深度学习(CNN RNN Attention)大规模文本分类问题综述和实践（附周一新闻四则，合 23k 字 17 页）

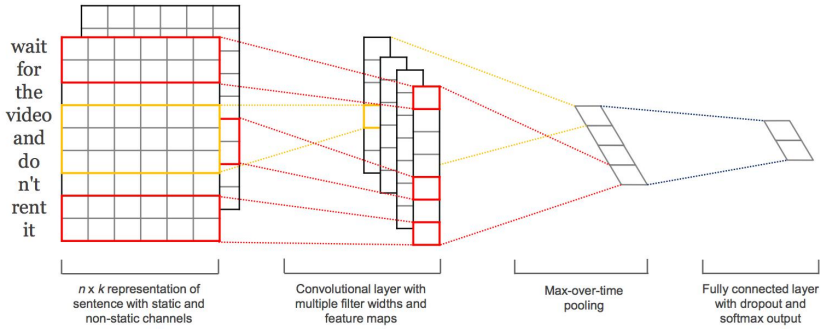
秦陇纪 10 汇编

目录

【作者介绍】清淞（花名），现居北京，阿里巴巴集团搜索排序部门算法相关工作。本科毕业于山东大学，研究生北邮机器学习方向。毕业后在阿里巴巴从事搜索排序算法相关工作。希望可以和业界学界多交流。

【摘要】深度学习(CNN RNN Attention)大规模文本分类问题综述和实践；Python 在 CV、NLP、ML 和 DM 等六大方面的资源汇总；附今日新闻四则 5k 字；合 23k 字 17 页。本文主要根据清淞昨天首发于《用深度学习（CNN RNN Attention）解决大规模文本分类问题 - 综述和实践》知乎深海遨游（原文网址 https://zhuanlan.zhihu.com/p/25928551）汇编而成。文末打赏后“阅读原文”可下载完整+周一新闻四则，合 23k 字 17 页 PDF 文档。

I. 深度学习(CNN RNN Attention)大规模文本分类问题综述和实践



近来在同时做一个应用深度学习解决淘宝商品的类目预测问题的项目，恰好硕士毕业时论文题目便是文本分类问题，趁此机会总结下文本分类领域特别是应用深度学习解决文本分类的相关的思路、做法和部分实践的经验。[1]

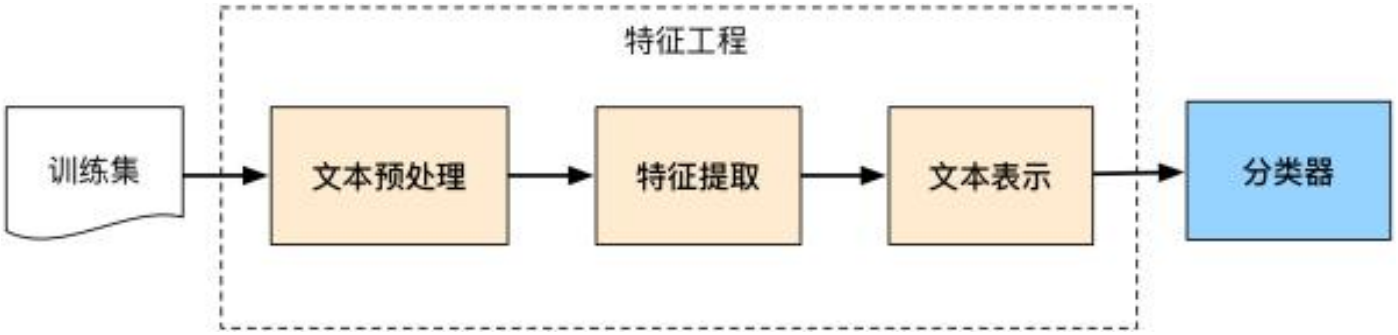
业务问题描述：

淘宝商品的一个典型的例子见下图，图中商品的标题是“夏装雪纺条纹短袖 t 恤女春半袖衣服夏天中长款大码胖 mm 显瘦上衣夏”。淘宝网后台是通过树形的多层的类目体系管理商品的，覆盖叶子类目数量达上万个，商品量也是 10 亿量级，我们是任务是根据商品标题预测其所在叶子类目，示例中商品归属的类目为“女装/女士精品>>蕾丝衫/雪纺衫”。很显然，这是一个非常典型的短文本多分类问题。接下来分别会介绍下文本分类传统和深度学习的做法，最后简单梳理下实践的经验。

一、传统文本分类方法

文本分类问题算是自然语言处理领域中一个非常经典的问题了，相关研究最早可以追溯到上世纪 50 年代，当时是通过专家规则（Pattern）进行分类，甚至在 80 年代初一度发展到利用知识工程建立专家系统，这样做的好处是短平快的解决 top 问题，但显然天花板非常低，不仅费时费力，覆盖的范围和准确率都非常有限。

后来伴随着统计学习方法的发展，特别是 90 年代后互联网在线文本数量增长和机器学习学科的兴起，逐渐形成了一套解决大规模文本分类问题的经典玩法，这个阶段的主要套路是人工特征工程+浅层分类模型。训练文本分类器过程见下图：



整个文本分类问题就拆分成了特征工程和分类器两部分，玩机器学习的同学对此自然再熟悉不过了

1.1 特征工程

特征工程在机器学习中往往是最耗时耗力的，但却极其的重要。抽象来讲，机器学习问题是把数据转换成信息再提炼到知识的过程，特征是“数据-->信息”的过程，决定了结果的上限，而分类器是“信息-->知识”的过程，则是去逼近这个上限。然而特征工程不同于分类器模型，不具备很强的通用性，往往需要结合对特征任务的理解。

文本分类问题所在的自然语言领域自然也有其特有的特征处理逻辑，传统分本分类任务大部分工作也在此处。文本特征工程分位文本预处理、特征提取、文本表示三个部分，最终目的是把文本转换成计算机可理解的格式，并封装足够用于分类的信息，即很强的特征表达能力。

1）文本预处理

文本预处理过程是在文本中提取关键词表示文本的过程，中文文本处理中主要包括文本分词和去停用词两个阶段。之所以进行分词，是因为很多研究表明特征粒度为词粒度远好于字粒度，其实很好理解，因为大部分分类算法不考虑词序信息，基于字粒度显然损失了过多“n-gram”信息。

具体到中文分词，不同于英文有天然的空格间隔，需要设计复杂的分词算法。传统算法主要有基于字符串匹配的正向/逆向/双向最大匹配；基于理解的句法和语义分析消歧；基于统计的互信息/CRF 方法。近年来随着深度学习的应用，WordEmbedding + Bi-LSTM+CRF 方法逐渐成为主流，本文重点在文本分类，就不展开了。而停止词是文本中一些高频的代词连词介词等对文本分类无意义的词，通常维护一个停用词表，特征提取过程中删除停用表中出现的词，本质上属于特征选择的一部分。

经过文本分词和去停止词之后淘宝商品示例标题变成了下图“ / ”分割的一个个关键词的形式：

夏装 / 雪纺 / 条纹 / 短袖 / t 恤 / 女 / 春 / 半袖 / 衣服 / 夏天 / 中长款 / 大码 / 胖 mm / 显瘦 / 上衣 / 夏

2）文本表示和特征提取

文本表示：

文本表示的目的是把文本预处理后的转换成计算机可理解的方式，是决定文本分类质量最重要的部分。传统做法常用词袋模型（BOW, Bag Of Words）或向量空间模型（Vector Space Model），最大的不足是忽略文本上下文关系，每个词之间彼此独立，并且无法表征语义信息。词袋模型的示例如下：

(0, 0, 0, 0, ... , 1, ... 0, 0, 0, 0)

一般来说词库量至少都是百万级别，因此词袋模型有个两个最大的问题：高纬度、高稀疏性。词袋模型是向量空间模型的基础，因此向量空间模型通过特征项选择降低维度，通过特征权重计算增加稠密性。

特征提取：

向量空间模型的文本表示方法的特征提取对应特征项的选择和特征权重计算两部分。特征选择的基本思路是根据某个评价指标独立的对原始特征项（词项）进行评分排序，从中选择得分最高的一些特征项，过滤掉其余的特征项。常用的评价有文档频率、互信息、信息增益、 x^2 统计量等。

