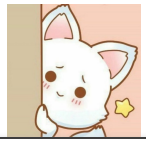
# 想让推荐和搜索引擎更聪明？基于知识图谱的篇章标签生成

夕小瑶的卖萌屋 1月5日

以下文章来源于丁香园大数据 ，作者丁香园大数据NLP

**丁香园大数据**

丁香园大数据是以数据为驱动、以产品为导向、致力于数据挖掘和商业探索的部门。她采用先进的大数据平台，以前沿的...



一只小狐狸带你解锁NLP/ML/DL秘籍

正文来源：丁香园大数据NLP

老板～我们的推荐系统笨笨的

你怎么对文档处理的这么糙！抽个关键词就应付过去了？

啊啊啊我错惹，那那，不用关键词用什么呢？

知识图谱用上了没？

概念词知道不？9012年了知道么！


## 前言

**篇章的标签生成是NLP领域的一项基础任务，目的是对文本更好地结构化，筛选重要的关键词，概括文本的中心语义。**因此，我们探索了一套标签生成流程，其中除了应用了已有的信息抽取技术之外，还将医疗知识图谱结构，实体显著性判断，concept抽取融入模型，实现业务增长。

关于标签生成，优化的方法大致有两种思路，第一种是在拥有一个较为完备的知识图谱后，如何使用知识图谱去指导标签抽取过程保持语义上的一致。举个栗子，比如通过词分布的分析，某篇文章的主题被定为在“**妇科**”相关疾病上，那么“**骨科**”的实体词就会被避免作为标签被抽出。这种思路在业界多以LDA的无监督打标签算法为主，利用知识表示向量、知识图谱结构或者其他统计信息对LDA模型进行改进，输出的结果为原文出现过的实体词，以下我们将它称之为 主题语义连贯的词分布标签方法；

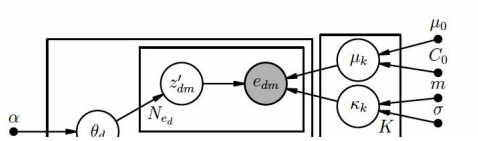
另一种思路是在知识图谱上做文章，比如专业的医学知识图谱上的实体词够精确，但有时由于词本身的含义不够泛化，并不适用于文章的表示，举个栗子，比如“**HP**”、“**胃镜**”、“**三联疗法**”这几个词的确贴合消化内科的主题，但是它没有“**幽门螺杆菌的治疗方法**”这样更加泛化的标签词来的直观，后者包含了更多的信息量，且更具**可解释性**。这方向需要结合更多的NLP技术，包括在业务场景中挖掘优质的concept短语，构建concept短语与实体词的taxonomy，利用文本子图中心度测量、随机游走路径、词频共现等做encoding，以LDA作为抽取器完成标签工作，以下我们将它为 Concept挖掘的标签方法 。

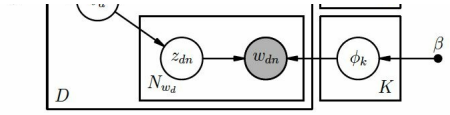
对这两种思路，我们调研了已有的相关研究，利用医疗知识图谱和医学垂直领域业务数据做了标签生成的尝试。

## 主题语义连贯的词分布主题模型

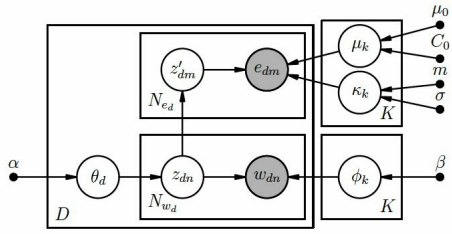
### 《Incorporating Knowledge Graph Embeddings into Topic Modeling》

概率主题模型可用于从文档集合中提取低维主题。然而，以往的模型往往产生无法解释的主题。近年来，已有许多基于知识的主题模型被提出，但它们不能很好的处理知识图中的三元组，大部分以must-link形式，或直接利用图谱中的上层概念，无法在向量空间中量化。本文将知识表示嵌入到到LDA中，将潜在的Dirichlet分配（一种广泛使用的主题模型）与实体向量编码的知识相结合，来提高了主题语义的一致性和连贯性。本文主要在两个以往研究（CI-LDA和Corr-LDA）上做了改进，上图为linkKGLDA，下图为corrKGLDA：






