问题建模

汉语分词问题可以视为序列标注问题的一种。待分词的句子是一个字符序列，分词结果是对该字符序列的每个字符打上一个标记，单字符构成的词标记为S；多字符构成的词的首字标记为B，词中字符标记为M，词尾字符标记为E。该问题即可描述为：求一个中文字符序列的最佳标记序列。

基于CRF的算法

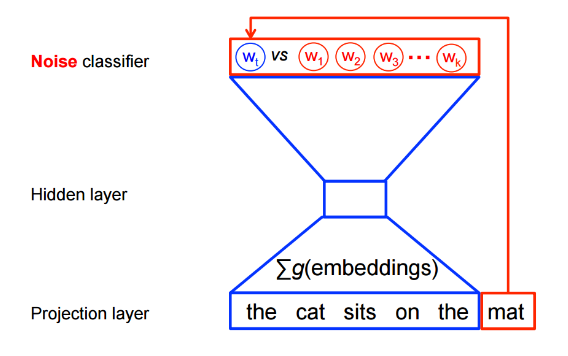
见课件上对CRF的描述

基于字向量、双向LSTM与CRF的算法

因为该问题的输入和输出都是序列，很自然地联想到长短时记忆网络（LSTM）。LSTM在序列建模上有很广泛的应用，如自然语言、语音和视频等。在方法上，我们阅读了*Bidirectional LSTM-CRF Models for Sequence Tagging*这篇论文，并打算复现论文的工作，用于我们最终的系统里面。

因为LSTM的输入是数值表示的张量，而分词的输入是汉语字符，显然汉字不能直接被输入到LSTM里面。一个比较单纯的做法是对汉字进行one-hot编码。也就是说，假设汉语里面有N个汉字字符，每个字符编码出来的向量就有N维，这个向量只有一个维度上是1，其余全是0，用1的位置来区分不同的汉字。这种编码方式的一个缺点是会导致数据过于稀疏，随着字符集的增长，向量的维度也线性增长；另一个缺点是无法反映汉字之间的相互关联，比如“百”和“千”在语义上都是表示数量的，那么应该让它们编码后的距离尽可能小，这样更大程度上保留汉字的语义信息。所以说，一个合适的编码方式对于基于神经网络的算法来说是至关重要的。

虽然PCA、t-SNE等算法可以用来对词语空间降维，并实现词嵌入的可视化与语义感应，但是基于深度学习的算法在表现上越来越有优势，所以我们使用了2013年由谷歌团队发明的词嵌入方法——word2vec来对汉字编码。

词语出现的上下文是很有用的无监督信息，word2vec利用了这一信息，训练一个判断词语间是否上下文相关的分类器，再取该分类器的编码部分作为词语的编码器。训练这个分类器不需要人工标注，只需要取来自相同上下文的词样本对作为正样本，随机采样样本对作为负样本即可构造数据集，具有半监督的特点。

要解释LSTM，首先要提到循环神经网络（RNN），RNN在多层神经网络的基础上增加了时序的概念。RNN单元会在一个时间周期内向外界输出一个状态，RNN的输出取决于上一个时间周期的状态和这个时间周期内的输入，可以表示为【公式】。单纯的RNN在训练的时候会遇到梯度消失和梯度爆炸等问题，所以一般人们都使用LSTM而不是RNN。LSTM相比RNN，增加了输入门，遗忘门，输出门，这些门的加入可以增加网络对序列中信息的选择能力，网络能学习出哪些信息应该保留，哪些信息应该被遗忘。双向LSTM在LSTM的基础上增加了反向的信息流动，在表现上一般优于单向的LSTM。

CRF

如果只使用双向LSTM，当前时刻的输出标签和上一时刻的输出标签互不影响的，也就是说，假如上一时刻输出是B之后，给予特定的输入，当前时刻输出S的可能性也是比较大的，但是“BS”这样的序列是不可能出现的，所以必须对标签序列进行约束。CRF可以承担这个任务，做法是在LSTM的输出之后接一个CRF层。在论文的实验表明加入CRF层能够提升序列标注的精度，证实了标签序列的建模对精度的提升是有效的。

