运用微博数据构建相关联的新闻概要

欧一锋(@Souler Ou) 、郑齐(@Tidyzq)

目录

摘要

多年来，构建基于中长篇文本的内容的概要一直都是信息检索的研究热点问题，对于很多科学工作者，尤其是机器学习，数据挖掘，信息检索分析的工作者来说，构建概要通常费时费力，效果还不甚理想。本文基于 **TG-SUM:基于Tweet的多文档概要数据集构建** 一文，运用了*线性处理*、*Rouge评测* 等方法，运用微博数据，挖掘出跟相关新闻相关的微博，给构建新闻文本与其他类似文本的概要提供了一种效率较好且行之有效的方法。

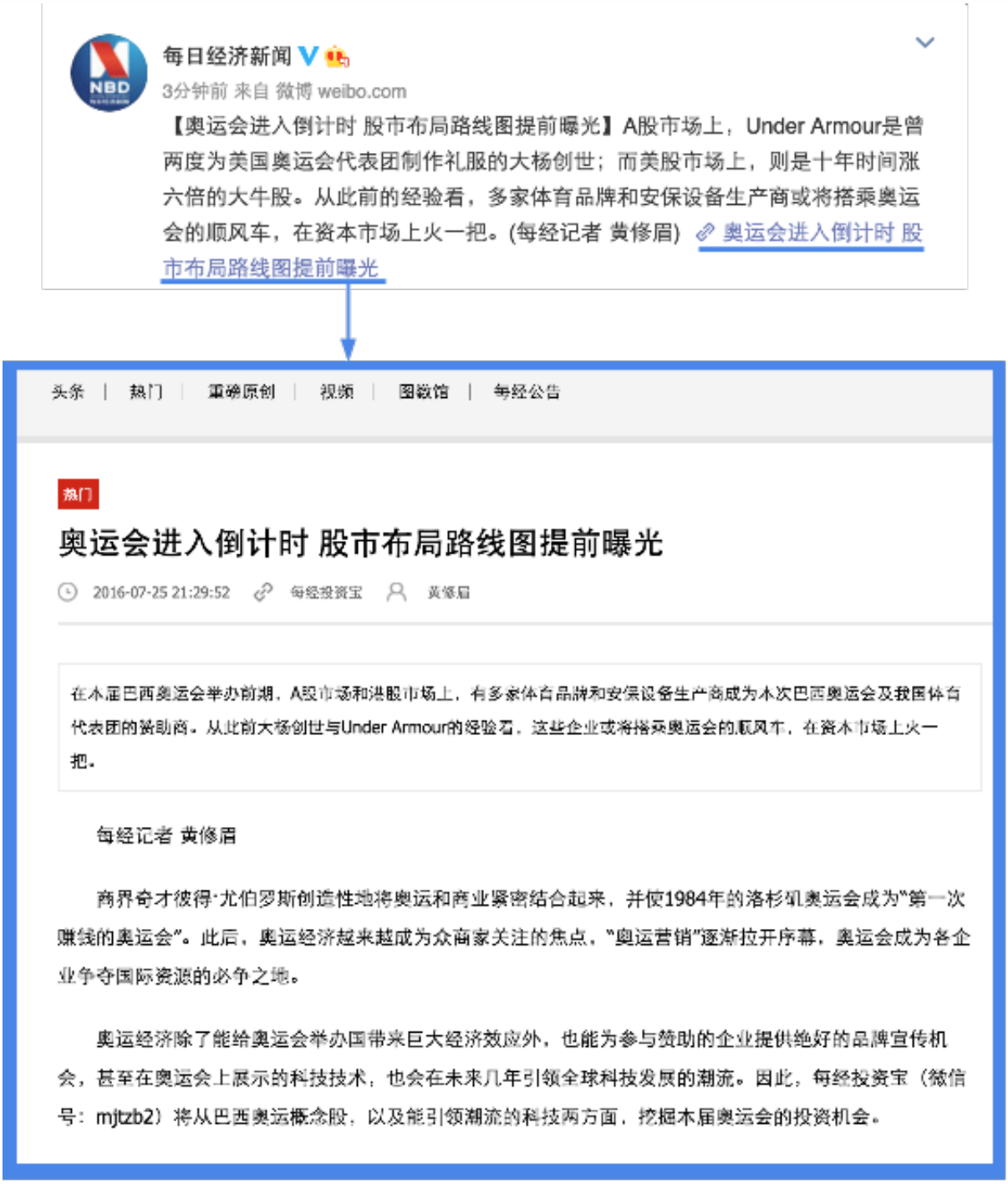
引言

众所周知，很多报纸或是其他的新闻出版物的编辑者所最头疼的一件事之一即是撰写新闻报道的概要，能否撰写出言简意赅的新闻摘要是评判一个新闻报纸是否能够被大众接受或者是否能够被更多的人所阅读的重要标准。我们可以看到，网上阅读源在互联网+时代数量显现爆炸性增长趋势，如果仍然选择使用人工撰写的方式写概要，显然十分不可取。我们知道，对于人工概要来说，不但 *极其耗时耗力* ，而且撰写出的内容 *不甚客观* 。所以自动概要生成，也成为了一个亟须解决的问题。

但是很不幸的是，*基于机器自动生成的概要* 一直以来表现都比较平平。一方面，这是由于用来提供给机器训练的人类编写的概要资料数量实在太少，造成了其生成概要的表现瓶颈。另一方面，基于机器学习的自动概要系统绝大多数情况下非常依赖于*特征选择*，这就导致了在比较多的情况下，非监督性学习反而会比监督性学习得到更好的表现。

从 **TG-SUM:基于Tweet的多文档概要数据集构建** 一文中，我们得到了一个比较行之有效的思路，那就是：通过参考社交媒体（对于中国大陆境内来说，主要是微博）的 *相关言论* ，构建出多文本概要参考集，进而通过对这些集合内的句子和词语进行线性规划寻找局部最优值而得出针对某个新闻的概要文本。在之前，有一系列的学者针对社交媒体上的社会化标注进行了基于NLP的研究。这些研究让我们在寻找局部最优值的时候能够提供很好的帮助。

> 在有关的社交帖子中具有如下性质：
>
