Topic model 主题模型 专知荟萃

基础入门

- 中英文博客以及基础文章

1. Topic model 介绍 简介：简要了解主题模型是什么意思，最基本的概念https://en.wikipedia.org/wiki/Topic_model
2. 概率主题模型简介 Introduction to Probabilistic Topic Models 简介：一步让你知道什么是lda，翻译了提出主题模型大神对概率主题模型的描述。中文文档更适合入门。David M. Blei所写的《Introduction to Probabilistic Topic Models》的译文 http://www.cnblogs.com/siegfang/archive/2013/01/30/2882391.html
3. 主题模型-LDA浅析：简述了LDA的基础概念，描述了模型的生成过程，帮助你进一步了解主题模型~！ http://blog.csdn.net/huagong_adu/article/details/7937616
4. Latent dirichlet allocation：开山之作LDA原论文。了解了主题模型的基础知识之后可以开始看原论文了。原文看不太懂也不要着急，可以先看个大概~ 作者：David M. Blei, Andrew Y. Ng, and Michael I. Jordan 顺便介绍一下Blei大神：David M. Blei Professor in the Statistics and Computer Science departments at Columbia University. Prior to fall 2014 he was an Associate Professor in the Department of Computer Science at Princeton University. His work is primarily in machine learninghttp://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
5. Rickjin 腾讯的rickjin大神：LDA数学八卦 简介：传说中的“上帝掷骰子”的来源之处。这篇文章是一个连载的科普性博客，作者是rickjin，文章分为7个章节，主要5个章节讲得是Gamma函数、Beta/Dirichlet函数、MCMC和Gibbs采样、文本建模、LDA文本建模，对于想要了解和LDA的同学来说，是一篇很好的入门教程，建议结合Blei的开山之作一起看。http://download.csdn.net/download/happyer88/8791493
6. LDA漫游指南 作者：马晨 清华大学在读博士，毕业于北京邮电大学硕士，曾任新浪网公司自然语言处理算法工程师。 简介：完全明白主题模型的必备之路：一份从数学基础开始讲的教程，与LDA数学八卦可以互为补充。 https://yuedu.baidu.com/ebook/d0b441a8ccbff121dd36839a.html###
7. 《Supervised topic models》：有监督主题模型，提出SLDA，实现有监督学习。 作者：David M. Blei https://research.googleblog.com/2016/09/show-and-tell-image-captioning-open.html
8. 《Fast collapsed gibbs sampling for latent dirichlet allocation》：快速推理算法，在参数推理上提出更容易理解的方法。更加实用。事实上，由于方法相对更好理解，现在更多的主题模型都采用这种方法求解。 作者：I Porteous，D Newman，A Ihler A Asuncion P Smythhttps://www.researchgate.net/publication/221653277_Fast_collapsed_Gibbs_sampling_for_latent_Dirichlet_allocation
9. **LDA-math-MCMC 和 Gibbs Sampling** 简介：rickjin大神对马尔科夫链蒙特卡洛采样和吉布斯采样的描述，讲的非常清晰明白。可以帮助大家更好的理解采样。https://cosx.org/2013/01/lda-math-mcmc-and-gibbs-sampling/
10. *用变分推理求解LDA模型的参数* 简介：LDA原文中采用的变分算法求解，想要了解变分算法可以看一下这篇文章。 作者：斯玮 Fantastic https://zhuanlan.zhihu.com/p/28794694
11. **早期文本模型的简介** https://zhuanlan.zhihu.com/p/28777266
12. **Gensim简介、LDA编程实现、LDA主题提取效果图展 示**https://zhuanlan.zhihu.com/p/28830480
13. **图模型学习** http://blog.csdn.net/pipisorry/article/category/6241251
14. Gaussian LDA: Gaussian LDA简介, 介绍主题模型和词向量结合的一些工作， 比较有意思，建议看一下 [http://blog.csdn.net/u011414416/article/details/51188483]