特征权重主要是经典的 TF-IDF 方法及其扩展方法，主要思路是一个词的重要度与在类别内的词频成正比，与所有类别出现的次数成反比。

3）基于语义的文本表示

传统做法在文本表示方面除了向量空间模型，还有基于语义的文本表示方法，比如 LDA 主题模型、LSI/PLSI 概率潜在语义索引等方法，一般认为这些方法得到的文本表示可以认为文档的深层表示，而

word embedding 文本分布式表示方法则是深度学习方法的重要基础，下文会展现。

1.2 分类器

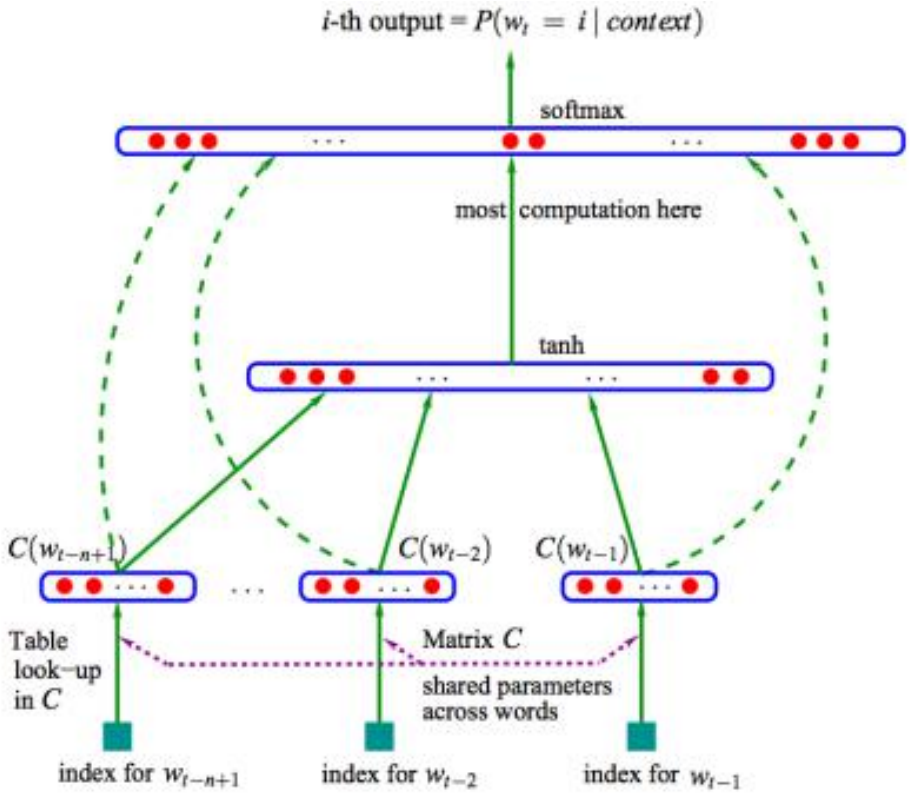
分类器基本都是统计分类方法了，基本上大部分机器学习方法都在文本分类领域有所应用，比如朴素贝叶斯分类算法（Naïve Bayes）、KNN、SVM、最大熵和神经网络等等，传统分类模型不是本文重点，在这里就不展开了。

二、深度学习文本分类方法

上文介绍了传统的文本分类做法，传统做法主要问题的文本表示是高纬度高稀疏的，特征表达能力很弱，而且神经网络很不擅长对此类数据的处理；此外需要人工进行特征工程，成本很高。而深度学习最初在之所以图像和语音取得巨大成功，一个很重要的原因是图像和语音原始数据是连续和稠密的，有局部相关性，。应用深度学习解决大规模文本分类问题最重要的是解决文本表示，再利用 CNN/RNN 等网络结构自动获取特征表达能力，去掉繁杂的人工特征工程，端到端的解决问题。接下来会分别介绍：

2.1 文本的分布式表示：词向量（word embedding）

分布式表示（Distributed Representation）其实 Hinton 最早在 1986 年就提出了，基本思想是将每个词表达成 n 维稠密、连续的实数向量，与之相对的 one-hot encoding 向量空间只有一个维度是 1，其余都是 0。分布式表示最大的优点是具备非常 powerful 的特征表达能力，比如 n 维向量每维 k 个值，可以表征 k^n个概念。事实上，不管是神经网络的隐层，还是多个潜在变量的概率主题模型，都是应用分布式表示。下图是 03 年 Bengio 在 A Neural Probabilistic Language Model 的网络结构：



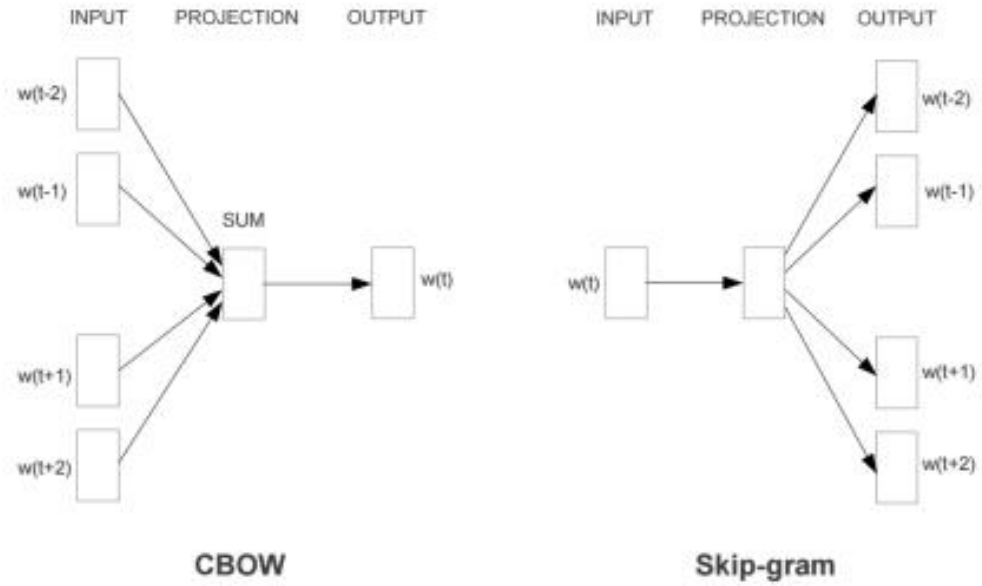
这篇文章提出的神经网络语言模型（NNLM，Neural Probabilistic Language Model）采用的是文本分布式表示，即每个词表示为稠密的实数向量。NNLM 模型的目标是构建语言模型：

$$\text{The objective is to learn a good model } f(w_t,\cdots,w_{t-n+1}) = \hat{P}(w_t|w_1^{t-1}),$$

词的分布式表示即词向量（word embedding）是训练语言模型的一个附加产物，即图中的 Matrix C。

尽管 Hinton 86 年就提出了词的分布式表示，Bengio 03 年便提出了 NNLM，词向量真正火起来是 google Mikolov 13 年发表的两篇 word2vec 的文章 Efficient Estimation of Word Representations in Vector Space 和 Distributed Representations of Words and Phrases and their Compositionality，更重要的是发布了简单好用的 word2vec 工具包，在语义维度上得到了很好的验证，极大的推进了文本分析的进程。下图是文中提出的 CBOW 和 Skip-Gram 两个模型的结构，基本类似于 NNLM，不同的是模型去掉了非线性隐层，预测目标不同，CBOW 是上下文词预测当前词，Skip-Gram

则相反。



CBOW **Skip-gram**

除此之外，提出了 Hierarchical Softmax 和 Negative Sample 两个方法，很好的解决了计算有效性，事实上这两个方法都没有严格的理论证明，有些 trick 之处，非常的实用主义。详细的过程不再阐述了，有兴趣深入理解 word2vec 的，推荐读读这篇很不错的 paper：word2vec Parameter Learning Explained。额外多提一点，实际上 word2vec 学习的向量和真正语义还有差距，更多学到的是具备相似上下文的词，比如“good”“bad”相似度也很高，反而是文本分类任务输入有监督的语义能够学到更好的语义表示，有机会后续系统分享下。

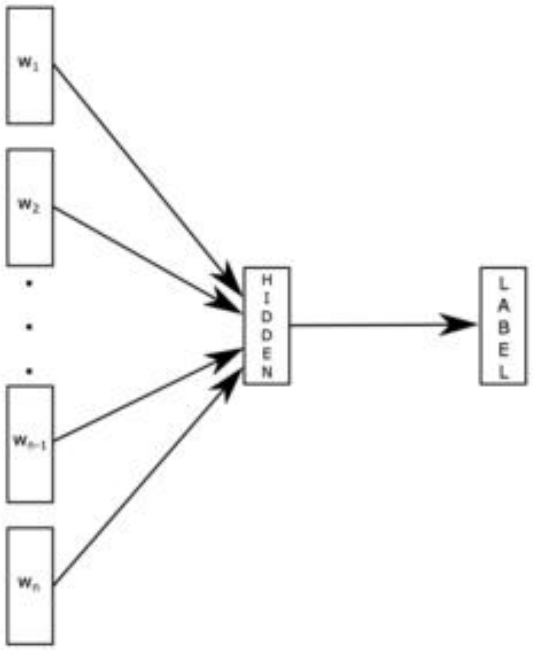
至此，文本的表示通过词向量的表示方式，把文本数据从高纬度高稀疏的神经网络难处理的方式，变成了类似图像、语音的的连续稠密数据。深度学习算法本身有很强的数据迁移性，很多之前在图像领域很适用的深度学习算法比如 CNN 等也可以很好的迁移到文本领域了，下一小节具体阐述下文本分类领域深度学习的方法。

