Topic Modeling

2022/4/29

浙江大学 管理学院

SCHOOL OF MANAGEMENT
ZHEJIANG UNIVERSITY

Content

- What is topic modeling ?
- How to deal with topic modeling?
  - Coding
  - Latent Dirichlet Allocation
- Related work of topic modeling
  - Demographic Information Topic Modeling
  - Decomposition (in a semantic way)

1 What is topic modeling ?

- In the **social sciences**, coding is an analytical process in which data, in both quantitative form (such as questionnaires results) or qualitative form (such as interview transcripts) are categorized to facilitate analysis. Prior to coding, an annotation scheme is defined. The coding scheme ensures that the codes are added consistently across the data set and allows for verification of previously tagged data.

- In **ML** and **NLP**, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic models are referred to as probabilistic topic models, which refers to statistical algorithms for discovering the latent semantic structures of an extensive text body.

2 How to deal with topic modeling?

Coding: https://en.wikipedia.org/wiki/Coding_(social_sciences)

LDA(Latent Dirichlet Allocation): https://zhuanlan.zhihu.com/p/31470216

- Dirichlet Distribution AND Beta Distribution
- MCMC AND Gibbs Sampling

So, can we explain LDA in a brief way?

2.1 Brief LDA

In fact, though LDA is the essential part of topic modeling, the mathematic definition of LDA is not necessary in this process at all. Learning by code is a simpler way to have an intuitive understanding of LDA. Here we use ‘Gensim’ (one of the most famous Python package to deal with topic modeling) as example.

```
[1]stopwords = codecs.open('stopwords.txt','r',encoding='utf8').readlines()
```

Define Stopwords

```
[2]dictionary = Dictionary(train_set)
[3]corpus = [ dictionary.doc2bow(text) for text in train_set]
```

What is it?

Training Text

Bag-Of-Word

```
[4]lda = LdaModel(corpus=corpus, id2word=dictionary, num_topics=20)
```

Model Genrator

2.2 Bag-of-Word

Example in Wikipedia:

(1) John likes to watch movies. Mary likes movies too.

(2) Mary also likes to watch football games.

→ “John”, “likes”, “to”, “watch”, “movies”, “Mary”, “likes”, “movies”, “too”, "Mary", "also", "likes", "to", "watch", "football", "games”

→ BoW1 = {"John":1, "likes":2, "to":1, "watch":1, "movies":2, "Mary":1, "too":1};

BoW2 = {"Mary":1, "also":1, "likes":1, "to":1, "watch":1, "football":1, "games":1};

2.2 Bag-of-Word

	John	Likes	To	Watch	Movies	Mary	…
Text 1	1	2	1	1	2	1	
Text 2	0	1	1	1		1	
…							
Total	**1**	**3**	**2**	**2**	**2**	**2**	

Which word can represent the whole text?

- John? Likes? (Actually, John!)

And why? → tf-idf: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

2.3 Model Generator

What happens in model generator?

```
[4]lda = LdaModel(corpus=corpus, id2word=dictionary, num_topics=20)
```

2 Matrix in model

- Topic-Word Matrix

	Topic 1	Topic 2	Topic 3	…
Word 1	freq(1, 1)	freq(1, 2)	freq(1, 3)	
Word 2	e.g. 0.96	e.g. 0.04	e.g.0.13	
…				

- Topic-Text Matrix

	Topic 1	Topic 2	Topic 3	…
Text 1	freq(1, 1)	freq(1, 2)	freq(1, 3)	
Text 2	e.g. 0.45	e.g. 0.16	e.g.0.52	
…				