进阶论文

- 实践以及一些变形方法

1. **如何计算两个文档的相似度（一）** 简介：52nlp上的文章，从最简单的tf-idf到SVD和LSI再到) LDA，可以说是形成了一条逻辑线，一步一步说明算法的发展过程，同时也方便对比各种算法的优缺点。另外，从实践的角度出发。迅速上手！用到了python里的gensim，这是一个非常好用的库，实践必不可少。 http://www.52nlp.cn/%E5%A6%82%E4%BD%95%E8%AE%A1%E7%AE%97%E4%B8%A4%E4%B8%AA%E6%96%87%E6%A1%A3%E7%9A%84
2. **如何计算两个文档的相似度（二）** 从gensim最基本的安装讲起，然后举一个非常简单的例子用以说明如何使用gensim，可以跟着教程做一下实验，肯定会有更好地体会 http://www.52nlp.cn/%E5%A6%82%E4%BD%95%E8%AE%A1%E7%AE%97%E4%B8%A4%E4%B8%AA%E6%96%87%E6%A1%A3%E7%9A%84
3. 文章说了很多实验的细节，讲了如何数据预处理，解决了很多理论类文章中不会提到的技术细节。NTLK是著名的Python自然语言处理工具包，在这也讲了怎么去用这些工具。 http://www.52nlp.cn/%E5%A6%82%E4%BD%95%E8%AE%A1%E7%AE%97%E4%B8%A4%E4%B8%AA%E6%96%87%E6%A1%A3%E7%9A%84
4. **A correlated topic model of science** Blei的大作，引入了主题之间的关联。考虑到了潜在主题的子集将是高度相关的。 http://www.cs.columbia.edu/~blei/papers/BleiLafferty2007.pdf（ppt） http://www-users.cs.umn.edu/~banerjee/Teaching/Fall07/talks/Muhammed_slides.pdf
5. **Topic Models over Text Streams: A Study of Batch and Online Unsupervised Learning.** 文本流推理 作者：A Banerjee ， S Basu http://www-users.cs.umn.edu/~banerjee/papers/07/sdm-topics-long.pdf
6. **Topical n-grams: Phrase and topic discovery, with an application to information retrieval** 在LDA基础上考虑了词与词之间的顺序 作者：X Wang ， A Mccallum ， X Weihttp://www.cs.cmu.edu/~xuerui/papers/ngram_tr.pdf

7. **Hierarchical Dirichlet processes.** 基于DirichletProcess的变形，即HDP模型，可以自动的学习出主题的数目。该方法：1、在一定程度之上解决了主题模型中自动确定主题数目这个问题，2、代价是必须小心的设定、调整参数的设置，3、实际中运行复杂度更高，代码复杂难以维护。 所以在实际中，往往取一个折中，看看自动确定主题数目这个问题对于整个应用的需求到底有多严格，如果经验设定就可以满足的话，就不用采用基于非参数贝叶斯的方法了，但是如果为了引入一些先验只是或者结构化信息，往往非参数是优先选择，例如树状层次的主题模型和有向无环图的主题模型 作者：Yee Whye Michael I. Jordan J Beal David M. Bleihttps://people.eecs.berkeley.edu/~jordan/papers/hdp.pdf
8. **Modeling online reviews with multi-grain topic models* 从用户评论数据中进行无监督主题抽取，考虑了一个多级背景主题模型：词~句子~段落~文档，解决了传统LDA模型提出的主题往往对应品牌而不是可以ratable的主题。 作者：I Titov ， R Mcdonaldhttp://delivery.acm.org/10.1145/1370000/1367513/p111-titov.pdf
9. **A joint model of text and aspect ratings for sentiment summarization.** 本文将一些具有结构化信息的特征融入到主题模型中，具体来说，我们同时关联两个生成过程，一个就是文档中词的生成，另一个就是这些结构化特征的生成。 作者：Titov ， Ivan ， McDonald ， Ryanhttp://www.aclweb.org/anthology/P08-1036
10. **Comparing twitter and traditional media using topic models.** 用于社交媒体研究的方法，提出Twtter-LDA，传统LDA并不适用于短文本，这篇论文解决了这一缺点。 作者：WX Zhao J Jiang，J Weng， J H EP Lim https://link.springer.com/chapter/10.1007%2F978-3-642-20161-5_34