> * 帖子中的绝大部分都可以指示与新闻相关的关键点（Key points）。
> * 以 "#" 或者 "@"开头的一系列标签会频繁出现在帖子中，会带来很多噪声
> * 由于长度限制，帖子基本上只描述了新闻中的一个特定的内容或者方面。



【图表示例新闻和链接微博的问题】

需要注意的是，对比新闻与微博，新闻中的句子通常写得较为简明且不容易产生歧义，但是微博中的句子可能会写得较为随意。所以我们决定不直接使用微博来构造我们的新闻摘要。 我们从 *文本和高质量相关微博* 中选取句子，尽可能多的囊括新闻中的关键点。在（Flippova和Alton提出的）所做成的模型中，选取了新闻的第一句话和帖子中的关键点用来构建新闻的概要，他们利用 *关键点* 删除了原始句子中一些不必要的成分。在 **TG-SUM:基于Tweet的多文档概要数据集构建** 一文中，引用了ROUGE矩阵，通过计算关键点的持有率来判断概要的优劣程度。同时使用了ILP（整数线性规划）方法来找到能达到ROUGE上界的句子集合。本文也继续沿用了这一方法。

新闻摘要的构建

我们主要把新闻摘要的构建分成四个步骤：数据收集，数据预处理，构建摘要，评估结果。

数据收集

数据收集直接使用python爬虫爬取特定新闻账号的原创微博，使用微博API即可自动爬取。爬取过程中只保留文本信息，不保留图片。

另外某些长微博（大于140词的微博）在爬取过程中会被自动截断并给予一个全文url。在爬取过程中，就可以使用正则判断微博是否被截断，然后根据全文url爬取全文。

我们选取了14个活跃新闻微博账号，爬取了其2016年5月至7月发布的全部原创微博，一共爬取了4379条微博。

爬取微博后我们根据微博中的url链接，爬取链接网页中的新闻信息。由于并非所有微博都有链接网页，而且某些新闻由于在爬取时已过期被删除，最终我们爬取到1489条新闻信息。

数据预处理

按照数据收集中的做法，我们把收集到的数据整理成两个csv文件，一个文件存放的是微博的内容，一个文件存放的是新闻内容。

预处理首先要解决的最大问题就是聚类问题。即如何给一个新闻文本匹配跟它说的最相似的内容的微博。要做的具体思路不外与信息检索关于文本聚类的问题的经典解决方法的思路类似。