数据准备与实现细节

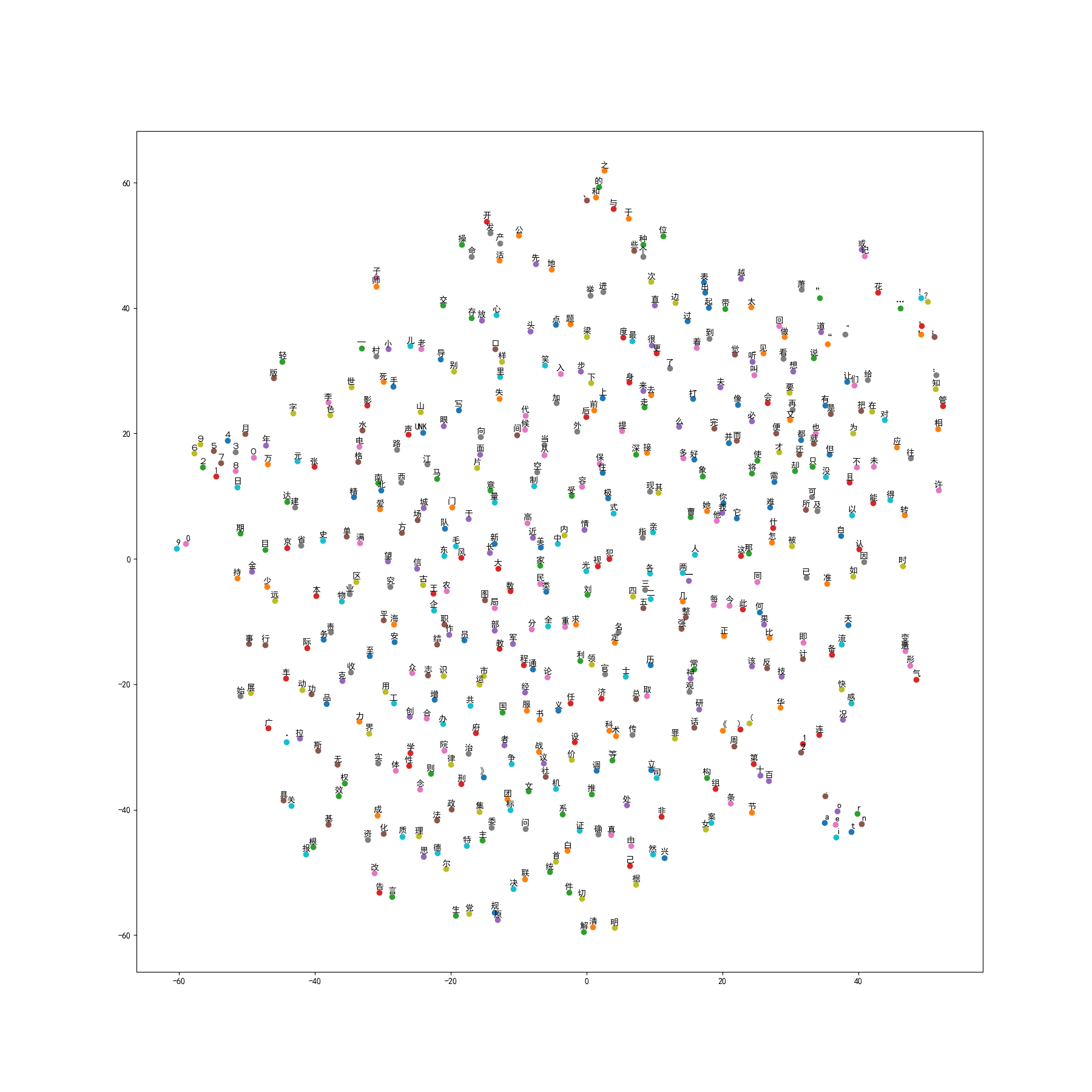
**CRF** 在训练CRF模型时，我们借助了北京大学计算语言研究所提供的1998年1月人民日报语料库。我们把语料库里面的每个词分割开来，并标上类似于“S“、”B“、”E“、”M“的标签。在设计CRF的函数模板时，我们仅考虑字符的unigram特征，没有加入词性作为特征。

**字向量** 训练字向量不需要带标签的数据，只需要提供足够的文本即可，所以我们准备了35本电子书，包含了小说、文史等各方面的内容，总计1000多万字，把这些文本作为训练字向量的数据。对于每个字符，我们只考虑两侧一个字的上下文关系。训练时，我们使用了Adam求解器，采用nce损失函数，负样本的个数为64。

**BiLSTM+CRF** 因为BiLSTM和CRF层直接可以传递梯度，所以这部分的训练是端到端的。训练这部分网络需要大量的监督数据，所以我们只能使用人民日报1998年1月份语料。在对语料的加工上，我们仅利用了词的划分关系，没有利用到词性标注。我们使用Adam作为求解器，学习率初始值为0.0005，损失不再下降后调整为0.0001。