(a) KGE-LDA(a)




(b) KGE-LDA(b)

两个模型的不同之处在于，前者为条件独立，后者为条件相关。具体的改进如下：

2. For each topic $k$ in $1 \ldots K$:
   (a) Draw $\phi_k \sim \text{Dir}(\beta)$.
   (b) Draw $\mu_k \sim \text{vMF}(\mu_0, C_0)$.
   (c) Draw $\kappa_k \sim \text{logNormal}(m, \sigma^2)$.

entity embedding $e_{dm} \sim \text{vMF}(\mu_{z'_{dm}}, \kappa_{z'_{dm}})$.

由于一些知识表示（TransE）是unit sphere，因此使用von Mises Fisher（VMF）分布对其进行建模。vmf分布被广泛用于模拟此类定向数据。此外，使用vmf分布代替多元高斯分布，可以更有效地进行推断。与传统LDA模型相比，增加一组参数：(μk, κk) ，主题k的vMF分布；以及edm，即文档中实体的知识表示向量。

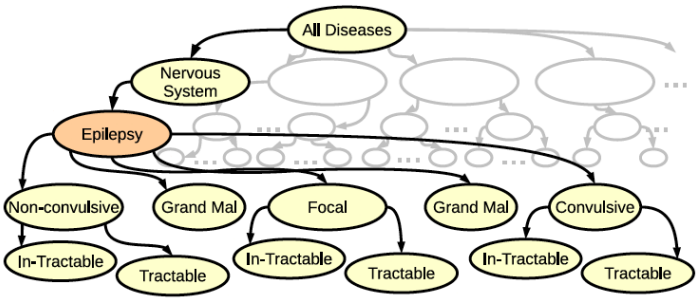
## 实现Concept挖掘的标签方法

目前的大多数concept标签方案，多是基于一定的统计数据，如：词对共现频数，词对覆盖率等。这些方法依赖业务场景下的query log，或者也可利用知识图谱结构进行层次聚类，中心化，路径预测等方式进行。下面将一一介绍相关研究。

### 《Graph-Sparse LDA: A Topic Model with Structured Sparsity》

本文引入了图稀疏LDA，这是一种分层主题模型，它使用单词之间关系的知识（例如：本体编码）。在模型中，主题被一些潜在的概念词概括，这些潜在概念词来自观察词在本体中的图。

GS-LDA在标准LDA模型中引入了一个额外的层次结构层：主题不是分布在观察到的词上，而是分布在概念词上，然后通过由词汇结构通知的噪声过程生成观察到的词。




例如上图所示，“癫痫”是一个很好的概念词，可以概括出子类癫痫。如果患者患有癫痫也可以解释“中枢神经系统紊乱”甚至“疾病”。

利用词典用的词或者实体总结文本时，通常都非常具体，而使用概念词总结一段文本，不仅可以说明具体的语义，同时能挖掘到更上层或者相似主题的语义。例如："抗病毒药物"和"抗逆转录病毒"，一个词和它的邻居词，可以被认为产生自一个核心概念。Graph-Sparse LDA模型假设一个主题有一组稀疏的概念词组成，或是后代，或是祖先。最后定义如下过程：

$$\pi_k \sim \text{IBP-Stick}(\gamma_B) \quad (5)$$
$$\rho_v \sim \text{Beta}(\gamma_A/V, 1) \quad (6)$$
$$\bar{B}_{nk} \mid \pi_k \sim \text{Bernoulli}(\pi_k) \quad (7)$$
$$\bar{A}_{kv} \mid \rho_v \sim \text{Bernoulli}(\rho_v) \quad (8)$$
$$B_n \mid \bar{B}_n \sim \text{Dirichlet}(\bar{B}_n \odot \alpha_B 1_K) \quad (9)$$
$$A_k \mid \bar{A}_k \sim \text{Dirichlet}(\bar{A}_k \odot \alpha_A 1_V) \quad (10)$$
$$z_{in} \mid B_n \sim \text{Discrete}(B_n) \quad (11)$$
$$\tilde{w}_{in} \mid z_{in}, \{A_k\} \sim \text{Discrete}(A_{z_{in}}) \quad (12)$$
$$P_v \sim \text{Dirichlet}(\mathcal{O}_v \odot \alpha_P 1_V) \quad (13)$$
$$w_{in} \mid \tilde{w}_{in}, P \sim \text{Discrete}(P_{\tilde{w}_{in}}) \quad (14)$$