更多Papers推荐

1. Multi-modal Multi-view Topic-opinion Mining for Social Event Analysis. 将主题模型用于多媒体分析，同时考虑了opinion，view，collection等因素 作者：Shengsheng Qian Tianzhu Zhang Changsheng Xu http://delivery.acm.org/10.1145/2970000/2964294/p2-qian.pdf
2. TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency RNN与主题模型结合，结合了主题模型的全局信息和RNN的局部特征。 作者：AB Dieng， C Wang， J Gao， J Paisley https://arxiv.org/pdf/1611.01702.pdf
3. Cross-media Topic Detection with Refined CNN based Image-Dominant Topic Model CNN与主题模型结合 作者：Z Wang， L Li ， Q Huanghttp://delivery.acm.org/10.1145/2810000/2806309/p1171-wang.pdf
4. Gaussian LDA for Topic Models with Word Embeddings word embedding 应用于LDA变形 作者：R Das， M Zaheer， C Dyer http://rajarshd.github.io/papers/acl2015.pdf

一些主题模型的应用场景

Papers for NLP

1. Topic modeling: beyond bag-of-words 为文本语料库建模提供了一种替代方法。 作者：Hanna M. Wallach http://delivery.acm.org/10.1145/1150000/1143967/p977-wallach.pdfhttps://people.cs.umass.edu/~wallach/talks/beyond_bag-of-words.pdf （ppt）
2. Topical n-grams: Phrase and topic discovery, with an application to information retrieval 本文介绍了主题n-gram即一个发现主题以及主题短语的主题模型。 作者：Andrew McCallum, Xing Wei University of Massachusettshttp://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4470313
3. A topic model for word sense disambiguation 用WORDNET（LDAWN）开发LDA 作者：JL Boyd-Graber ， DM Blei ， X Zhu http://www.aclweb.org/anthology/D07-1109

Papers for opinion mining

1. Topic sentiment mixture: modeling facets and opinions in weblogs 定义了Weblogs主题情感分析的问题，并提出了一种概率模型来同时捕捉主题和情绪的混合。 作者：Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, Chengxiang Zhaihttp://delivery.acm.org/10.1145/1250000/1242596/p171-mei.pdf
2. A joint model of text and aspect ratings for sentiment summarization 提出一个统计模型，能够在文本中发现相应的主题，并从支持每个方面评级的评论中提取文本证据。 作者：Titov， Ivan， McDonald， Ryan http://www.aclweb.org/anthology/P08-1036
3. Current State of Text Sentiment Analysis from Opinion to Emotion Mining 较新的文章，很全面的介绍了opinion挖掘的当前状况。 作者：OR Zaiane http://delivery.acm.org/10.1145/3060000/3057270/a25-yadollahi.pdf

Papers for retrieval

1. LDA-based document models for ad-hoc retrieval 在语言建模框架内提出基于LDA的文档模型，并对几个TREC集合进行评估。 作者：X Wei ， WB Crofthttp://delivery.acm.org/10.1145/1150000/1148204/p178-wei.pdf
2. Probabilistic Models for Expert Finding 设计算法找到某个领域的专家。 作者：Hui Fang ChengXiang Zhai https://link.springer.com/chapter/10.1007%2F978-3-540-71496-5_38
3. Thread-based probabilistic models for expert finding in enterprise Microblogs. 提出一个概率文件候选模型，该模型可以在企业微博中找到更多专家。 作者：Zhe Xu Jay Ramanathan Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, United States https://ac.els-cdn.com/S0957417415004406/1-s2.0-S0957417415004406-main.pdf