实验与分析

我们首先进行了字向量的训练，我们把字符映射到了128维的向量空间，然后随机取了500个字符用t-sne降维到2维进行可视化，得到下图。可以看出，具有相同语义的字符间的距离很近，与Word2vec的描述是相符的。



我们把取出数据中的90%作为训练集训练模型，10%作为测试集计算模型的在测试集上的正确率、召回率和F1值。为了对比方法之间的差异性，我们分别评估CRF和深度方法的表现。同时，为了证明字向量的重要性，我们对比了随机字向量、自己训出来的字向量和POLYGLOT提供的字向量的结果，如下表所示

|  |  |  |  |
| --- | --- | --- | --- |
|  | 正确率 | 召回率 | F1值 |
| CRF | 0.946392 | 0.945538 | 0.945965 |
| BiLSTM+CRF（随机字向量） | 0.774613 | 0.758323 | 0.766381 |
| BiLSTM+CRF（自己训出来的字向量） | 0.878501 | 0.886340 | 0.882403 |
| BiLSTM+CRF（POLYGLOT提供的字向量） | 0.897666 | 0.903313 | 0.900480 |

可以看出，CRF模型在精确度上高于深度模型。随机生成的字向量在结果上明显比不上精心训练得到的字向量。由于我们训练字向量的数据没有POLYGLOT那么丰富，这种差距也在最终的模型上得到体现。所以，字向量表示的好坏关系到分词结果的好坏。

程序文档

见程序目录下的README.html

微博分词结果（部分）

CRF

兰陵/瓜/贩子/到/沂水/批发/西瓜/被/强买/强/卖/的/后续/关注/】沂/水县/四十/里堡镇/男子/李某某/、/王某/军/、/王某/平/从事/水果/生意/，/今年/5月/15/日/，/三/人/在/网上/发布/西瓜/购销/信息/，/明确/价格/0.5元/每/斤/且/保证/质量/。/兰陵县/客户/赵/某某/从/网上/看到/信息/之后/，/认为/价格/合适/协/商/购销/部分/西瓜/。/5月/17/日/9时/许/，/赵/某某/从/县/赶往/沂水县/李某某/等/人/的/西瓜/摊位/。/经/查看/西瓜/质量/一般/，/赵/某某/不/想/要/了/。/李某某/等/三/人/以/赵/某某/购买/西瓜/达成/意向/，/这/才/从/瓜/农手中/收购/的/，/不/能不要/了/，/而且/近期/由于/西瓜/减产/，/必须/价格/上涨/，/否则/不准/离开/。/无奈/之下/，/赵/某某/装车/西瓜/4000余/斤/，/不论/大/小生熟/全部/装车/了/，/每/斤/还/涨价/0.4元/。/离开/西瓜/交易/地/之后/，/赵/某某/选择/报警/求助/。/接到/警情/之后/，/四十/里堡/派出所/民警/立即/开展/工作/并/将/案情/通报县/局/扫黑/专业/队/。/所/对/联手/开展/案件/侦破/工作/，/经/大量/调查/取证/，/走访/群众/，/网上网/下/工作/，/掌握/大量/证据/。/7月/12/日/，/民警/将/嫌疑人/李某某/等/三/人/抓获/。/经讯/问/，/犯罪/嫌疑人/李某某/、/28/岁/；/王某/军/、/31/岁/、/王某/平/、/31/岁/，/供述/了/在/网上/发布/虚假/信息/，/引诱客户/到来/之后/再/涨价/强迫/交易/的/犯罪/事实/，/采用/相同/方式/作案/多/起/。/

【家人/急/寻/！/兰陵/斜沟/一/学生/离/家/出/走/】/姓名/：/李/永强/年龄/：/15/岁/芦/柞镇/斜沟村/人/，/兰陵县/第二/中学/上学/。/昨天/下午/六点/在/斜沟家中/离家/出/走/，/有/看到/他/的/同学/，/以及/见到/他/的/好心人士/多/加留意/！/家长/万分/着急/！/感恩/感谢/好心人士/帮忙/有/知道/者/请/打电话/：/15092866284

