基于潜在狄利克雷分配模型的语义搜索方法

张桐

CONTENTS

01

原理介绍

LDA

Latent Dirichlet allocation（潜在狄利克雷分配）是一种用于离散数据集合的概率生成模型。

常被用于在自然语言处理中对语料库进行建模，进而实现文档聚类、主题挖掘、情感分析、推荐系统等功能；在图像处理领域也有应用。

关键概念：

- 词语：一个离散数据
  例如 "北京"、"天安门"

- 主题：词语的概率分布
  例如 {"北京": 0.7, "天安门": 0.3}

- 文档：离散数据的集合
  例如 {"北京", "天安门"}

- 语料库：文档的集合
  例如 {{"北京", "天安门"}, {"北京"}}

LDA：基本思想

在 LDA 中，一篇文档的生成过程为：

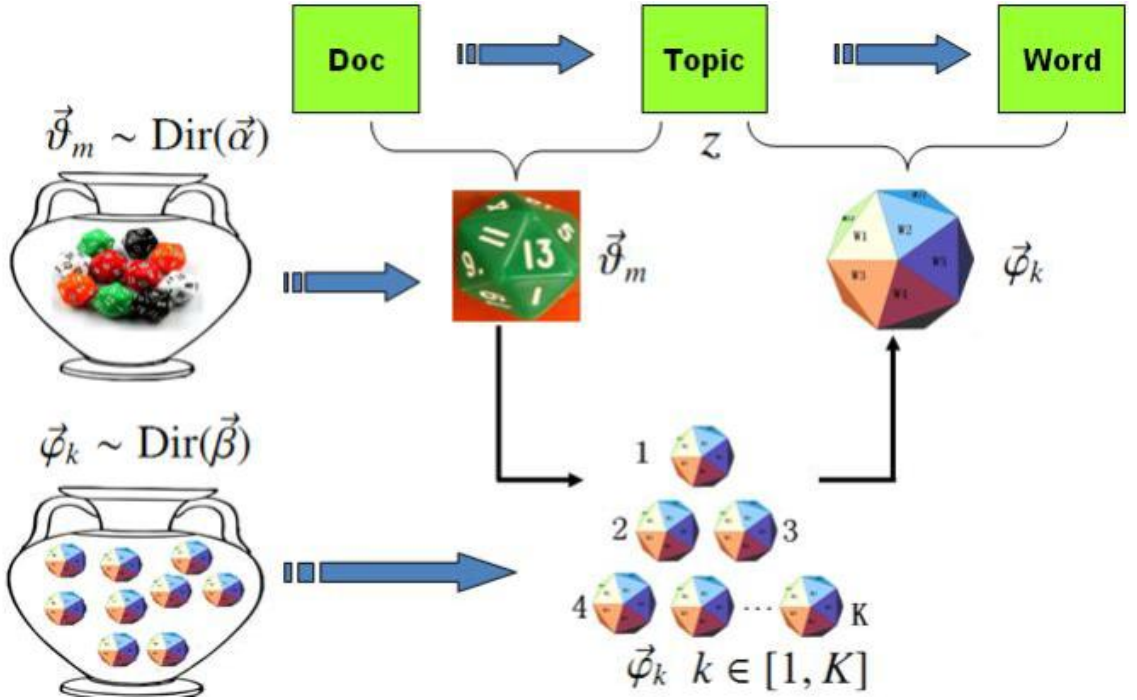
1. 根据参数 β 生成 K 个主题 ϕ
   例如 [
   ```
   {"经济": 0.5, "钱": 0.5},
   {"政治": 0.3, "国家": 0.7},
   {"体育": 0.8, "足球": 0.2}
   ]
   ```

2. 根据参数 α 生成主题编号的分布 θ
   例如 {0: 0.3, 1: 0.3, 2: 0.4}

3. 根据主题的分布 θ 生成主题编号 z
   例如 2

4. 根据主题 ϕ_z 生成一个词语
   例如 "足球"

5. 重复 3~4 步，直到达到要求词语数



LDA：基本思想

在 LDA 中，一篇文档的生成过程为：

1. 根据参数 β 生成 K 个主题 ϕ
   例如 [
   {"经济": 0.5, "钱": 0.5},
   {"政治": 0.3, "国家": 0.7},
   {"体育": 0.8, "足球": 0.2}
   ]