Papers for information extraction

1. Employing Topic Models for Pattern-based Semantic Class Discovery 从语义类的角度出发，做信息提取。具体可以参考ppt 作者：Huibin Zhang Nankai University Mingjie Zhu University of Science and Technology of China huming Shi Ji-Rong Wen Microsoft Research Asia http://www.aclweb.org/anthology/P09-1052https://pdfs.semanticscholar.org/604b/c2fb02b48d6d106215955a6a30629314df14.pdf (ppt)
2. Combining Concept Hierarchies and Statistical Topic Models 提供一个通用的数据驱动框架，用于从大量文本文档中自动发现高级知识。 作者：C Chemudugunta ， P Smyth ， M Steyvers http://delivery.acm.org/10.1145/1460000/1458337/p1469-chemudugunta.pdf
3. An Unsupervised Framework for Extracting and Normalizing Product Attributes from Multiple Web Sites 开发了一个无监督的框架，用于从源自不同站点的多个网页同时提取和归一化产品的属性。 作者：Tak-Lam Wong Wai Lam The Tik-Shun Wong The Chinese University of Hong Kong, Hong Kong, Hong Konghttp://delivery.acm.org/10.1145/1400000/1390343/p35-wong.pdf

Tutorials

1. Courses 哥伦比亚大学给出的教程，David M. Blei的课程http://www.cs.columbia.edu/~blei/courses.html
2. LDA数学八卦 传说中的“上帝掷骰子”的来源之处。这篇文章是一个连载的科普性博客，作者是rickjin，文章分为7个章节，主要5个章节讲得是Gamma函数、Beta/Dirichlet函数、MCMC和Gibbs采样、文本建模、LDA文本建模，对于想要了解和LDA的同学来说，是一篇很好的入门教程，建议结合Blei的开山之作一起看。 作者：Rickjin 腾讯的rickjin大神http://download.csdn.net/download/happyer88/8791493
3. LDA漫游指南 完全明白主题模型的必备之路：一份从数学基础开始讲的教程，与LDA数学八卦可以互为补充。 作者：马晨 清华大学在读博士，毕业于北京邮电大学硕士，曾任新浪网公司自然语言处理算法工程师。https://yuedu.baidu.com/ebook/d0b441a8ccbff121dd36839a.html###
4. MIT自然语言处理第三讲：概率语言模型 很系统的英文教程，这里给出了第一部分，后边几部分在52nlp也有翻译，可以对照去看看。 作者：Regina Barzilay (MIT,EECS Department, November 15, 2004) /52nlp上的翻译版本 http://people.csail.mit.edu/regina/6881/http://www.52nlp.cn/mit-nlp-third-lesson-probabilistic-language-modeling-first-part
5. 斯坦福大学深度学习与自然语言处理第二讲：词向量 这里给出了整个深度学习与自然语言处理的连接。很适合想要做主题模型与深度学习相结合的人看。 作者：Richard Socher 斯坦福大学青年才俊 http://cs224d.stanford.edu/
6. topic_modeling_tutorial 除了基本的概念还包括在python上实现的流程。指导编程实现。 作者：piskvorky https://github.com/piskvorky/topic_modeling_tutorial

综述

1. Probabilistic Topic Models: Origins and Challenges 权威综述，介绍了很多基本的主题模型，还包括这些模型之间渐进的关系 作者：David M. Bleihttp://www.cs.columbia.edu/~blei/talks/Blei_Topic_Modeling_Workshop_2013.pdf
2. Probabilistic Topic Models 作者：David M. Bleihttp://www.cs.columbia.edu/~blei/talks/Blei_MLSS_2012.pdf
3. 通俗理解LDA主题模型 相对简单一些的中文综述，可以帮助读者迅速理解各种基本概念。 作者：v_JULY_v http://blog.csdn.net/v_july_v/article/details/41209515

视频教程

1. Probabilistic topic models http://delivery.acm.org/10.1145/2110000/2107741/tutorial-6-part1.mp4
2. Probabilistic topic models http://delivery.acm.org/10.1145/2110000/2107741/tutorial-6-part2.mp4
3. a 2008 talk on dynamic and correlated topic models applied to the journal Science .http://www.cs.columbia.edu/~blei/talks/Blei_Science_2008.pdf