2.3 Model Generator

Here is an example of topic vectors.

```
topic #0 (0.050): 0.024*主编 + 0.019*时尚 + 0.018*网友 + 0.016*女人 + 0.015*编辑 + 0.015*评论 + 0.013*点击 + 0.012*邮箱 + 0.011*打造 + 0.011*丽丽
topic #1 (0.050): 0.022*汽车 + 0.014*车型 + 0.010*发动机 + 0.009*设计 + 0.006*系统 + 0.006*万元 + 0.005*动力 + 0.005*配置 + 0.005*轿车 + 0.005*上市
topic #2 (0.050): 0.067*基金 + 0.037*投资 + 0.032*投资者 + 0.019*风险 + 0.013*股票 + 0.012*证券 + 0.011*机构 + 0.011*建议 + 0.010*参考 + 0.010*操作
topic #3 (0.050): 0.011*合作 + 0.007*飞机 + 0.007*咨询 + 0.006*部队 + 0.006*酒店 + 0.005*婚礼 + 0.005*直升机 + 0.004*频道 + 0.004*报道 + 0.004*刘嘉玲
topic #4 (0.050): 0.022*比赛 + 0.007*球员 + 0.007*球队 + 0.006*时间 + 0.006*体育 + 0.006*北京 + 0.005*中国队 + 0.005*对手 + 0.005*欧洲杯 + 0.005*奥运会
topic #5 (0.050): 0.010*位置 + 0.010*首页 + 0.009*腾讯 + 0.009*正文 + 0.008*所在 + 0.008*短期 + 0.007*日期 + 0.007*趋势 + 0.006*今日 + 0.006*详细
topic #6 (0.050): 0.017*工作 + 0.013*发展 + 0.008*社会 + 0.007*建设 + 0.006*管理 + 0.005*国家 + 0.005*能力 + 0.004*一个 + 0.004*制度 + 0.004*组织
topic #7 (0.050): 0.021*一个 + 0.004*电影 + 0.004*记者 + 0.004*生活 + 0.003*喜欢 + 0.003*一种 + 0.003*导演 + 0.003*事情 + 0.003*希望 + 0.003*朋友
topic #8 (0.050): 0.036*公司 + 0.011*投资 + 0.011*股份 + 0.011*上市公司 + 0.011*亿元 + 0.011*集团 + 0.010*上市 + 0.009*资产 + 0.008*股权 + 0.008*证券
topic #9 (0.050): 0.013*中国 + 0.010*活动 + 0.009*灾区 + 0.008*北京 + 0.007*现场 + 0.006*香港 + 0.006*地震 + 0.006*上海 + 0.006*支持 + 0.005*新浪
topic #10 (0.050): 0.047*中国 + 0.031*美国 + 0.011*日本 + 0.010*国际 + 0.010*越南 + 0.009*国家 + 0.008*经济 + 0.007*世界 + 0.006*报道 + 0.005*全球
topic #11 (0.050): 0.040*奥运 + 0.034*传递 + 0.023*火炬 + 0.017*圣火 + 0.015*北京 + 0.015*奥运会 + 0.014*火炬手 + 0.009*记者 + 0.009*重庆 + 0.008*北京奥运
topic #12 (0.050): 0.040*银行 + 0.022*点击 + 0.015*发行 + 0.012*香港 + 0.012*贷款 + 0.012*选择 + 0.011*论坛 + 0.010*速度 + 0.010*即可 + 0.009*按钮
topic #13 (0.050): 0.015*价格 + 0.010*企业 + 0.009*增长 + 0.007*房地产 + 0.007*影响 + 0.006*经济 + 0.006*市场 + 0.006*政策 + 0.005*项目 + 0.005*增加
topic #14 (0.050): 0.027*信息 + 0.016*地震 + 0.012*企业 + 0.012*服务 + 0.011*记者 + 0.010*灾区 + 0.009*社会 + 0.009*分类 + 0.009*新闻 + 0.008*网评
topic #15 (0.050): 0.018*考生 + 0.014*高考 + 0.012*学生 + 0.010*学校 + 0.009*考试 + 0.008*专业 + 0.008*孩子 + 0.007*记者 + 0.006*信息 + 0.006*志愿
topic #16 (0.050): 0.052*公司 + 0.029*有限公司 + 0.021*股份 + 0.013*董事会 + 0.013*公告 + 0.013*股东 + 0.010*年度 + 0.009*议案 + 0.009*本次 + 0.009*董事
topic #17 (0.050): 0.025*市场 + 0.015*美元 + 0.010*上涨 + 0.009*下跌 + 0.007*指数 + 0.007*经济 + 0.007*预期 + 0.006*价格 + 0.006*通胀 + 0.006*反弹
topic #18 (0.050): 0.016*企业 + 0.014*公司 + 0.013*市场 + 0.013*中国 + 0.012*产品 + 0.009*发展 + 0.009*生产 + 0.007*国内 + 0.007*技术 + 0.006*品牌
topic #19 (0.050): 0.012*治疗 + 0.008*医院 + 0.007*女性 + 0.006*国际 + 0.004*平衡 + 0.004*康复 + 0.004*转会 + 0.004*公司 + 0.004*荷尔蒙 + 0.004*回到
```