2.2 深度学习文本分类模型

词向量解决了文本表示的问题，该部分介绍的文本分类模型则是利用 CNN/RNN 等深度学习网络及其变体解决自动特征提取（即特征表达）的问题。

1）fastText

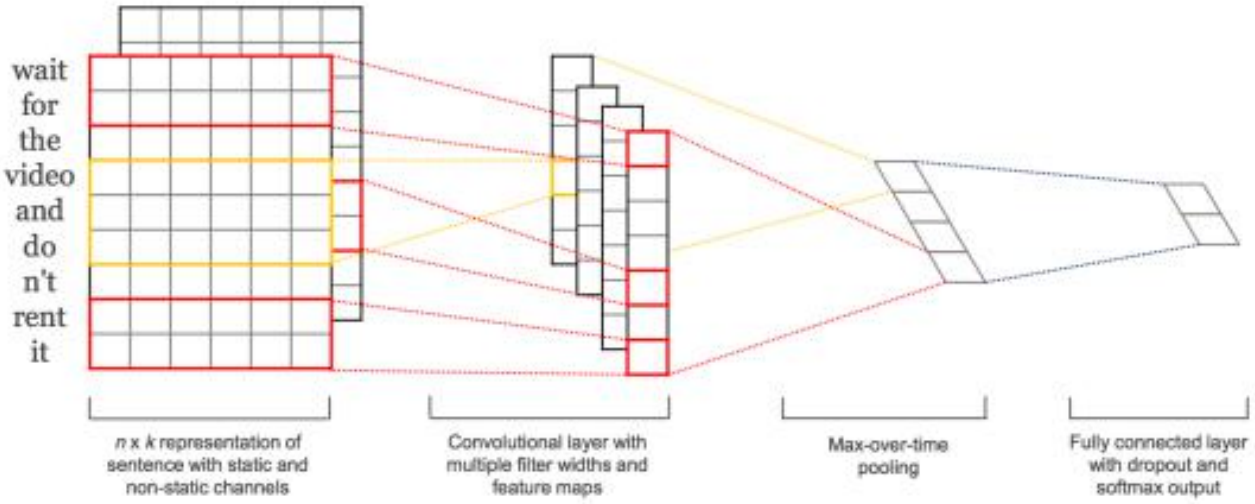
fastText 是上文提到的 word2vec 作者 Mikolov 转战 Facebook 后 16 年 7 月刚发表的一篇论文 Bag of Tricks for Efficient Text Classification。把 fastText 放在此处并非因为它是文本分类的主流做法，而是它极致简单，模型图见下：



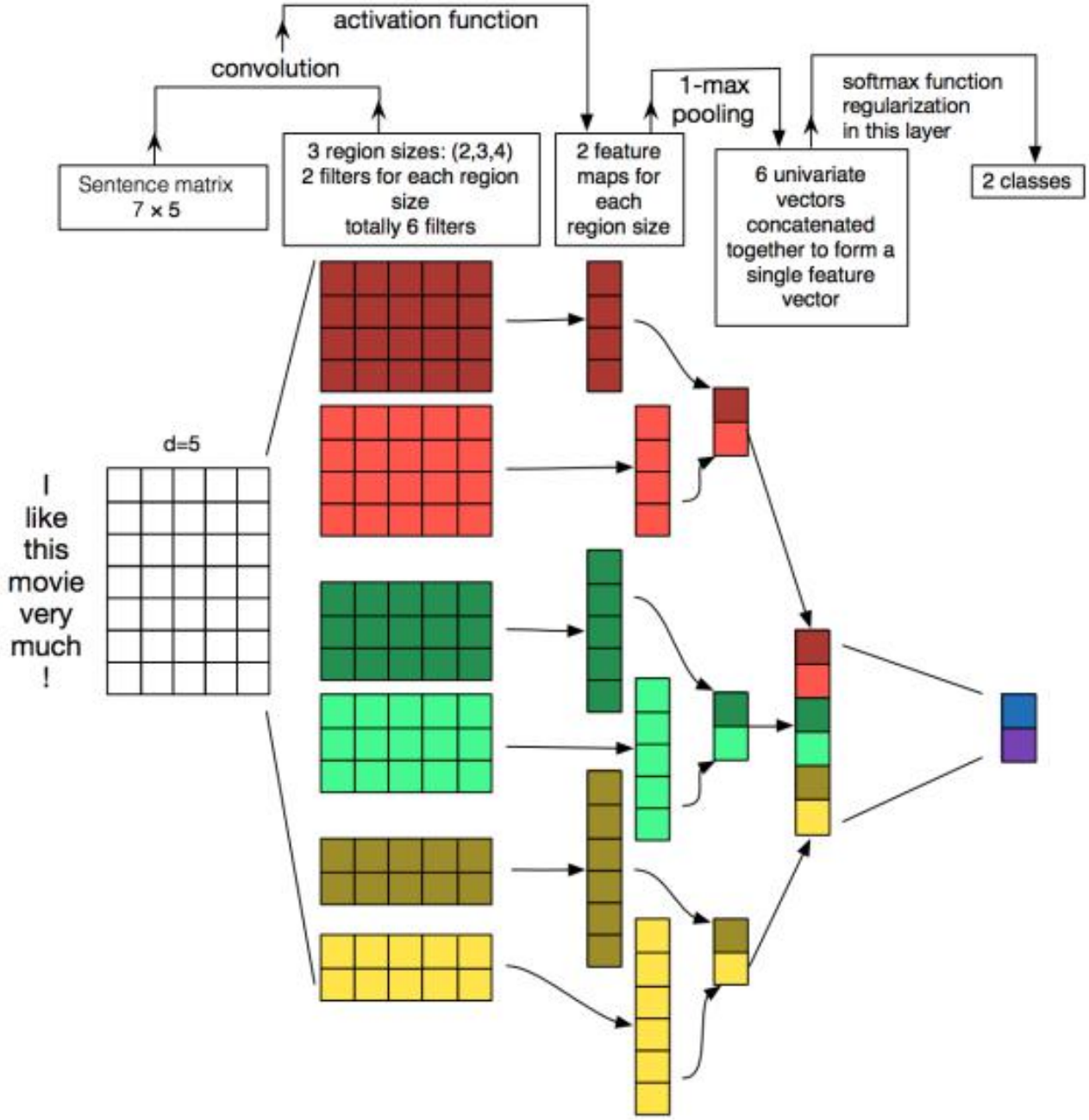
原理是把句子中所有的词向量进行平均(某种意义上可以理解为只有一个 avg pooling 特殊 CNN)，然后直接接 softmax 层。其实文章也加入了一些 n-gram 特征的 trick 来捕获局部序列信息。文章倒没太多信息量，算是“水文”吧，带来的思考是文本分类问题是有一些“线性”问题的部分[from 项亮]，也就是说不必做过多的非线性转换、特征组合即可捕获很多分类信息，因此有些任务即便简单的模型便可以搞定了。

2）TextCNN

本篇文章的题图选用的就是 14 年这篇文章提出的 TextCNN 的结构（见下图）。fastText 中的网络结果是完全没有考虑词序信息的，而它用的 n-gram 特征 trick 恰恰说明了局部序列信息的重要意义。卷积神经网络（CNN Convolutional Neural Network）最初在图像领域取得了巨大成功，CNN 原理就不讲了，核心点在于可以**捕捉局部相关性**，具体到文本分类任务中可以利用 CNN 来提取句子中类似 n-gram 的关键信息。



TextCNN 的详细过程原理图见下：



TextCNN 详细过程：第一层是图中最左边的 7 乘 5 的句子矩阵，每行是词向量，维度=5，这个可以类比为图像中的原始像素点了。然后经过有 filter_size=(2,3,4) 的一维卷积层，每个 filter_size 有两个输出 channel。第三层是一个 1-max pooling 层，这样不同长度句子经过 pooling 层之后都能变成定长的表示了，最后接一层全连接的 softmax 层，输出每个类别的概率。

特征：这里的特征就是词向量，有静态（static）和非静态（non-static）方式。static 方式采用比如 word2vec 预训练的词向量，训练过程不更新词向量，实质上属于迁移学习了，特别是数据量比较小的情况下，采用静态的词向量往往效果不错。non-static 则是在训练过程中更新词向量。推荐的方式是 non-static 中的 fine-tunning 方式，它是以预训练（pre-train）的 word2vec 向量初始化词向量，训练过程中调整词向量，能加速收敛，当然如果有充足的训练数据和资源，直接随机初始化词向量效果也是可以的。

通道（Channels）：图像中可以利用 (R, G, B) 作为不同 channel，而文本的输入的 channel 通常是不同方式的 embedding 方式（比如 word2vec 或 Glove），实践中也有利用静态词向量和 fine-tunning 词向量作为不同 channel 的做法。

一维卷积（conv-1d）：图像是二维数据，经过词向量表达的文本为一维数据，因此在 TextCNN 卷积用的是一维卷积。一维卷积带来的问题是需要设计通过不同 filter_size 的 filter 获取不同宽度的视野。

Pooling 层：利用 CNN 解决文本分类问题的文章还是很多的，比如这篇 A Convolutional Neural

Network for Modelling Sentences 最有意思的输入是在 pooling 改成 (dynamic) k-max pooling ，pooling 阶段保留 k 个最大的信息，保留了全局的序列信息。比如在情感分析场景，举个例子：