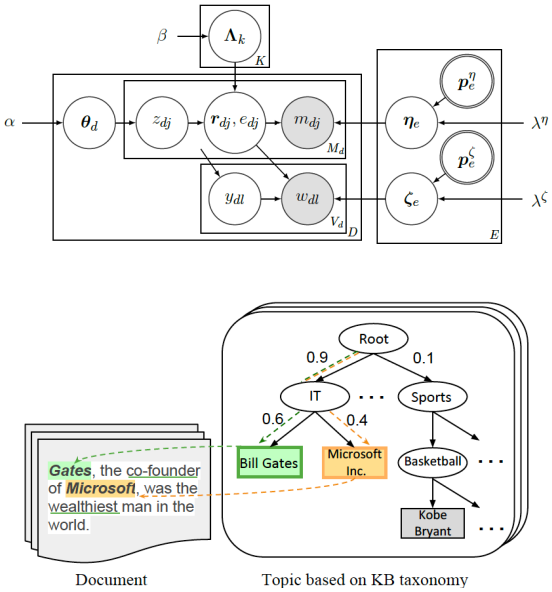
推导了一个B、B-、A、A-和P（以及添加和删除主题）的 blocked-Gibbs sampler。然而，单是吉布斯抽样并不能使主题概念词矩阵中的稀疏性足够快。混合速度很慢，因为阻塞的吉布斯取样器设置Akw~=0的唯一时间是没有w~计数分配给任何文档中的主题K时。当有海量文档时，不太可能达到零计数，采样器稀疏主题概念词矩阵A的速度会很慢。因此本文引入了一个MH procedure，通过在A和P上的 joint moves，鼓励主题概念词矩阵A向更大稀疏的方向移动。分布如下：

$$a_{MH} = 1 \wedge \frac{p(X \mid B, A', P')\, p(A')\, p(P')\, Q(A, P \mid A', P')}{p(X \mid B, A, P)\, p(A)\, p(P)\, Q(A', P' \mid A, P)}$$

首先，对A'进行智能分割合并移动。其次，试图通过提出一个P'来保持似然函数尽可能恒定：AP=A'P'。这样，先验p(A)和p(P)将对移动产生很大的影响。

### 《Grounding Topic Models with Knowledge Bases》

这篇工作认为尽管最近的研究试图利用各种知识源来改进主题建模，但它们要么承担着仅将主题表示为单词或短语分布，要么通过将主题与预先定义的知识库（知识库）实体进行一对一的绑定，建立主题模型，牺牲了主题建模的灵活性。因此提出了一种基于taxnomy层次结构随机游走特征的LDA，目的在于将taxnomy的语义和结构化特征全部考虑进来。



与以往LDA不同的是，模型中加入了四个新的变量，其中rdj代表随机游走路径，edj代表taxonomy中的概念，mdj代表文档中的实体，ydl为单词index。从根节点顶层概念词c0开始，通过随机游走筛选子节点。过程结束直到到达叶子节点。因此这个随机游走给每一个entity（概念）分配了一个从根到叶子的路径。基于主题k，可以计算到达每个实体的随机行走的概率，从而获得主题k下实体的分布。同样，对于每个类别节点c，可以计算一个概率kc，表示c被包含在随机行走路径中的可能性。

除了随机游走得到的结构特征，本文同样利用了维基百科的page数据，得到实体，单词，概念之间的共现频数，作为先验信息。大文本语料库和知识库的推理是复杂的，为了保证实践中的效率，文章提出需要注意以下几个方面：

**（a）**所有实体的路径总数可能非常大，使得随机游走公式的计算非常庞大。因此，本文使用命名实体字典为每个文档选择候选实体，在采样时只考虑这些实体的路径。实验表明，该近似方法对建模性能的影响可以忽略不计，同时大大降低了采样的复杂度，使推理成为现实。