2.4 Related Paper

	Topic 1	Topic 2	Topic 3	...
Word 1	freq(1, 1)	freq(1, 2)	freq(1, 3)	
Word 2	e.g. 0.96	e.g. 0.04	e.g.0.13	
...				

表 1：1982-1989 年养育文本主题模型结果（最重要的 11 个关键词）

Topic 1 科学喂养	Topic 2 疾病治疗	Topic 3 生活习惯
婴儿	孩子	孩子
食物	小儿	父母
维生素	治疗	家长
小儿	皮肤	儿童
营养	家长	妈妈
儿童	患儿	幼儿
孩子	症状	时间
发育	感染	吃饭
食品	儿童	活动
蛋白质	疾病	习惯
牛奶	医院	生活

表 2：主题-文档数分布

主 题	Topic 1 科学喂养	Topic 2 疾病治疗	Topic 3 生活习惯	总 计
文档数	180	247	178	605

表 3：二次主题模型结果（最重要的 11 个关键词）

“科学喂养”类				
1. 食物营养	2. 生长发育	3. 婴儿喂哺	4. 饮食知识	5. 营养性疾病
孩子	婴儿	婴儿	小儿	维生素
儿童	动作	母亲	孩子	儿童
食物	孩子	母乳	食物	贫血
维生素	发育	母乳喂养	饮食	铁
营养	手	奶粉	牛奶	蛋白质
发育	活动	营养	消化	食品
蛋白质	练习	乳汁	家长	孩子
作用	声音	奶瓶	食欲	毫克
脂肪	训练	喂哺	便秘	吸收
功能	厘米	免疫	断奶	钙
脑细胞	体重	乳儿糕	冷饮	佝偻病

“疾病治疗”类			“生活习惯”类			
1. 常见疾病	2. 婴幼儿护理	3. 发育疾病	1. 生活规律	2. 用眼卫生	3. 情绪心理	4. 饮食习惯
治疗	皮肤	孩子	孩子	孩子	孩子	孩子
患儿	婴儿	脊柱	时间	妈妈	口吃	父母
家长	感染	发育	培养	视力	家长	家长
症状	家长	牙齿	动作	家长	父母	吃饭
医生	腹泻	儿童	生活	眼睛	情绪	食欲
发作	婴幼儿	骨折	活动	爸爸	治疗	习惯
疾病	尿布	家长	习惯	看电视	心理	健康
咳嗽	细菌	损伤	父母	噪音	发育	方法
发现	消毒	畸形	睡觉	音乐	遗尿	养成
肺炎	烫伤	发现	吃饭	近视	精神	吃零食
发热	洗澡	手术	学习	听力	紧张	厌食

[1]陈倩. 社会转型中的大众媒介育儿话语变迁_陈倩(1)[D]. 上海大学, 2019.

3 Related work of topic modeling

- Demographic Information Topic Modeling
  - Jiang, S., Qian, X., Shen, J., Fu, Y., & Mei, T. (2015). Author topic model-based collaborative filtering for personalized POI recommendations. IEEE transactions on multimedia, 17(6), 907-918.
  - Trusov, M., Ma, L., & Jamal, Z. (2016). Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. Marketing Science, 35(3), 405-426.

- Decomposition (in a semantic way)
  - Scale-invariant feature transform (SIFT)

https://en.wikipedia.org/wiki/Bag-of-words_model_in_computer_vision