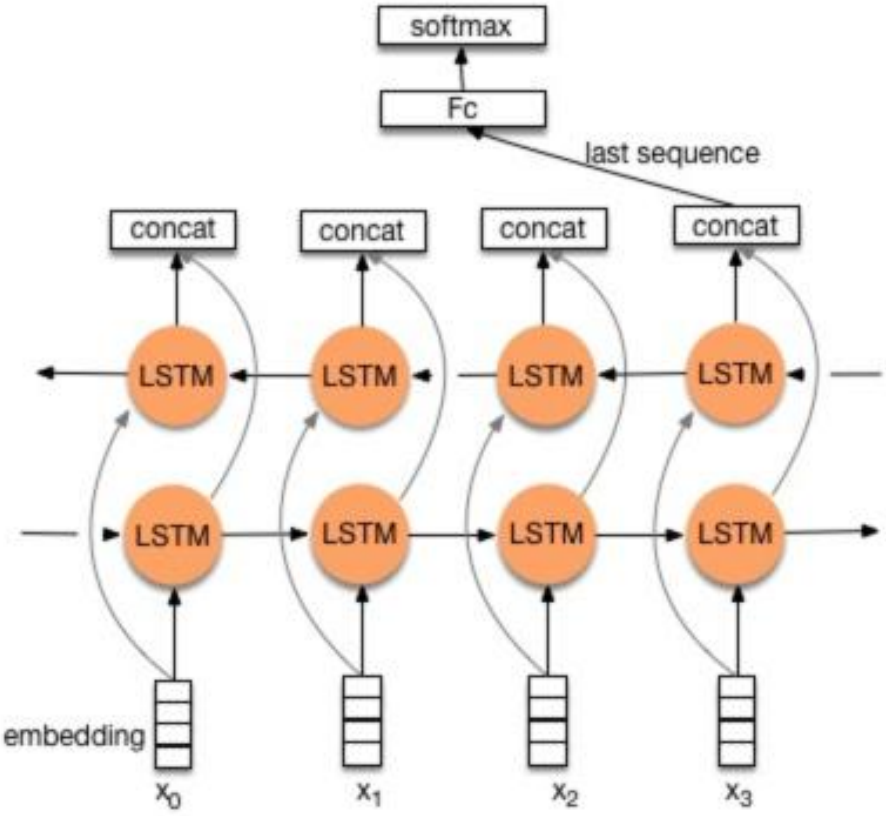
“ 我觉得这个地方景色还不错，但是人也实在太多了 ”

虽然前半部分体现情感是正向的，全局文本表达的是偏负面的情感，利用 k-max pooling 能够很好捕捉这类信息。

3）TextRNN

尽管 TextCNN 能够在很多任务里面能有不错的表现，但 CNN 有个最大问题是固定 filter_size 的视野，一方面无法建模更长的序列信息，另一方面 filter_size 的超参调节也很繁琐。CNN 本质是做文本的特征表达工作，而自然语言处理中更常用的是递归神经网络（RNN, Recurrent Neural Network），能够更好的表达上下文信息。具体在文本分类任务中，Bi-directional RNN（实际使用的是双向 LSTM）从某种意义上可以理解为可以捕获变长且双向的的 "n-gram" 信息。

双向 LSTM 算是在自然语言处理领域非常一个标配网络了，在序列标注/命名体识别/seq2seq 模型等很多场景都有应用，下图是 Bi-LSTM 用于分类问题的网络结构原理示意图，黄色的节点分别是前向和后向 RNN 的输出，示例中的是利用最后一个词的结果直接接全连接层 softmax 输出了。



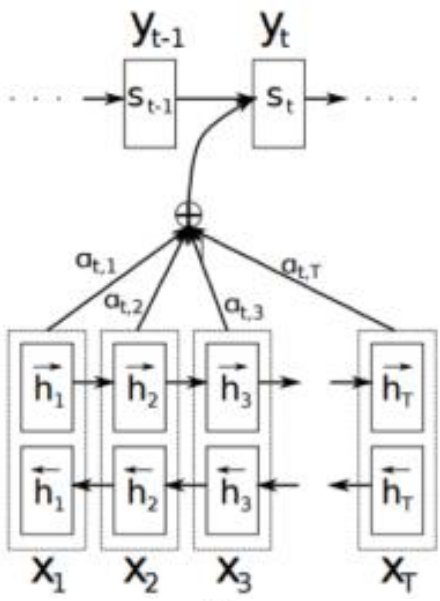
4）TextRNN + Attention

CNN 和 RNN 用在文本分类任务中尽管效果显著，但都有一个不足的地方就是不够直观，可解释性不好，特别是在分析 badcase 时候感受尤其深刻。而注意力（Attention）机制是自然语言处理领域一个常用的建模长时间记忆机制，能够很直观的给出每个词对结果的贡献，基本成了 Seq2Seq 模型的标配了。实际上文本分类从某种意义上也可以理解为一种特殊的 Seq2Seq，所以考虑把 Attention 机制引入近来，研究了下学术界果然有类似做法。

Attention 机制介绍：

详细介绍 Attention 恐怕需要一小篇文章的篇幅，感兴趣的可参考 14 年这篇 paper NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE。

以机器翻译为例简单介绍下，下图中 x_t是源语言的一个词，y_t 是目标语言的一个词，机器翻译的任务就是给定源序列得到目标序列。翻译 y_t的过程产生取决于上一个词 y_{t-1}和源语言的词的表示 h_j（x_j 的 bi-RNN 模型的表示），而每个词所占的权重是不一样的。比如源语言是中文 “我 / 是 / 中国人” 目标语言 “i / am / Chinese”，翻译出“Chinese”时候显然取决于“中国人”，而与“我 / 是”基本无关。下图公式，$alpha_{ij}$则是翻译英文第 i个词时，中文第 j个词的贡献，也就是注意力。显然在翻译“Chinese”时，“中国人”的注意力值非常大。