2. 根据参数 α 生成主题编号的分布 θ
   例如 {0: 0.3, 1: 0.3, 2: 0.4}

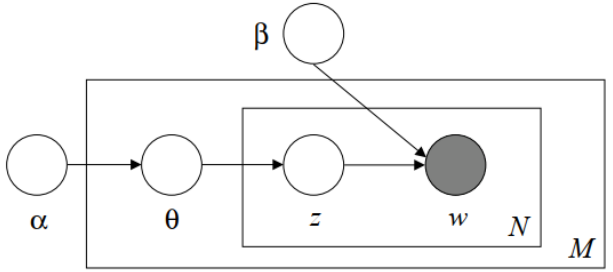
3. 根据主题的分布 θ 生成主题编号 z
   例如 2

4. 根据主题 ϕ_z 生成一个词语
   例如 "足球"

5. 重复 3~4 步，直到达到要求词语数

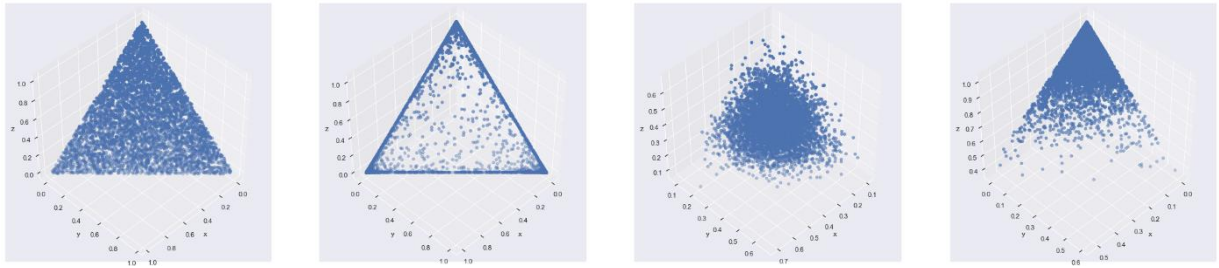
数学语言：

1. $\phi \sim Dirichlet(\beta)$
2. $\theta \sim Dirichlet(\alpha)$
3. $z \sim Multinomial(\theta)$
4. $w \sim Multinomial(\phi_z)$



Dirichlet 分布是 Multinomial 的共轭先验分布，即，在先验 Dirichlet 分布观察了服从 Multinomial 分布的随机变量 x 后，后验分布仍能保持 Dirichlet 分布，超参数 α 变为：

$$\alpha + \sum_{i=1}^{n} x_i$$



Gibbs 抽样

根据定义可以得到 $\theta, \vec{z}, \vec{w}$ 的联合分布：

$$p(\theta, \vec{z}, \vec{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

对 θ 积分可得 $\vec{w}, \vec{z}$ 的联合分布：

$$p\left(\vec{w}, \vec{z}\middle|\alpha, \beta\right) = \int p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta))\, d\theta$$

根据联合分布可以得到相应的 Gibbs 抽样公式：

$$p(z_i = k|\vec{z}_{\neg i}\ , \vec{w}) \propto \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}\left(n_{m,\neg i}^{(k)} + \alpha_k\right)} \cdot \frac{\left(n_{k,\neg i}^{(t)} + \beta_t\right)}{\sum_{t=1}^{V}\left(n_{k,\neg i}^{(t)} + \beta_t\right)}$$

其中 z_i 表示第 i 个词的主题，$\vec{z}_{\neg i}$ 表示其它词语的主题，$n_m^{(k)}$ 表示 m 文档中出现 k 主题的次数，$n_k^{(t)}$ 表示 k 主题中出现 t 词的次数。

相似度度量：主题分布-主题分布

- 余弦相似度

$$S_C(\theta_i, \theta_j) = \frac{\theta_i \cdot \theta_j}{\|\theta_i\|\|\theta_j\|}$$

- Kullback-Leibler 散度（相对熵）

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

分类分布：

$$D_{KL}(\theta_i||\theta_j) = \sum_{k=1}^{n} \theta_{ik} \log \frac{\theta_{ik}}{\theta_{jk}}$$

Jensen-Shannon 散度：

$$D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)$$

分类分布：

$$D_{JS}(\theta_i||\theta_j) = \frac{1}{2} D_{KL}(\theta_i||\frac{\theta_i + \theta_j}{2}) + \frac{1}{2} D_{KL}(\theta_j||\frac{\theta_i + \theta_j}{2})$$