今天/太/热/了/还是/找/个/树阴凉底/避暑/吧/再/去/工地/搬/砖/的/话/恐怕/会/热/的/干哕/

【/三/个/“/熊/孩子/骑/一/电动车/，/撞/上/路边/大货车/】7月/11/日/15/时/30/分/，/在/临沂/兰陵县/火车站/以南/约/400/米/处/路段/，/一/辆/两/轮/电动车/突然/撞/上/路边/停放/的/一/辆/大货车/，/电动车/重重/摔倒/。/兰陵/交警/大队/接到/报警/后/，/民警/迅速/赶到/事发现场/，/伤者/已/送/往/该县/人民/医院/。/骑/电动车/的/小/女孩/面部/严重/挫伤/，/另/一/乘车/小/女孩/的/腿/部/受伤/，/正在/接受/检查/治疗/。/还有/一个/乘车/女孩/有些/擦伤/。/经过/进一步/了解/得知/，/3个/小/女孩/同/骑/一/辆/两/轮/电动车/外出/玩耍/，/最/大/的/15/岁/(/骑车人/)/，/最/小/的/只/有/7岁/，/在/返回/家中/的/路上/发生/交通/事故/。/3/个/孩子/外出/时/，/其/家长/不/知道/，/只是/发现/家中/的/电动车/不见/了/，/直到/听说/孩子/出/了/事/。/民警/提示/，/家长/作为/孩子/的/第一/监护人/，/要/时刻/关注/孩子/的/出行/安全/。/

BiLSTM+CRF（POLYGLOT提供的字向量）、

【/兰陵瓜/贩子/到/沂水/批发/西瓜/被/强买/强卖/的/后续/关注/】/沂水县/四十/里/堡镇/男子/李/某某/、/王/某军/、/王/某平/从事/水果/生意/，/今年/5月/1/5日/，/三/人/在/网上/发布/西瓜/购销/信息/，/明确/价格/0/./5元/每斤/且/保证/质量/。/兰陵县/客户/赵/某某/从/网上/看到/信息/之后/，/认为/价格/合适/协商/购销/部分/西瓜/。/5月/1/7日/9/时许/，/赵/某某/从/县/赶/往/沂水县/李/某/某/等/人/的/西瓜/摊位/。/经查/看/西瓜/质量/一般/，/赵/某某/不/想/要/了/。/李/某某/等/三/人/以/赵/某某/购买/西瓜/达成/意向/，/这/才/从/瓜/农手/中/收购/的/，/不/能/不/要/了/，/而且/近期/由于/西瓜/减产/，/必须/价格/上涨/，/否则/不准/离开/。/无奈/之下/，/赵/某某/装车/西瓜/40/00/余/斤/，/不论/大小/生熟/全部/装车/了/，/每斤/还/涨价/0/./4元/。/离开/西瓜/交易/地/之后/，/赵/某某/选择/报警/求助/。/接到/警情/之后/，/四十/里/堡派/出/所民/警立/即/开展/工作/并/将/案情/通报/县局/扫黑/专业队/。/所/对/联手/开展/案件/侦破/工作/，/经/大量/调查/取证/，/走访/群众/，/网上/网下/工作/，/掌握/大量/证据/。/7月/12日/，/民警/将/嫌疑人/李/某某/等/三/人/抓获/。/经讯/问/，/犯罪/嫌疑人/李/某某/、/2/8岁/；/王/某军/、/3/1岁/、/王/某平/、/3/1岁/，/供述/了/在/网上/发布/虚假/信息/，/引诱/客户/到来/之后/再/涨价/强迫/交易/的/犯罪/事实/，/采用/相同/方式/作案/多起/。/

【/家人/急寻/！/兰陵/斜沟/一/学生/离家/出走/】/姓名/：/李/永强/年龄/：/15/岁/芦柞镇/斜沟/村人/，/兰陵县/第二/中学/上学/。/昨天/下午/六/点/在/斜沟家/中/离家/出走/，/有/看到/他/的/同学/，/以及/见到/他/的/好心/人士/多加/留意/！/家长/万分/着急/！/感恩/感谢/好心/人士/帮忙/有/知道者/请/打电话/：/1509286/6284/

今天/太热/了/还/是/找/个/树阴/凉底/避暑/吧/再/去工/地/搬砖/的/话/恐怕/会/热/的/干哕/

【/三/个/“/熊孩子/骑/一/电动车/，/撞/上/路边/大/货车/】/7月/1/1日/15/时/3/0分/，/在/临沂/兰陵县/火车站/以/南约/400/米处/路段/，/一/辆/两/轮/电动车/突然/撞/上/路边/停放/的/一/辆/大/货车/，/电动车/重重/摔倒/。/兰陵/交警/大队/接到/报警/后/，/民警/迅速/赶到/事发/现场/，/伤者/已/送往/该县/人民/医院/。/骑电/动车/的/小/女孩/面部/严重/挫伤/，/另/一/乘/车小/女孩/的/腿部/受伤/，/正在/接受/检查/治疗/。/还有/一个/乘车/女孩/有些/擦伤/。/经过/进一步/了解/得知/，/3/个/小/女孩/同骑/一/辆/两/轮/电动车/外出/玩耍/，/最/大/的/15/岁/(/骑车人/)/，/最/小/的/只有/7岁/，/在/返/回家/中/的/路上/发生/交通/事故/。/3/个/孩子/外/出时/，/其/家长/不/知道/，/只是/发现家/中/的/电动车/不见/了/，/直到/听说/孩子/出/了/事/。/民警/提示/，/家长/作为/孩子/的/第一/监护人/，/要/时刻/关注/孩子/的/出行/安全/。/