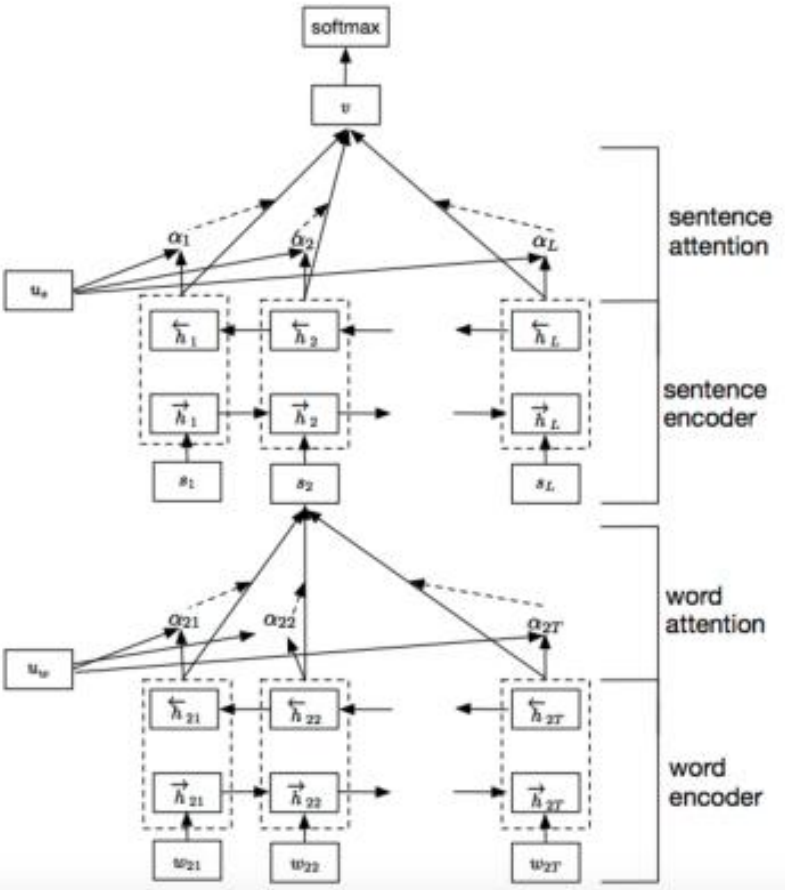


$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j. \qquad (5)$$

Attention 的核心 point 是在翻译每个目标词（或 预测商品标题文本所属类别）所用的上下文是不同的，这样的考虑显然是更合理的。

TextRNN + Attention 模型：

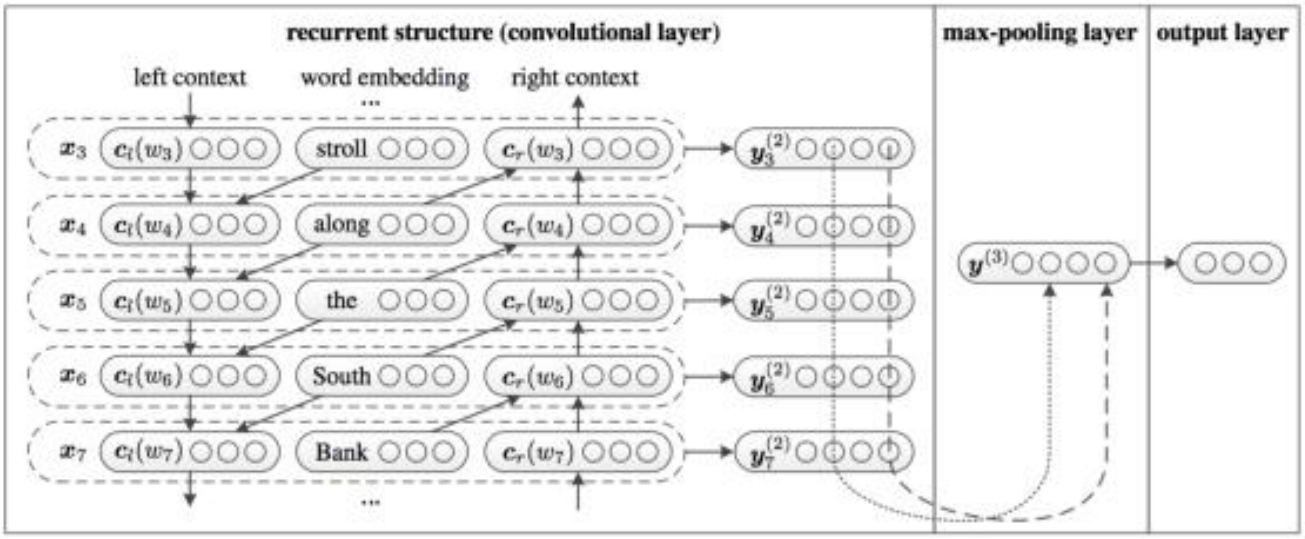
我们参考了这篇文章 Hierarchical Attention Networks for Document Classification，下图是模型的网络结构图，它一方面用层次化的结构保留了文档的结构，另一方面在 word-level 和 sentence-level。淘宝标题场景只需要 word-level 这一层的 Attention 即可。



加入 Attention 之后最大的好处自然是能够直观的解释各个句子和词对分类类别的重要性。

5）TextRCNN（TextRNN + CNN）

我们参考的是中科院 15 年发表在 AAAI 上的这篇文章 Recurrent Convolutional Neural Networks for Text Classification 的结构：



利用前向和后向 RNN 得到每个词的前向和后向上下文的表示：

$$\boldsymbol{c}_l(w_i) = f(W^{(l)}\boldsymbol{c}_l(w_{i-1}) + W^{(sl)}\boldsymbol{e}(w_{i-1})) \quad (1)$$

$$\boldsymbol{c}_r(w_i) = f(W^{(r)}\boldsymbol{c}_r(w_{i+1}) + W^{(sr)}\boldsymbol{e}(w_{i+1})) \quad (2)$$

这样词的表示就变成词向量和前向后向上下文向量 concat 起来的形式了，即：

$$\boldsymbol{x}_i = [\boldsymbol{c}_l(w_i); \boldsymbol{e}(w_i); \boldsymbol{c}_r(w_i)] \quad (3)$$

最后再接跟 TextCNN 相同卷积层，pooling 层即可，唯一不同的是卷积层 filter_size = 1 就可以了，不再需要更大 filter_size 获得更大视野，这里词的表示也可以只用双向 RNN 输出。

三、一点经验

理论和实践之间的 Gap 往往差异巨大，学术 paper 更关注的是模型架构设计的新颖性等，更重要的是新的思路；而实践最重要的是在落地场景的效果，关注的点和方法都不一样。这部分简单梳理实际做项目过程中的一点经验教训。

模型显然并不是最重要的：不能否认，好的模型设计对拿到好结果的至关重要，也更是学术关注热点。但实际使用中，模型的工作量占的时间其实相对比较少。虽然再第二部分介绍了 5 种 CNN/RNN 及其变体的模型，实际中文本分类任务单纯用 CNN 已经足以取得很不错的结果了，我们的实验测试 RCNN 对准确率提升大约 1%，并不是十分的显著。最佳实践是先用 TextCNN 模型把整体任务效果调试到最好，再尝试改进模型。

理解你的数据：虽然应用深度学习有一个很大的优势是不再需要繁琐低效的人工特征工程，然而如果你只是把他当做一个黑盒，难免会经常怀疑人生。一定要理解你的数据，记住无论传统方法还是深度学习方法，数据 sense 始终非常重要。要重视 badcase 分析，明白你的数据是否适合，为什么对为什么错。

关注迭代质量 - 记录和分析你的每次实验：迭代速度是决定算法项目成败的关键，学过概率的同学都很容易认同。而算法项目重要的不只是迭代速度，一定要关注迭代质量。如果你没有搭建一个快速实验分析的套路，迭代速度再快也只会替你公司心疼宝贵的计算资源。建议记录每次实验，实验分析至少回答这三个问题：为什么要实验？结论是什么？下一步怎么实验？

超参调节：超参调节是各位调参工程师的日常了，推荐一篇文本分类实践的论文 A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification，里面贴了一些超参的对比实验，如果你刚开始启动文本分析任务，不妨按文章的结果设置超参，怎么最快的得到超参调节其实是一个非常重要的问题，可以读读 萧瑟的这篇文章 深度学习网络调参技巧 - 知乎专栏。

一定要用 dropout：有两种情况可以不用：数据量特别小，或者你用了更好的正则方法，比如 bn。实际中我们尝试了不同参数的 dropout，最好的还是 0.5，所以如果你的计算资源很有限，默认 0.5 是一个很好的选择。

fine-tuning 是必选的：上文聊到了，如果只是使用 word2vec 训练的词向量作为特征表示，我赌你一定会损失很大的效果。

未必一定要 softmax loss： 这取决与你的数据，如果你的任务是多个类别间非互斥，可以试试着训练多个二分类器，我们调整后准确率还是增加了>1%。

类目不均衡问题：基本是一个在很多场景都验证过的结论：如果你的 loss 被一部分类别 dominate，对总体而言大多是负向的。建议可以尝试类似 boostrap 方法调整 loss 中样本权重方式解决。

避免训练震荡：默认一定要增加随机采样因素尽可能使得数据分布 iid，默认 shuffle 机制能使得训练结果更稳定。如果训练模型仍然很震荡，可以考虑调整学习率或 mini_batch_size。

没有收敛前不要过早的下结论：玩到最后的才是玩的最好的，特别是一些新的角度的测试，不要轻易否定，至少要等到收敛吧。

四、写在最后

几年前校招面阿里时，一面二面聊的都是一个文本分类的项目（一个新浪微博主题分类的学校课题项目），用的还是文中介绍的传统的做法。面试时对特征项处理和各个分类器可谓如数家珍，被要求在白板上写了好几个特征选择公式，短短几年传统做法已经被远远超越，不得不感慨深度学习的发展。

值得感慨的一方面是今天技术的发展非常快，故步自封自然是万万万万不可取，深知还有很多理论尚且不懂还要继续深读 paper；另一方面，理解理论原理和做好项目间实际非常有巨大的 gap，特别是身处工业界的同仁们，学术圈值得钻但要把握分寸，如果仅仅追逐技术深度，不免容易陷入空中阁楼。

最后老规矩再次安利下我们 team 的招聘，对淘宝搜索排序和自然语言处理方向感兴趣的同学欢迎邮件我 qingsong.huaqs@taobao.com，来淘宝，一起成长！

以上，感谢阅读。

深度学习（Deep Learning）自然语言处理文本分类

知乎跟帖（14 小时前）节选：

 Weijie Huang 写得很好啊 我 thesis 也是做这个 topic 对比了一下 cnn 下中文文本和拼音文本的效率提升问题。中文效果蛮好的，dictionary 影响很大，pre-trained 意义不大 lookup table 也能胜任。。。1 赞 13 小时前

 清淞（作者）回复 Weijie Huang 我们做下来看 pretrain 意义的确也不大，文章分享来读读啊 13 小时前

 Weijie Huang 回复清淞（作者）链接在这里 https://arxiv.org/abs/1611.04358。 intro 有小 bug 是第一个用在 raw character words 而不是 chinese character，基准文是从 Zhang 的 2015NIPS 开始看起的 感觉他总结的蛮好的 这个 topic 我觉得基本扫完那 5-6 篇(之前的 CNN 韩国 phd 的,nips15 这篇的 之后 fb 两个组分别做的 fasttext 和超级深度的..额外点了技能点的有 zhang 同组的 cnn+rnn 的,富士通的 graph 做的,,,)就知道文本分类用于深度学习怎么做了...pretrain 这几篇按照瞬间顺序一开始是花样 pretrain 然后是 pretrain 然后 lookup table 我觉得还是很有道理的...1 赞 13 小时前[2]

 GreatxXz 原来淘宝后台是根据商品标题预测其所在叶子类目的，我一直以为是完全由店铺商家自主选取分类的。通过您的文章，再次感受到了神经网络的技术魅力~：）9 小时前

II. Python 在 CV、NLP、ML 和 DM 等六大方面的资源汇总

文章来源：转载自 GitHub 人工智能深度内参[3]

计算机视觉

SimpleCV：开源计算机视觉框架，可以访问如 OpenCV 等高性能计算机视觉库使用 Python 编写，可以在 Mac、Windows 以及 Ubuntu 上运行。（http://simplecv.org/）