**（b）**通过修剪低级的具体类别节点（其最短的根到节点路径长度超过阈值），进一步减少层次深度。作者发现这样一个“粗糙”的实体本体足以满足需求。

**（c）**为了计算路径的概率，使用动态规划来避免冗余计算。

**（d）**初始化实体和路径分配以确保良好的起点。

### 《Modeling Documents by Combining Semantic Concepts with Unsupervised Statistical Learning》

以往的研究中，或是只用文档中的词来描述主题，或是在taxnomy中找到合适的concept描述主题，本文虽然只是在传统LDA模型上做了很小的改动，但是它可以实现文档词分布和taxnomy concept共存的主题表达形式。将文档中的所有单词（不仅仅是实体）映射到一组本体概念上，学习单词和概念的概率模型，并且使用完全无监督的方法，而无需任何监督标记。

$$p(w_i|d) = \sum_{j=1}^{C} p(w_i|c_j)p(c_j|d).$$

将把这种模型称为概念模型。在概念模型中，属于概念的词由人类先验地（例如，作为本体的一部分）定义，并且仅限于（通常）总体词汇的一小部分。相反，在主题模型中，词汇表中的所有单词都可以与任何特定主题关联，但具有不同的概率。在上面的公式中，概念模型的未知参数是单词概念概率p（wi|cj）和概念文档概率p（cj|d）。作者的目标（在主题模型中）是从适当的语料库中估计这些。例如，注意概率p（cj|d）可以解决前面提到的标记问题，因为每个这样的分布都告诉我们文档d表示的概念cj的混合。

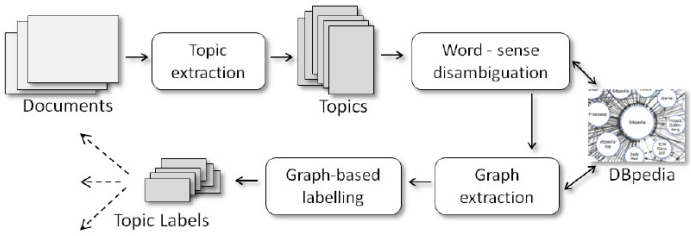
使用吉布斯抽样将概念分配给文档中的单词，使用与在主题模型相同的抽样方程，但是有一个额外的约束，即一个单词只能分配给它在本体中关联的概念。除了约束外，学习算法与主题模型的标准学习完全相同，最终的结果是语料库中的每个词都被赋予了本体中的一个概念。反过来，这些分配可以直接估计上面等式中的某些项。为了估计特定概念cj的p（wi|cj），我们通过抽样算法判断语料库中有多少单词可以分配给概念cj，并将这些计数标准化，以得到概率分布p（wi|cj）。为了估计特定文档d的p（cj|d），计算每个概念分配给文档d中单词的次数，然后再次规范化和平滑计算以获得p（cj|d）。下表显示了一组单词的学习概率（按概率排序）的例子。

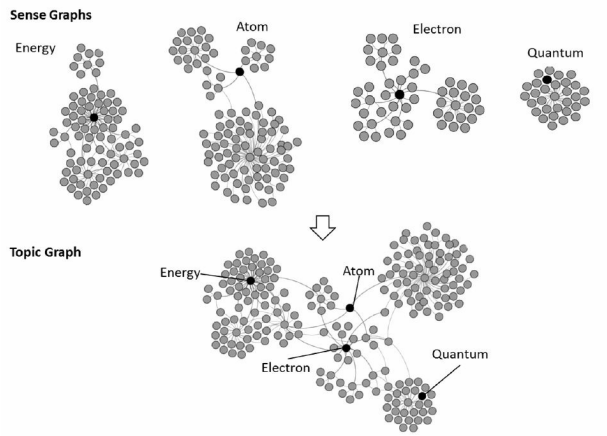
| FARMING & FORESTRY | | EARTH & OUTER SPACE | |
|---|---|---|---|
| crops | (0.135) | earth | (0.226) |
| plant | (0.076) | sky | (0.107) |
| grow | (0.050) | space | (0.082) |
| land | (0.040) | sun | (0.066) |
| fertilizers | (0.038) | scientists | (0.046) |
| soil | (0.037) | planets | (0.033) |
| earth | (0.034) | universe | (0.033) |
| farming | (0.034) | stars | (0.032) |