代码

1. Topic modeling software https://github.com/Blei-Lab
2. blei的github主页，有大量代码 lda-c (Latent Dirichlet allocation) LDA代码 http://www.cs.columbia.edu/~blei/lda-c/index.html
3. Supervised topic models for classification 有监督LDAhttp://www.cs.cmu.edu/~chongw/slda/
4. R package for Gibbs sampling in many models 吉布斯采样代码 https://cran.r-project.org/web/packages/lda/
5. online lda 在线lda http://www.cs.princeton.edu/~blei/downloads/onlineldavb.tar

6. Online inference for the HDP Hierarchical Dirichlet processes.http://www.cs.cmu.edu/~chongw/software/onlinehdp.tar.gz

7. Collaborative modeling for recommendation 关联主题模型http://www.cs.cmu.edu/~chongw/citeulike/

8. Dynamic topic models and the influence model 动态主题模型https://code.google.com/archive/p/princeton-statistical-learning/downloads

领域专家

1. David M. Blei
   - LDA开上鼻祖，哥伦比亚大学统计与计算机科学系教授。曾在普林斯顿大学计算机科学系担任副教授。 他的工作主要是机器学习。他的博客中包含很多主题模型的知识，也可以很快地了解主题模型的发展方向。 http://www.cs.columbia.edu/~blei/

2. Ivan Titov Иван Титов 图模型方面的专家，有许多高水平论文。博客中有很多好的资源可以使读者了解主题模型的发展。 http://www.ivan-titov.org/

3. Eric xing
   - My principal research interests lie in the development of machine learning and statistical methodology, andlarge-scale computational system and architecture, for solving problems involving automated learning, reasoning, and decision-making in high-dimensional, multimodal, and dynamic possible worlds in artificial, biological, and social systems. http://www.cs.cmu.edu/~epxing/

4. 朱军
   - My research focuses on developing statistical machine learning methods to understand complex scientific and engineering data. My current interests are in latent variable models, large-margin learning, Bayesian nonparametrics, and deep learning. Before joining Tsinghua in 2011, I was a post-doc researcher and project scientist at the Machine Learning Department in Carnegie Mellon University.http://ml.cs.tsinghua.edu.cn/~jun/index.shtml

5. Alexander J. Smola
   - Professor, Carnegie Mellon University, CEO, Marianas Labs, 亚马逊云服务(AWS)机器学习总监
   - http://alex.smola.org/
   - Alex Smola，1996 年毕业于慕尼黑工业大学，获物理学硕士学位。1998年在柏林工业大学取得计算机科学博士学位。之后，他在澳大利亚国立大学担任研究院和研究小组组长。2004 - 2008年，Alex Smola 在NICTA研究中心统计机器学习项目担任项目负责人。2008 年，他加入雅虎，后于 2012 年加入谷歌从事研究工作。他是加州大学伯克利分校的兼职教授，目前担任卡内基梅隆大学机器学习教授。2015 年，他与人联合创立了 Marianas 实验室。2016 年，Alex 加入亚马逊，目前担任亚马逊 AWS 机器学习总监。迄今为止共发表超过 200 篇论文并参与编写 5 本学术专著。他的研究兴趣包括：算法的可扩展性，SVM、高斯过程和条件随机场等核方法，统计建模以及用户建模、文档分析、时序模型等各种机器学习应用。
   - 他最近的一篇工作比较有意思， **Latent LSTM Allocation: Joint Clustering and Non-Linear Dynamic Modeling of Sequence Data** 把LDA 和 LSTM 结合一起，赞。

初步版本，水平有限，有错误或者不完善的地方，欢迎大家提建议和补充，会一直保持更新，敬请关注http://www.zhuanzhi.ai 和关注**专知公众号**，获取第一手AI相关知识