自然语言处理

NLTK：一个领先的平台，用来编写处理人类语言数据的 Python 程序。（http://www.nltk.org/）

Pattern：Python 可用的 web 挖掘模块，包括自然语言处理、机器学习等工具。（http://www.clips.ua.ac.be/pattern）

TextBlob：为普通自然语言处理任务提供一致的 API，以 NLTK 和 Pattern 为基础，并和两者都能很好兼容。（http://textblob.readthedocs.io/en/dev/）

jieba：中文断词工具。（https://github.com/fxsjy/jieba#jieba-1）

SnowNLP：中文文本处理库。（https://github.com/isnowfy/snownlp）

loso：另一个中文断词库。（https://github.com/fangpenlin/loso）

genius：基于条件随机域的中文断词库。（https://github.com/duanhongyi/genius）

nut：自然语言理解工具包。（https://github.com/pprett/nut）

通用机器学习

Bayesian Methods for Hackers：Python 语言概率规划的电子书。（https://github.com/CamDavidsonPilon/%E3%80%82Probabilistic-Programming-and-Bayesian-Methods-for-Hackers）

MLlib in Apache Spark：Spark 下的分布式机器学习库。（http://spark.apache.org/docs/latest/mllib-guide.html）

scikit-learn：基于 SciPy 的机器学习模块。（http://scikit-learn.github.io/stable）

graphlab-create：包含多种机器学习模块的库（回归、聚类、推荐系统、图分析等），基于可以磁盘存储 DataFrame（http://graphlab.com/products/create/docs/）

BigML：连接外部服务器的库。（https://bigml.com/）

pattern：Python 的 web 挖掘模块。（https://github.com/clips/pattern）

NuPIC：Numenta 公司的智能计算平台。（https://github.com/numenta/nupic）

Pylearn2：基于 Theano 的机器学习库。（https://github.com/lisa-lab/pylearn2）

hebel：Python 编写的使用 GPU 加速的深度学习库。（https://github.com/hannes-brt/hebel）

gensim：主题建模工具。（https://github.com/RaRe-Technologies/gensim）

PyBrain：另一个机器学习库。（https://github.com/pybrain/pybrain）

Crab：可扩展的、快速推荐引擎。（https://github.com/muricoca/crab）

python-recsys：Python 实现的推荐系统。（https://github.com/ocelma/python-recsys）

thinking bayes：关于贝叶斯分析的书籍。（https://github.com/AllenDowney/ThinkBayes）

Restricted Boltzmann Machines：Python 实现的受限波尔兹曼机。（https://github.com/echen/restricted-boltzmann-machines）

Bolt：在线学习工具箱。（https://github.com/pprett/bolt）

CoverTree：cover tree 的 Python 实现，scipy.spatial.kdtree 便捷的替代。（https://github.com/patvarilly/CoverTree）

nilearn：Python 实现的神经影像学机器学习库。（https://github.com/nilearn/nilearn）

Shogun：机器学习工具箱。（https://github.com/shogun-toolbox/shogun）

Pyevolve：遗传算法框架。（https://github.com/perone/Pyevolve）

Caffe：考虑了代码清洁、可读性及速度的深度学习框架。（http://caffe.berkeleyvision.org/）

breze：深度及递归神经网络的程序库，基于 Theano。（https://github.com/breze-no-salt/breze）

数据分析/数据可视化

SciPy：基于 Python 的数学、科学、工程开源软件生态系统。（https://www.scipy.org/）

NumPy：Python 科学计算基础包。（http://www.numpy.org/）

Numba：Python 的低级虚拟机 JIT 编译器，Cython and NumPy 的开发者编写，供科学计算使用。（http://numba.pydata.org/）

NetworkX：为复杂网络使用的高效软件。（https://networkx.github.io/）

Pandas：这个库提供了高性能、易用的数据结构及数据分析工具。（http://pandas.pydata.org/）

Open Mining：Python 中的商业智能工具（Pandas web 接口）。（https://github.com/mining/mining）

PyMC：MCMC 采样工具包。（https://github.com/pymc-devs/pymc）
zipline：Python 的算法交易库。（https://github.com/quantopian/zipline）
PyDy：全名 Python Dynamics，协助基于 NumPy、SciPy、IPython 以及 matplotlib 的动态建模工作流。（http://www.pydy.org/）
SymPy：符号数学 Python 库。（https://github.com/sympy/sympy）
statsmodels：Python 的统计建模及计量经济学库。（https://github.com/statsmodels/statsmodels）
astropy：Python 天文学程序库，社区协作编写。（http://www.astropy.org/）
matplotlib：Python 的 2D 绘图库。（http://matplotlib.org/）
bokeh：Python 的交互式 Web 绘图库。（https://github.com/bokeh/bokeh）
plotly：Python and matplotlib 的协作 web 绘图库。（https://plot.ly/python/）
vincent：将 Python 数据结构转换为 Vega 可视化语法。（https://github.com/wrobstory/vincent）
d3py：Python 的绘图库，基于 D3.js。（https://github.com/mikedewar/d3py）
ggplot：和 R 语言里的 ggplot2 提供同样的 API。（https://github.com/yhat/ggpy）
Kartograph.py：Python 中渲染 SVG 图的库，效果漂亮。（https://github.com/kartograph/kartograph.py）
pygal：Python 下的 SVG 图表生成器。（http://pygal.org/en/stable/）
pycascading（https://github.com/twitter/pycascading）

杂项脚本/iPython 笔记/代码库

pattern_classification：（https://github.com/rasbt/pattern_classification）
thinking stats 2：（https://github.com/Wavelets/ThinkStats2）
hyperopt：（https://github.com/hyperopt/hyperopt-sklearn）
numpic：（https://github.com/numenta/nupic）
2012-paper-diginorm：（https://github.com/dib-lab/2012-paper-diginorm）
ipython-notebooks：（https://github.com/ogrisel/notebooks）
decision-weights：（https://github.com/CamDavidsonPilon/decision-weights）
Sarah Palin LDA：Sarah Palin 关于主题建模的电邮。（https://github.com/Wavelets/sarah-palin-lda）
Diffusion Segmentation：基于扩散方法的图像分割算法集合。（https://github.com/Wavelets/diffusion-segmentation）
Scipy Tutorials：SciPy 教程，已过时，请查看 scipy-lecture-notes。（https://github.com/Wavelets/scipy-tutorials）
Crab：Python 的推荐引擎库。（https://github.com/marcelcaraciolo/crab）
BayesPy：Python 中的贝叶斯推断工具。（https://github.com/maxsklar/BayesPy）
scikit-learn tutorials：scikit-learn 学习笔记系列。（https://github.com/GaelVaroquaux/scikit-learn-tutorial）
sentiment-analyzer：推特情绪分析器。（https://github.com/madhusudancs/sentiment-analyzer）
group-lasso：坐标下降算法实验，应用于（稀疏）群套索模型。（https://github.com/fabianp/group_lasso）
mne-python-notebooks：使用 mne-python 进行 EEG/MEG 数据处理的 IPython 笔记。（https://github.com/mne-tools/mne-python-notebooks）
pandas cookbook：使用 Python pandas 库的方法书。（https://github.com/jvns/pandas-cookbook）
climin：机器学习的优化程序库，用 Python 实现了梯度下降、LBFGS、rmsprop、adadelta 等算法。（https://github.com/BRML/climin）

Kaggle 竞赛源代码

wiki challange：Kaggle 上一个维基预测挑战赛 Dell Zhang 解法的实现

(https://github.com/hammer/wikichallenge)

kaggle insults：Kaggle 上”从社交媒体评论中检测辱骂“竞赛提交的代码
(https://github.com/amueller/kaggle_insults)

kaggle_acquire-valued-shoppers-challenge：Kaggle 预测回头客挑战赛的代码
(https://github.com/MLWave/%E3%80%82kaggle_acquire-valued-shoppers-challenge)

kaggle-cifar：Kaggle 上 CIFAR-10 竞赛的代码，使用 cuda-convnet
(https://github.com/zygmuntz/kaggle-cifar)

kaggle-blackbox：Kaggle 上 blackbox 赛代码，关于深度学习
(https://github.com/zygmuntz/kaggle-blackbox)

kaggle-accelerometer：Kaggle 上加速度计数据识别用户竞赛的代码。
(https://github.com/zygmuntz/kaggle-accelerometer)

kaggle-advertised-salaries：Kaggle 上用广告预测工资竞赛的代码
(https://github.com/zygmuntz/kaggle-advertised-salaries)

kaggle amazon：Kaggle 上给定员工角色预测其访问需求竞赛的代码
(https://github.com/zygmuntz/kaggle-amazon)

kaggle-bestbuy_big：Kaggle 上根据 bestbuy 用户查询预测点击商品竞赛的代码（大数据版）。
(https://github.com/zygmuntz/kaggle-bestbuy_big)

kaggle-bestbuy_small：Kaggle 上根据 bestbuy 用户查询预测点击商品竞赛的代码（小数据版）。
(https://github.com/zygmuntz/kaggle-bestbuy_small)

Kaggle Dogs vs. Cats：Kaggle 上从图片中识别猫和狗竞赛的代码。
(https://github.com/kastnerkyle/kaggle-dogs-vs-cats)

Kaggle Galaxy Challenge：Kaggle 上遥远星系形态分类竞赛的优胜代码。
(https://github.com/benanne/kaggle-galaxies)

Kaggle Gender：Kaggle 竞赛，从笔迹区分性别。
(https://github.com/zygmuntz/kaggle-gender)

Kaggle Merck：Kaggle 上预测药物分子活性竞赛的代码（默克制药赞助）
(https://github.com/zygmuntz/kaggle-merck)

Kaggle Stackoverflow：Kaggle 上 预测 StackOverflow 网站问题是否会被关闭竞赛的代码
(https://github.com/zygmuntz/kaggle-stackoverflow)

wine-quality：预测红酒质量
(https://github.com/zygmuntz/wine-quality)

III. 文本语义解析能不能深度学习训练？

秦陇纪 10 言：对知识里面文本语义的简化只能从建立庞大又复杂的知识库和标记语言，完备知识图谱入手。类似 AlphaGo 和 FineArt 的策略网络，在语义解析和语义推理方面，句子里面不同语序和词尾变化等是不是可以借鉴图像分类的方法呢？对文本语义特征的挖掘和训练方法，也许可以参考对图像和策略网络做分类训练的方法，这个领域也很难做、也非常耗时。

参考文献

[1] 清淞. 用深度学习（CNN RNN Attention）解决大规模文本分类问题 - 综述和实践[EB/OL]. https://zhuanlan.zhihu.com/p/25928551,2017-03-26.