**Table 3.** Two example concepts from the CIDE thesaurus

**《Unsupervised Graph-based Topic Labelling using DBpedia》**

这是一个完全将LDA作为抽取功能组件的模型，topic labeling过程完全基于聚类和图的中心化操作。




可以看到，每个实体的语义图只能表示一种非常宽泛的概念，并不能体现各个实体概念之间的关系，直接作为标签会使每个概念都孤立起来，文本的语义不仅不一致，还会非常离散。相反本文方法是建立一个由多个实体子图构成的主题图，从中可以分析大图中每个节点对主题图的语义中心度贡献（因此模型的假设前提是：在图中起重要作用的节点也与种子概念有重要的语义关系）。最终从这些语义中心度贡献高的节点中选择标签。本文应用了几种语义中心性措施：Focused Closeness Centrality，Focused Information Centrality，Focused Betweenness Centrality，来计算候选标签的语义中心度。

**《On Conceptual Labeling of a Bag of Words》**

本文利用了probase来进行concept tagging，与上述的研究不同，他没有用到任何主题模型，主要的方法是用信息论的方法来权衡对词袋的语义覆盖度，输出覆盖最广但标签最少的单词。

$$L^*(x|C) = \min(L(x), -\log P_{NML}(x|C))$$

$$\hat{P}(x|C)$$

$$= \min(L(x), -\log \frac{\hat{P}(x|C)}{\sum_{x'} \hat{P}(x'|C)})$$

$$\hat{P}(x|C) = \max_{c \in C} P(x|c)$$

使用MDL（最小长度描述原则）作为选择最好concept的标准，通过最大限度减少描述长度来实现当前概念集C的更新。迭代终止时，描述长度不能再减少了。由于编码长度在每次迭代时单调递减，因此保证了该算法收敛。虽然没有用到任何主题模型，但算法仍然可以通过三种操作（增删改）和MDL收敛的方式，自动的决定主题个数。前提是需要有海量的上下级概念对频数。

## 实际工作中的尝试

在实际工作的尝试中，我们的整体算法流程分为两部分：核心主题关键词抽取以及获取更为抽象的concept短语。

符合一定主题的关键词抽取：在原有的linkKGLDA模型基础上，除了采用知识表示，还对图谱之外的词赋予deepwalk向量，是模型更好的应对图谱之外的词。

更为抽象的concept短语：首先构建基于帖子的知识图谱，除了利用业务词典，丁香园论坛结构，搭建图谱上层，中下层图谱通过层次主题模型，concept，关键词抽取进行搭建。

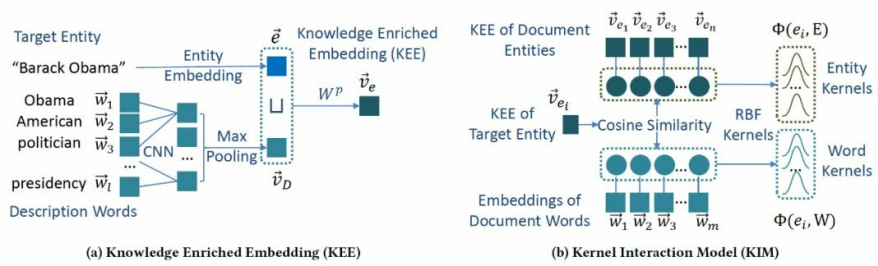
通过的concept抽取模型（可参考：医疗健康领域的短文本解析探索），我们从query和帖子标题中抽取到了300万的concept词语，那么如何才能找到一个帖子真正说的主题，并将文本中的最关键主题词连接到相应的concept上呢，这里我们要借助以下论文中提到的方法：

### 《Towards Better Text Understanding and Retrieval through Kernel Entity Salience Modeling》

由于一篇文章涉及到的实体多种多样，但一般会存在几个最关键的实体，例如：

医生你好，昨天像你提问的宝宝腹泻问题，昨天一共拉了6次，后期都是绿便，晚9点吃最后一顿奶，到今早上5点吃一次，8点多吃一次，然后又拉了2次，但已经不绿了，是黄色稀便，乳糖酶还没有给她吃，昨天加了益生菌，儿歌的两款乳糖酶，我买的葡萄糖这款，你看下对吗？现在是不是已经好转了？我还是应该买纯乳糖酶乳粉？

文中出现了腹泻，绿便，奶等一系列表述疾病过程和食物相关的实体（可通过linkKGLDA识别出来），但中心实体"腹泻"和"乳糖酶乳粉"在识别结果中的排序可能并不是top1，这会使得后续的concept对应工作产生一定的噪音。