相似度度量：文档-主题分布

除了使用主题分布来度量文档相似度外，我们也可以使用一篇文档的主题分布生成另一篇文档的概率来作为相似度的度量：

$$S_p = \prod_{w \in W_i} \sum_{k=1}^{n} p(w|z_k) p(z_k|W_j)$$

其中 W_i 是被生成的文档，W_j 则是另一篇文档。

02

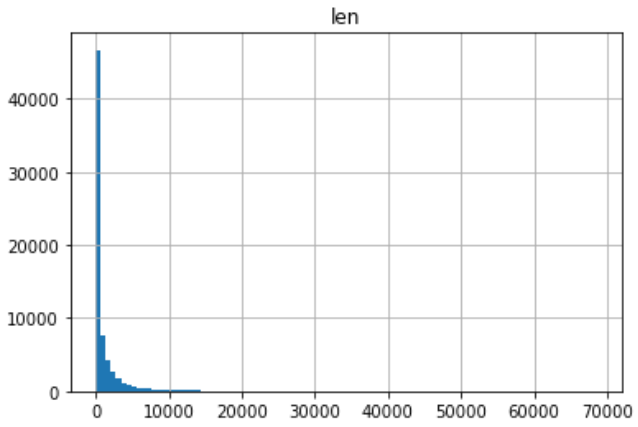
实验结果

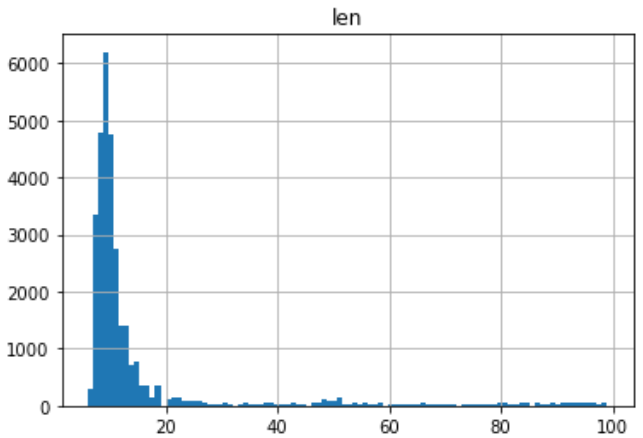