[2] Weijie Huang. Character-level Convolutional Network for Text Classification Applied to Chinese Corpus[EB/OL]. https://arxiv.org/abs/1611.04358,2017-03-26.

[3] github，人工智能深度内参. Python 在 CV、NLP、ML 和 DM 等六大方面的资源汇总[EB/OL]. https://mp.weixin.qq.com/s?__biz=MzIzODQ5Njc3Ng==&mid=2247483813&idx=5&sn=7f4d57466bc1a8490ba71fc2ed5a6c21&chksm=e9393cc2de4eb5d4b7c3dfc28d8113d3e5c8448b317ff3e70220ba0f524052232d284643f851,2017-03-26.

[4] 秦陇纪. 数据科学与大数据技术专业概论；人工智能研究现状及教育应用；文献数据共词分析；大数据简化之技术体系[EB/OL]. 数据简化 DataSimp（微信公众号）,2017-08-23.

文末打赏后“阅读原文”可百度网盘下载完整版（带跳转链接）PDF 文档。

Appx. 附录(5251 字)

内附. 2017 年 3 月 27 日（星期一）农历丁酉年二月卅新闻四则汇编(4912 字)

附 i. 早报,3 月 27 日,星期一

1、最高检派员审查辱母案 将调查警察执法是否渎职;
2、北京“商住房”全面限购：在建在售商办项目不得卖给个人;
3、朝鲜警告将先发制人 以特种作战粉碎美韩“斩首”图谋;
4、南京一公墓推出扫墓新方式：微信直播代扫，可观看实时画面;
5、北京法院撤销停售 iPhone 6 决定：确认不侵权;
6、聊城回应于欢案：对警察不作为、高利贷、涉黑开展调查
7、南京南站一男子被卡在站台缝隙里 被救后无生命体征;
8、林毅夫：2030 年中国将成世界第一大经济体;
9、Uber 无人驾驶汽车发生撞车事故 测试项目暂停;
10、广东中山市购房实施限购 非户籍购新房需半年社保;
11、刚果一群武装分子袭击当地警方 斩首 40 名警察;
12、北京首次实施免费自然葬：罐体可降解 不保留骨灰;
【微语】要成功，需要朋友，要取得巨大的成功，需要敌人。

附 ii. 2017 年 3 月 27 日周一读报！一切美好从“珍惜”开始！

1、2016 年我国“独角兽”企业已达 131 家，其中技术驱动型占绝大多数。目前已经有 16 个城市出现了“独角兽”企业，其中，北京、上海、深圳、杭州成为我国“独角兽”企业主要集聚区域，均超过 10 家。（“独角兽”企业是指成立 10 年以内、估值超过 10 亿美元、获得过私募投资且尚未上市的企业。）

2、西安外国语大学被疑多收学费，校方回应：已累计退费 960 万，但 2012 年度以前多收的学费，学生毕业，联系困难，正在收集账户信息。（近日，有学生反映西安外国语大学艺术类专业收取学费 11000 元，超过物价部门规定的 9000 元标准。）

3、北京“商住房”全面限购：在建在售商办项目不得卖给个人。已销售的商办类项目再次上市出售时，可出售给企事业单位、社会组织，也可出售给个人。个人购买应符合以下条件：①名下在京无住房和商办类房产记录的；②在申请购买之日起，在京已连续五年缴纳社会保险或者连续五年缴纳个人所得税。

4、心碎！沪天潼路浙江北路路口大巴撞 ofo 单车，骑车男童身亡。据悉，该男孩今年 10 岁左右，读小学 4 年级，此事或系目前上海首例不满 12 岁未成年人使用共享单车致死案例。

5、聊城回应于欢案：对警察不作为、高利贷、涉黑开展调查。(律师：法院支持的民间借贷最高利率为 24%，对年利率超过 36%的，超过部分的利息约定无效。本案中，双方约定月息 10%，相当于年利率 120%，远超过法定的最高利率，显然不合法的。于欢母亲于 2014 年 7 月借款 100 万，截至讨债之日，还款的数额无论是 187 万加房产，还是 152 万，按照 36%的最高利率计算，实际已经还清本息。之后发生的追债行为是违法的。）

6、林郑月娥以 777 张有效选票当选香港特区第五任行政长官。成为香港首位女特首。(有媒体迅速关注到其家庭情况，并称长子林节思目前在北京任职小米科技营运经理。资料显示，林节思 2016 年 4 月加盟小米任营运经理，从事软件开发。）

7、联合国：由于亚洲地区收入水平普遍提高，数亿人口开始成了数字世界的一份子，但随之而来的电子垃圾问题却无法忽视，最近五年亚洲地区的电子垃圾暴增 63%。

8、中国人觉得自己崛起了，依然不想多管别国闲事。皮尤调查显示，对于“我们国家的世界影响力比 10 年前高”，75%的中国人认为中国世界影响力比十年前不知道高到哪里去了。关于“我们国家应该帮助其他国家解决难题”，只有 22%的中国人表示赞同。56%中国人期待政府把注意力放在国内还没解决的麻烦上。

9、俄罗斯汽车产业分析机构 Autostat 表示，2 月份中国汽车在俄罗斯的销售额下降到 1900 辆，与 2016 年相比，同比下降 28%。2 月中国汽车在俄销售量冠军是力帆。此品牌汽车卖出 1161 辆，显示出 11%的年增长率。

10、乐天为中国分部紧急注资 3600 亿韩元，乐天玛特在中国的经营业务实际上已陷入瘫痪。乐天玛特相关人士解释称，“由于停业，在华门店没有销售额。但是采购商品和支付工资等方面需要资金，因此决定投资”。乐天集团会长辛东彬表示，“我热爱中国，也想继续在中国开展事业”。

11、日媒：华裔男子孙伟伟违法雇佣上百名卖淫者东京卖淫，其中多人是来自中国的女留学生。目前，孙伟伟因被指控违法雇佣禁止涉足色情店内服务的人员而被拘捕。估计在 2008 年 1 月至今年 2 月期间，该色情店累计营业额高达 5 亿日元(约合人民币 3090 万元)以上。(日本法律禁止留学生进入色情领域打工服务。）

12、人生如棋局，有时眼见的胜负，转眼就峰回路转。棋子总是越来越少，人生总是越来越短，所谓的举手无回，也如同人生的不可重来，珍惜当下，不留遗憾，且行且珍惜。

美好一天从“珍惜”开始！

附 iii. 2017 年 3 月 27 日（丁酉鸡年二月三十）周一 / 早读分享：

1、【最高检派员调查＂辱母杀人案＂】于欢到底是不是正当防卫？办案民警有无渎职失职？… 上亿条评论刷屏的背后，公众在＂尊重专业判决 vs 法官不近人情＂之间，争议巨大。随着该案的深入调查，我们期待法律给出一个令人信服的正义理据，或做出正义的修订。所谓的＂天理难容＂就是公众的关切！

2、【林郑月娥当选香港特区第五任行政长官】昨天上午，香港特区行政长官选举结果揭晓，林郑月娥获得 777 票，曾俊华获得 365 票，胡国兴获得 21 票。林当选。

3、【周小川：货币政策不是万金油 宽松已经到达周期的尾部】＂经过多年的量化宽松，我们已经到达周期的尾部，货币政策不再是宽松政策。＂央行行长周小川表示，货币政策不是万金油，不要以为量化宽松的货币政策可以治好每个国家的不同疾病。中国政府不会依赖于＂直升机撒钱＂式刺激措施。

4、【樊纲：要为＂非正规＂就业正名】国民经济研究所所长樊纲 3 月 25 日说，从 1 电子商务和共享经济的大趋势来看，未来＂非正规＂就业的人会越来越多，要从社会保障、社会观念、组织归属等方面提供支持。

5、【北京市：中小学不得与地产商合作办学】北京市教委：本书所有中小学校未经市教委批准不得到外地办学，各中小学不得与

房地产商合作办学。另外：预计到 2019 年，全市 91 所优质高中将安排 18325 个招生计划用于校额到校工作。届时，全市 411 所一般公办初中的 3 万余名学生将可享受政策。