本文主要研究实体对文章的显著程度，通过结合文章上下文和实体知识表示（KEE），和Kernel Interaction Model (KIM) 模型，对实体-文章对进行排序，从而得到实体在文章中的显著程度。

实验数据利用的是远监督标注，利用文章和文章标题，以及已有算法（ner，名词短语抽取，实体链接，linkKGLDA概率等）得到训练数据。

### 《A User-Centered Concept Mining System for Query and Document

**《[illegible] Understanding at Tencent》**

本篇文章在tagging docment上的方法有两种，实现的前提条件是需要有足够数量和高质量的query log，以及知识图谱。

整个过程分为两种方法：基于概率和基于匹配：

**基于条件概率：**

文章3.1节描述了如何利用抽取到的主题关键词去对应到一组合适的concept，主要利用的还是主题关键词上下文与concept的条件概率推理。与《Grounding Topic Models with Knowledge Bases》不同，这种方法可以得到图谱中不存在，但是搜索中经常出现的concept，并且也不需要大量上下位词的共现频数。

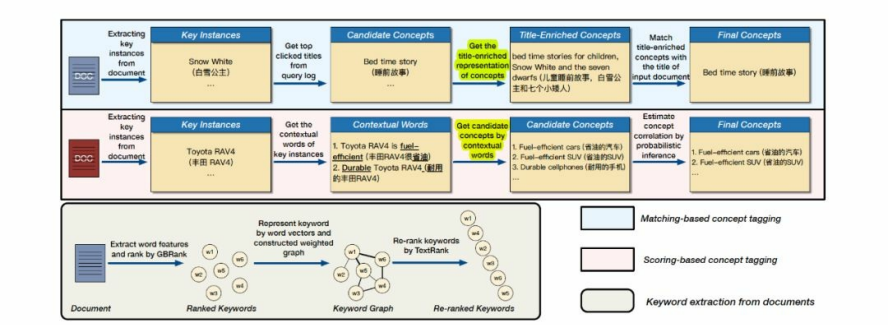


Figure 3: The overall procedures of concept tagging for documents. We combine both a matching-based approach with a scoring-based approach to handle different situations.

$$p(\mathbf{c}|d) = \sum_{i=1}^{|E^d|} p(\mathbf{c}|\mathbf{e}_i^d)p(\mathbf{e}_i^d|d),$$

$$p(\mathbf{c}|\mathbf{e}_i^d) = \sum_{j=1}^{|X_{E^d}|} p(\mathbf{c}|\mathbf{x}_j)p(\mathbf{x}_j|\mathbf{e}_i^d)$$

$$p(\mathbf{c}|\mathbf{x}_j) = \begin{cases} \frac{1}{|\mathbf{C}^{\mathbf{x}_j}|}, & \text{if } \mathbf{x}_j \text{ is a substring of } \mathbf{c}, \\ 0 & \text{otherwise.} \end{cases}$$

整个p（c|d）的计算过程重点在于第三步，即想要计算p(c|x)，必须存在以上下文xj作为子串的概念词(这样的概念词来自于query log)，如： 文中提到的概念词"省油的汽车"和文档中"丰田RAV4"的上下文"省油，耐用"。虽然这样的概念词在医疗专业词汇和搜索中很少见，但这实际上是实体属性和概念属性的一种交集，在没有足够数量和高质量的query log的情况下，我们可以采用属性抽取相关工作的研究结果。这样做更有利于做医疗领域的相关问题，当然也可参照之前的历史文章(taxnomy构建)和上面介绍的《On Conceptual Labeling of a Bag of Words》计算概率值。

**基于匹配：**

1.首先利用GBRank，textRank，word2vec算法，得到一定数量的关键词（instance）

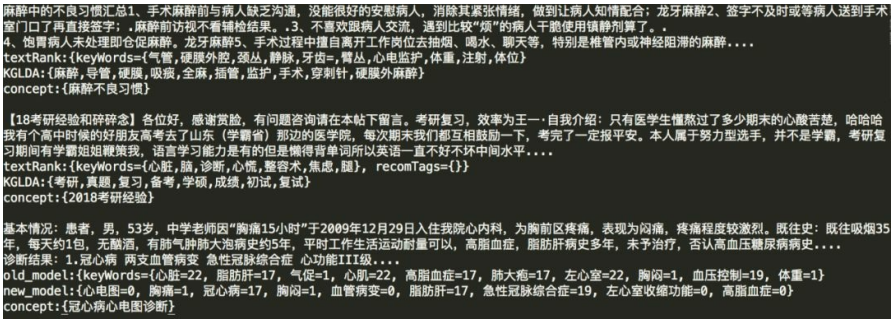
2.利用检索或者web table找到与instance相关的概念候选。每一个concept，用topN搜索结果标题文本信息来丰富concept表示（tf-idf向量）。