语料库

中文维基百科词条

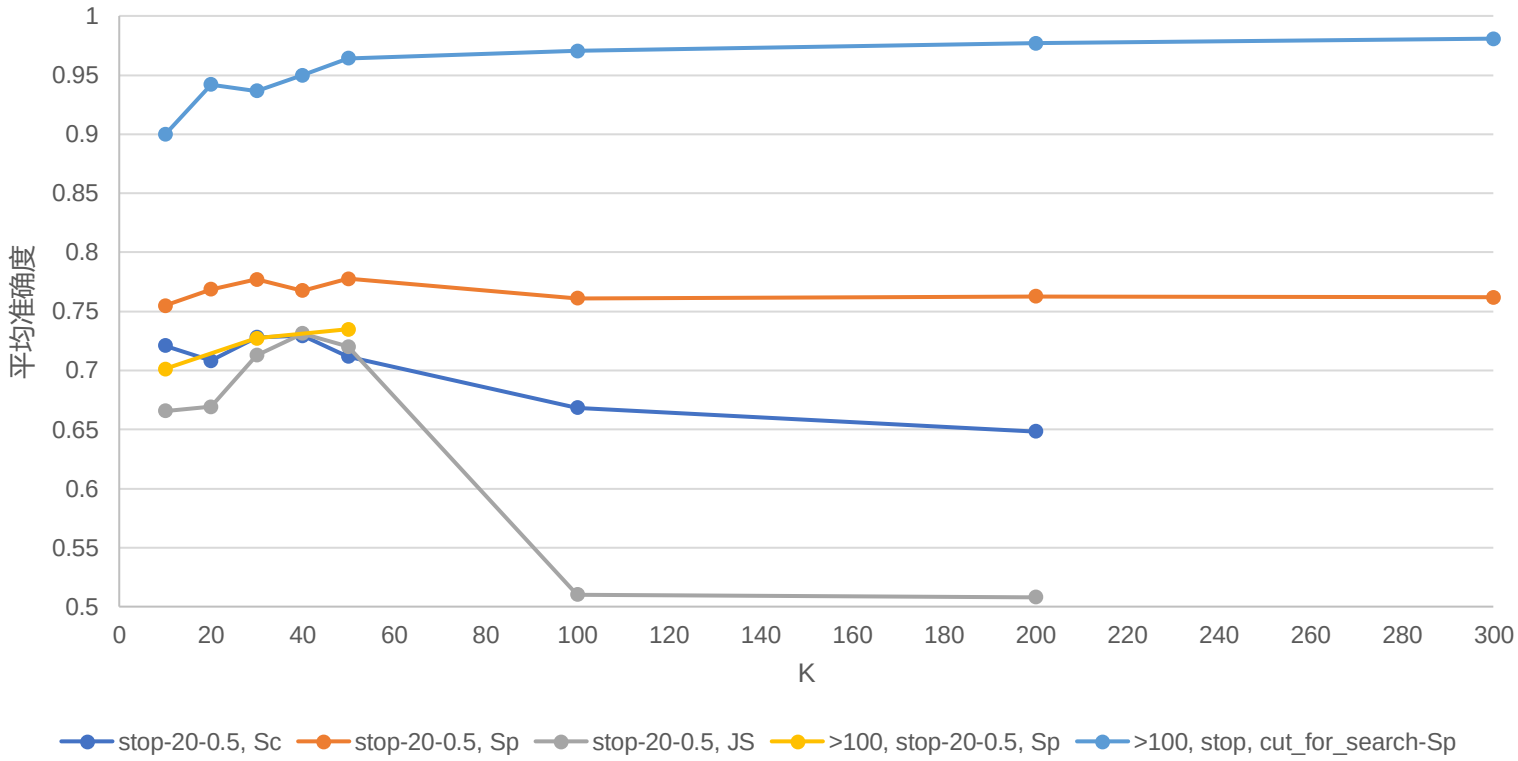
预处理：

1. 语料切分
2. XML 格式转换
3. 简繁转换
4. 分词



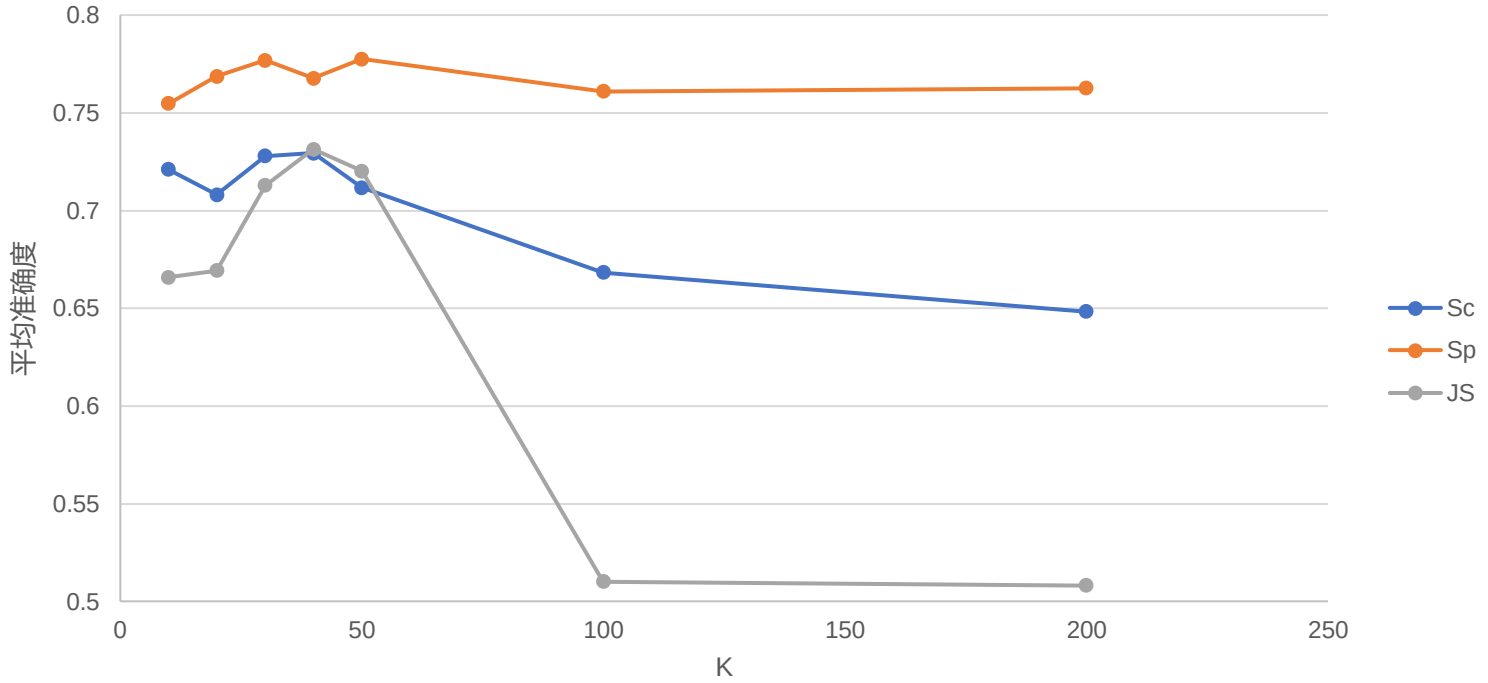


主题数



- 主题数会对语义搜索的准确度产生影响
- 在不同的相似度度量和分词策略下，最佳的主题数也会有所不同

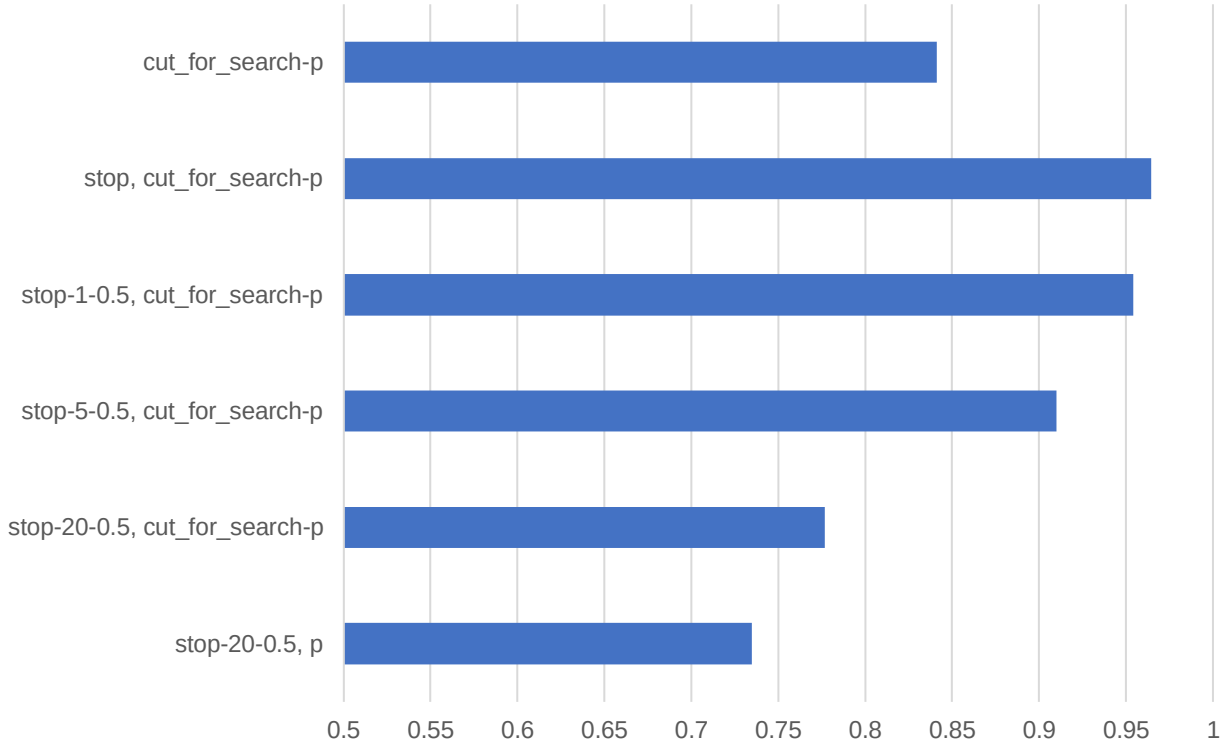
相似度度量



生成概率 > 余弦相似度 > Jensen-Shannon 散度

对于短文本来说，推断出的主题分布会丢失部分语义，导致匹配效果不佳，而使用生成概率作为相似度度量避免了对短文本进行主题分布推断

分词策略



- 过滤低频词会影响冷门文档的语义搜索效果
- 过滤高频词和停用词不影响语义搜索效果，但会导致训练同等准确度的模型耗费更多时间
- 查询文本中的错误中文分词会影响语义搜索效果，考虑所有分词路径能获得更好的效果