BiLSTM+CRF（自己训出来的字向量）

兰/陵瓜/贩子/到/沂水/批发/西瓜/被/强买/强卖/的/后续关/注】/沂水县/四十/里/堡镇/男子/李/某某/、/王/某军/、/王/某平/从/事/水果/生意/，/今年/5月/15日/，/三/人/在/网上/发布/西瓜/购销/信息/，/明确/价格/0./5元/每斤/且/保证/质量/。/兰陵县/客户/赵/某某/从网/上/看到/信息/之后/，/认为/价格/合适/协商/购销/部分/西瓜/。/5月/17/日9/时许/，/赵/某某/从/县/赶往/沂水县/李/某某/等/人/的/西瓜/摊位/。/经查/看/西瓜/质量/一般/，/赵/某某/不/想/要/了/。/李/某某/等/三/人/以/赵某/某购/买/西瓜/达成/意向/，/这/才/从/瓜/农手/中/收购/的/，/不/能/不/要/了/，/而且/近期/由于/西瓜/减产/，/必须/价格/上涨/，/否则/不准/离开/。/无奈/之下/，/赵/某某/装车/西瓜/40/00/余斤/，/不论/大小/生熟/全部/装车/了/，/每斤/还/涨价/0.4/元/。/离开/西瓜/交易/地/之后/，/赵/某某/选择/报警/求助/。/接到/警情/之后/，/四十/里/堡派/出所/民警/立即/开展/工作/并/将/案情/通报/县局/扫黑/专业队/。/所对/联手/开展/案件/侦破/工作/，/经/大量/调查/取证/，/走访/群众/，/网上/网下/工作/，/掌握/大量/证据/。/7月/12日/，/民警/将/嫌疑/人/李/某某/等/三/人/抓获/。/经/讯问/，/犯罪/嫌疑人/李/某某/、/2/8岁/；/王/某军/、/3/1岁/、/王/某平/、/3/1岁/，/供述/了/在/网上/发布/虚假/信息/，/引诱/客户/到/来/之后/再/涨价/强迫/交易/的/犯罪/事实/，/采用/相同/方式/作案/多/起/。/

【家/人/急寻/！/兰陵/斜沟/一/学/生离/家出/走】/姓名/：/李/永强/年龄/：/15/岁芦/柞镇/斜沟/村人/，/兰陵县/第二/中学/上学/。/昨天/下午/六点/在/斜沟家/中离/家/出走/，/有/看到/他/的/同学/，/以及/见到/他/的/好心/人士/多加/留意/！/家长/万分/着/急/！/感恩/感谢/好心/人士/帮忙/有/知道者/请打/电话/：/15/09/286/62/84/

今天/太热/了/还是/找个/树阴/凉底/避暑/吧/再/去/工地/搬砖/的/话恐/怕/会热/的/干哕/

【/三/个/“/熊孩子/骑/一/电/动车/，/撞上/路边/大货/车】/7月/11日/15/时3/0分/，/在/临沂/兰陵县/火车站/以/南约/40/0米/处/路段/，/一/辆/两/轮/电动/车突然/撞上/路边/停放/的/一/辆/大/货车/，/电动车/重重/摔倒/。/兰陵/交警/大队/接到/报警/后/，/民警/迅速/赶到/事发/现场/，/伤者/已/送往/该县/人民/医院/。/骑电/动车/的/小/女孩/面部/严重/挫伤/，/另/一/乘车/小/女孩/的/腿部/受伤/，/正/在/接受/检查/治疗/。/还有/一个/乘车/女孩/有些/擦伤/。/经过/进一步/了解/得知/，/3个/小/女孩/同骑/一/辆/两/轮/电动/车外/出/玩耍/，/最/大/的/15/岁(/骑车/人)/，/最/小/的/只/有/7岁/，/在/返回家/中/的/路上/发生/交通/事故/。/3个/孩子/外出/时/，/其/家长/不/知道/，/只是/发现家/中/的/电动车/不/见/了/，/直到/听/说/孩子/出/了/事/。/民警/提示/，/家长/作为/孩子/的/第一/监护/人/，/要/时刻/关注/孩子/的/出行/安全/。/