6、【社保卡未来将增加新功能】根据人社部《关于加快推进社会保障卡应用的意见》，未来社保卡可分九大类 102 项功能。包括可作为享受公共就业服务、就业扶持政策和参保登记、医疗费用报销等业务办理的身份凭证；当银行卡使用，具有现金存取、转账等服务；可当公交卡刷；可缴纳、领取养老金；就医和异地费用结算；定点买药刷卡；可代发工资等。

7、【大商所调整相关品种交易保证金标准和涨跌停板幅度】3 月 30 日结算时起，将铁矿石品种期货合约涨跌停板幅度和最低交易保证金标准调整至 9%和 11%。4 月 5 日恢复交易后，自铁矿石品种持仓量最大的两个合约未出现单边市的第一个交易日结算时起，铁矿石品种期货合约涨跌停板幅度和最低交易保证金标准分别恢复至 8%和 10%。

8、【吴晓求：美股和中国房价是最大的泡沫】中国人民大学副校长吴晓求在博鳌亚洲论坛年会上表示，现在全球存在的两大资产泡沫就是美国的股市和中国的楼市。北京市的四环房价都涨到了每平方米 10 万元或以上，这种趋势是难以为继的。

9、【北京商改住产品限购靴子落地：个人不得购买】26 日晚间北京市发布《关于进一步加强商业、办公类项目管理的公告》。个人购买可售的商办类项目，需名下在京无住房和商办类房产记录的，并且需在申请购买之日起，在京已连续五年缴纳社会保险或者连续五年缴纳个人所得税。此外，商业银行暂停对个人购买商办类项目的个人购房贷款。

10、【仙言潮声】

人在旅途，一定要记得遵从天理、良心和秩序，上车必须买票，吃饭一定要买单，不多给，也不少给！与人争吵，可以讲理论事，但不要出言羞辱。讲不通，耸耸肩膀摊摊手，转身离开……美好的时光，不能被自己浪费了！

美好的一天从珍惜自己开始！

附 iv. 2017 年 3 月 27 日（星期一）农历丁酉年二月卅“癸丑日”

每天三分钟　知晓天下事

A、【国内】

1）[博鳌论坛]　张高丽：支持外商投资企业在国内上市和发债；周小川回应货币宽松导致楼市泡沫：这不是预期后果；货币政策不是万金油，近些年公众过度关注；大咖诊脉世界经济，为全球化“正视听”“一　带一路”赋予经济全球化新内涵，“中国理念”激发世人新期待；博鳌亚洲论坛发布促进经济全球化宣言；

2）香港特区第五任行政长官选举结果 26 日揭晓，林郑月娥(女)以 777 票当选第五任香港特首，梁振英衷心祝贺林郑月娥女士当选，表示与候任行政长官做好交接，并全力支持新一届政府的筹组工作；林郑月娥：将竭尽所能维护“一国两制”；

3）[法制与反腐]　中纪委副书记李书磊兼任中央追逃办主任；南昌市纪委印发《通知》严明清明期间纪律要求；驻陕西省委办公厅纪检组组长因公款吃喝被免；山东聊城全面调查警察不作为、高利贷、涉黑犯罪等问题；

4）铁总回应东南高铁调价：部分票价仍明显低于公路；

5）西安“问题电缆”涉事公司财产被查封，对其银行存款 105 万余元予以冻结；

6）四川渠县选拔 46 名退役军人任贫困村“第一书记”，优秀退役军人成脱贫攻坚排头兵；

7）人民日报：食堂浪费现象调查，中央机关有人饭菜没动就倒掉；

8）[军事]　驻京十五所军队医院全部参加北京市医药分开综合改革，官兵和老干部免费医疗政策不变；

9）[港澳]　国务院港澳办：林郑月娥符合“4 个标准”；香港中联办：对林郑月娥当选表示祝贺；香港海关检控运送新加坡军车到香港的船运公司和船长，案件被押后至 5 月 19 日进行审理；

10）[台湾]　张志军：只有坚持“九二共识”，两岸关系才能行稳致远、造福民众；洪秀柱：国民党一定要打破酱缸、大佬和宫廷文化；赏萤火虫成台湾新流行；陈德铭在博鳌论坛隔空邀田弘茂：欢迎到他自己国家的大陆来。

B、【国际】

1）李克强出席澳大利亚华侨华人举行的欢迎晚宴并致辞，呼吁留学生报效祖国；李克强携夫人程虹抵达惠林顿机场，开始对新西兰进行正式访问；

2）马云首次回应 NASA 计划：阿里正在向未来投资，我们投资的每项技术，都是为了赋能小企业、赋能年轻人、赋能发展中国家；

3）美共和党撤回新医保案，特朗普施行“新政”再遭重挫；

4）美财长：即将推税改方案，对个人和企业税全面改革；

5）俄驻车臣部队遭遇恐怖袭击，双方交火士兵 6 死 3 伤，普京要求团结各方力量，共同打击恐怖主义；

6）杜特尔特邀中国海军舰队访菲，称要亲自登解放军舰船；菲总统称愿与中国共享南海资源：我们连设备都买不起；纪念《罗马条约》签署 60 周年欧盟特别峰会签署《罗马宣言》；

7）朝鲜军方回应美韩“斩首行动”演习：将以朝鲜式“特殊作战”粉碎美韩阴谋；

8）乐天为中国分部紧急注资 3600 亿韩元，继续开拓业务。

C、【财经证券】

1）央行货币政策委员会委员樊纲：外储降至 2 万亿美元可能是好事；

2）住建部拟提出新一轮棚改计划，于 2020 年前完成改造；

3）上海清算所：加强金融设施互联互通，推动债市开放；

4）证监会将对有能力但长期不分红上市公司加强监管。

D、【文教体娱】

1）基础学科拔尖学生培养试验计划研讨会在山东大学召开，教育部高等教育司司长张大良在研讨会现场发言；

2）北京市教委：中小学不得与房地产商合作办学；北京出台措施为学区房降温，幼升小摇号确定学位；

3）山东今夏起高考不再分一二本，取消少数民族加分；南京中小学将有望开设舞蹈、戏剧、戏曲等课程，到 2020 年，艺术素质测评将计入中考成绩；

4）国足伊朗首训遭对手盘外招，训练场草皮堪比菜地；国足战伊朗迎三大利好，12 强赛将重燃出线希望；惠若琪在南京赏樱花，身高 1 米 92，人丛中回头率最高。

E、【生活服务】

1）北京 4 月 10 日至 7 月 9 日星期一至星期五限行机动车车牌尾号分别为：3 和　8、4 和 9、5 和 0、1 和 6、2 和 7，纯电动汽车除外；

2）上海地铁：已全面排查 5 年内线路，未发现奥凯电缆；

3）广州公务员考试开始，今年首设京穗两地考区揽人才；

4）琼中“三月三”发出邀请函，请到最美乡村深呼吸；

5）浙江将探索农户“三权”自愿有偿退出机制，或将其作为进城落户条件。

F、【健康养生】

1）《中国居民膳食指南》2016 版中建议，坚果每周摄入量为 50-70 克，平均一天的摄入量为 10 克左右；假如是吃核桃，去皮的核桃肉的重量，还原成带皮的核桃，大约为 2-3 个。

2）春季是养肝明目的最佳季节，常吃香椿能起到清肝明目的作用，对视力下降及眼干眼疲劳有很好的恢复作用；春季流感容易引起发热，用鱼腥草煮水能快速退烧，尤其对感冒初期效果奇佳，非常适合老人和小孩子饮用。

（编辑：西安知非　自新华、中新、腾讯、凤凰、东方财富网）

外附 v. 数据简化 DataSimp 社区译文志愿者招募启事

“**数据简化 DataSimp**”社区翻译组、媒体组缺少志愿者，①设计黑白静态和三彩色动态社区 LOGO 图标；②翻译美欧 IT 大数据、人工智能、编程开发技术文章的至少投一篇高质量首译美欧数据科学技术论文，方可正式成为数据简化 DataSimp 社区贡献者。非诚勿扰，季度无贡献者自动退出。请扫下面的二维码，加入**数据简化 DataSimp 社区**（实名制微信群，拉人请修改昵称为：姓名-单位-职务）。

数据简化DataSimp社区

（实名制微信群，拉人请修改昵称为：姓名-单位-职务）

Data Simplification and Sciences Wechat and Toutiao Public Account, QinDragon2010@qq.com, 2017.02.27Mon, Xi’an, Shaanxi, China:

LIFE

Life begins at the end of your comfort zone.

-- Neale Donald Walsch

THE DAY

The strength of purpose and the clarity of your vision，along with the tenacity to pursue it，is your underlying driver of success.

-- Ragy Tomas

长按下面二维码“识别图中二维码”关注公众号：数据简化 DataSimp（搜索此名称也行）。

文末打赏后“阅读原文”可下载此文 PDF/论文/压缩包等文档。

（西安秦陇纪 10 数据简化 DataSimp 综合汇编，欢迎有志于**数据简化**之传媒、技术的实力伙伴加入全球“数据简化 DataSimp”社区！欢迎转载注明出处：**秦陇纪 10 数据简化 DataSimp 公众号、头条号“数据简化 DataSimp**、科学 Sciences”汇译编，投稿邮箱 QinDragon2010@qq.com）