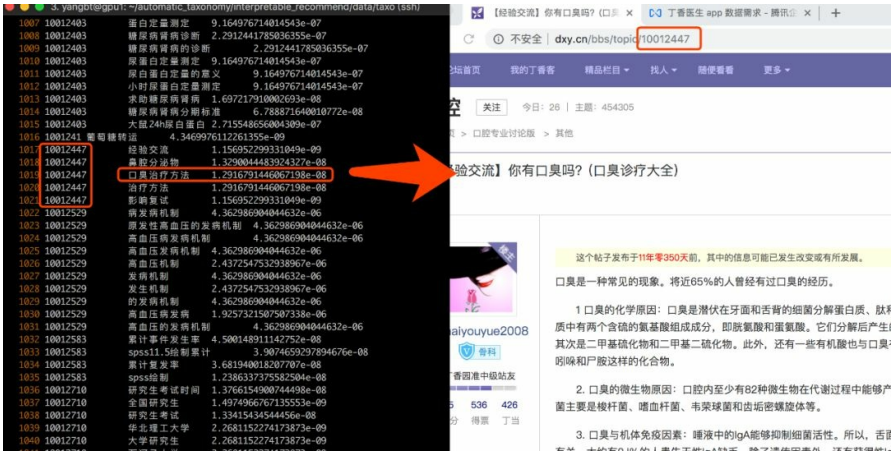
4.将concept表示与文档title tf-idf向量做相似度计算，超过一定阈值，打相应标签。

## 实验对比

## 实验对比

在丁香园论坛帖子的文本数据上，我们应用了前文调研的主题抽取、实体显著性判断、短语挖掘、concept召回等方法，所获得的标签在主题贴近度和可解释性上都有显著提高：




## 总结

标签生成任务虽然在NLP领域非常常见，但是想要获得高质量的标签词，在推荐、搜索、问答等工业场景下应用，背后其实集成了众多NLP基础工作。标签生成的上游包含了分词、命名实体识别、医学概念归一化、消岐、concept质量优化等工作。只有稳固的基础才能把楼盖得更高。

其次，知识图谱就是模型。合理的图谱结构、丰富的数据量将决定最终结果的好坏。在产业界尤其需要关注实际业务下的知识体系构建，一套知识图谱并不一定能满足所有的业务线，比如在医学科普文章下表现良好的图谱，若应用在考研、招聘类的文本下，反而会因为抽出过多的医学专业词汇而偏离主题。可以与业务部门协同补全知识图谱，或者用一些统计学方法加以补充。

最后，标签词并不需要一定作为文本的一种"**显式**"的展示方式，作为长文本的一种更优的结构化数据，它有各种各样的"**隐式**"用法，比如作为特征输入到下游的文本分类、标题生成甚至融入到推荐系统策略中，我们会在今后陆续分享各种有趣的玩儿法。

可能喜欢

- 跨平台NLP/ML文章索引
- NLP数据集构建
- 大话关系抽取
- 如何斩下NLP算法岗offer?
- python开发套件推荐

求关注 求投喂 拉你进高端群哦~

## 参考文献


- Incorporating Knowledge Graph Embeddings into Topic Modeling
- Graph-Sparse LDA: A Topic Model with Structured Sparsity
- Grounding Topic Models with Knowledge Bases
- Modeling Documents by Combining Semantic Concepts with Unsupervised Statistical Learning
- Unsupervised Graph-based Topic Labelling using DBpedia
- On Conceptual Labeling of a Bag of Words
- A User-Centered Concept Mining System for Query and Document Understanding at Tencent
- Towards Better Text Understanding and Retrieval through Kernel Entity Salience Modeling
- End-to-End Neural Ad-hoc Ranking with Kernel Pooling
- Automatic Event Salience Identification