模型效果

平均准确度 0.9934

“Windows”

	score	title
909	0.082623	Windows XP
361	0.054110	Windows 98
364	0.046134	Microsoft Windows
518	0.042010	Windows 2.0
366	0.041702	Windows 2000
363	0.041036	Windows NT
359	0.040210	Windows 3.1x
360	0.039409	Windows 95
318	0.034390	Windows 1.0
317	0.033375	微软
841	0.029918	文件扩展名
384	0.029840	Microsoft Office
362	0.024675	蓝屏死机
473	0.018100	DirectX
40	0.014506	操作系统
250	0.013086	Delphi
365	0.011742	DOS
475	0.011730	OpenGL
300	0.010910	Visual Basic
746	0.009846	电视

“计算机”

	score	title
719	0.000080	计算机语言
145	0.000076	编程语言
50	0.000023	计算机程序
371	0.000022	NTFS
67	0.000021	计算语言学
486	0.000021	第一代编程语言
330	0.000020	阿基米德
49	0.000017	程序设计
40	0.000015	操作系统
31	0.000015	数据结构
391	0.000012	Forth
432	0.000012	信息管理系统
4	0.000011	计算机科学
975	0.000010	软件
246	0.000010	数学家
704	0.000009	算法
836	0.000009	聂耳
357	0.000009	多用户
273	0.000008	自然语言
30	0.000007	计算

“数学”

	score	title
246	0.036766	数学家
23	0.018464	心理学
57	0.013222	运算数学
136	0.012739	离散数学
0	0.012165	数学
432	0.011743	信息管理系统
990	0.010759	生物学家
285	0.010258	空间科学
138	0.010184	数理逻辑
14	0.007001	信息科学
825	0.005620	邪教
163	0.005544	严重急性呼吸系统综合症
287	0.005520	汉斯·莫拉维克
665	0.005188	家政学
425	0.005168	皮埃尔-西蒙·拉普拉斯
883	0.004730	允禵
741	0.004482	AutoCAD
592	0.004433	工程学
724	0.004380	氧
451	0.004241	弗兰西斯·培根

“中国”

	score	title
71	0.026731	华侨华人
302	0.026729	中国人
539	0.023569	知识产权
986	0.020306	约翰·哈比森
861	0.018618	洛杉矶
105	0.013442	黑龙江省
831	0.012939	中国朝代
370	0.012937	朝鲜的称号
96	0.012691	安徽省
107	0.012543	湖北省
115	0.012486	四川省
32	0.012248	中华人民共和国
149	0.012044	法国
791	0.011602	武则天
721	0.011364	华国锋
446	0.011356	四人帮
101	0.011350	中华人民共和国各省级行政区人口列表
389	0.011348	江泽民
60	0.011196	中华人民共和国历史
516	0.011147	胡锦涛

03

总结

总结

本实验研究了基于潜在狄利克雷分配模型的语义搜索方法的实际效果，对不同相似度度量、分词策略和主题数下的模型效果进行了对比和分析，提供了一种对中文语料库进行语义搜索的可行方法。

后续研究方向：

- 主题数的自动选取
- 与基于 BERT、word2vec 等词嵌入（word embedding）模型的语义搜索方法进行比较
- 将词嵌入模型与 LDA 模型结合，替换 LDA 所使用的整数词编码
- 增大实验语料规模

参考文献：

- Blei, David M. “Latent Dirichlet Allocation,” 2003, 30.
- Towne, W. Ben, Carolyn P. Rosé, and James D. Herbsleb. “Measuring Similarity Similarly: LDA and Human Perception.” *ACM Transactions on Intelligent Systems and Technology* 8, no. 1 (January 31, 2017): 1–28. https://doi.org/10.1145/2890510.
- 靳志辉. “LDA数学八卦”. 2013.

Thanks.

张桐

2022.12.19