免每个词都做GPU操作，需要筛选机制
即：只对某个主题高度相关的单词，做GPU促进操作

主题导出

短文本经过GPU-DMM推导的条件概率，采样出一个主题，赋给该文档

如果当前单词是其主题下的高概率单词，该模型会借助GPU策略增强单词的词义相似词在该主题下的概率



第四部分　特征计算

自信息量

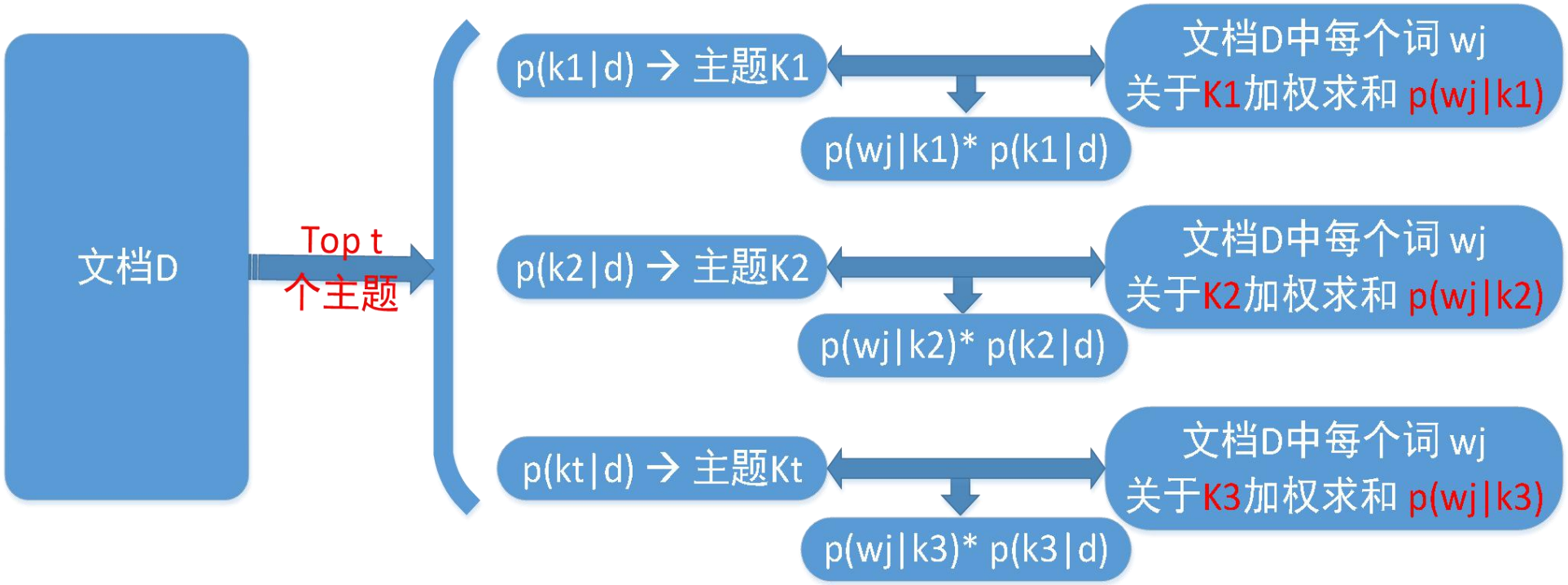
- 自信息量函数 $I(x) = -\log p(x)$

- 自信息量的两种含义
  - 事件 X 发生前：I(X) 表示 X 发生的不确定性
  - X 发生以后： I(X) 表示 X 能提供的信息量

- I(X) 应用于关键词抽取时
  - 某个词包含的信息量越大，说明该词越能代表该文档

- 词语的语义信息量
  - 利用词语和文档的主题信息来计算

主题特征计算



$$P_{topic}(d, w_j) = \sum_{i=1}^{t} P(w_j \mid k_i) \cdot P(k_i \mid d)$$

$$I = -\ln P_{topic}$$

结合统计特征计算

- TF-IDF 值、首次出现位置

$$p_{statistic}(d, w_j) = \frac{TF \times IDF}{firstOCC}$$

$$IDF = \log_2(\frac{D}{\#n(w_j \in d_i)} + 1)$$

$$p(d, w_j) = \alpha \times \frac{1}{p_{statistic}} + (1 - a) \times (-\ln\ p_{topic})$$

一波调参，a 约为 0.7 时，准确率最高

第四部分　待解决问题

未登录词识别

- 有研究显示：60%的分词错误都是未登录词切分错误
- 即使现阶段精确率达到90%以上的分词系统，切分精度依然不能保证。而文档的关键词往往是由新词或术语组成，新词、术语的低识别率，严重影响关键词的抽取效果

- 现有研究：
  - **有监督的**：如通过CRF、SVM、决策树等方法，抽取中文未登录词，利用抽取的未登录词对分词结果改进
  - **无监督的**：如合并高频词和信息熵等方法