解决聚类问题

分词问题

分词问题是首要问题。因为与英文不一样，中文的分词并不能简单的通过空格来进行划分，所以我们使用了 *基于NLP的分词器* 来协助我们进行中文分词。通过python的 *Jieba*中文分词器，我们能够将微博和新闻根据单词来分隔开。

词数向量的构建

此处由于词数向量比较简单，将文本切割开之后借助python的Counter即可完成词数向量（由于在使用，读取，分析的时候，运用向量的形式远不如词典类型的数据结构方便，所以我们此处词数向量的实现运用了词典类型的数据结构）

停词表的构建

由于在中文中，很多的特殊符号，介词，量词，程度副词，频率副词以及一些其他的虚词会对文本的分类造成很大的噪声干扰，使得原来并不匹配的两个文本匹配到一起。所以本处构建了中文的一个停词表，在切割后加入词数向量之前，依据停词表把常用的虚词（与文本的特征没有太大关系的词语）给筛选出去。

计算两文本间的相似度

传统的计算两文本间的相似度习惯于使用TF-IDF对数据进行正规化之后，使用 *欧几里得距离*、*余弦距离* 、*带权杰卡德距离* 计算两个词数向量之间的相似度。然而我们经过对数据的初步分析之后发现，该方法并不可取。原因在于，尽管使用了停词表，在去除了一大堆没有什么实际价值的虚词之后，新闻和微博的词数向量维度相差太大，在运行基于向量的距离计算时，由于维度相差太大，对微博词数向量升维处理的时候，除了造成很大的精度误差外，也可能会带来一系列的问题。所以我们构建了一个新的距离计算公式。使用该公式能够有效的将向量正规化，并解决由于维度差问题产生的最近邻居寻找失误。

寻找最近邻居

此处使用的方法与经典的文本分类中的最近邻居寻找方法相类似，不再赘述。

微博与新闻相似度计算

TF-IDF算法的优点是简单快速，结果比较符合实际情况。缺点是，单纯以"词频"衡量一个词的重要性，不够全面，有时重要的词可能出现次数并不多。而且，这种算法无法体现词的位置信息，出现位置靠前的词与出现位置靠后的词，都被视为重要性相同，这是不正确的。此处给出的是一种通过对次数向量中的不同关键词分别加权来解决处理新闻和微博的聚类问题的计算距离的一种可行的新方法。

具体分为如下几步：

步骤1、首先计算所有新闻向量的TF-IDF值。

步骤2、构建一个所有新闻的词向量，将各词的TF-IDF累加

$$\underset{wordi}{\rightarrow} = [0,1,1,0,0,\ldots,0,2] \Rightarrow \text{升维后的文本}i\text{的词向量的}TF\text{值}$$

$$\underset{whole}{\rightarrow} = \sum_{i}^{i<size(news)} \underset{word}{\rightarrow}$$

步骤3、将步骤二中获得的总词向量除以维度内最大项，获得正规化总词向量。

$$whole_r = \underset{whole}{\rightarrow} / max(\underset{whole}{\rightarrow}) \Rightarrow \text{正规化的总词向量}$$

步骤4、计算距离，使用 *新闻i* 和 *微博j* 的词向量求出交集：

$$inter_k = \underset{news_i}{\rightarrow} \bigcap \underset{weibo_j}{\rightarrow}$$

步骤5、通过对于交集中存在的元素，依据其在两亲（新闻i以及微博j）以及在总新闻的词向量中的的TF-IDF值，分别赋予不同的权值，并完成计算。

其中*whole_r*表示为正规化过的总词TF-IDF向量 , *whole*则为没正规化的向量。

tf_w 、*tf_n*则分别表示该条微博和该篇新闻的词数向量

$$weight_i = \begin{cases} whole[i] & , \quad (whole_r[i] < 0.1, tf_w[i] > 0 \quad and \quad tf_n[i] > 0) \\ whole_r[i] & , \quad (whole_r[i] > 0.1, tf_w[i] > 0 \quad and \quad tf_n[i] > 0) \\ 0 & , \quad (tf_n[i] <= 0) \\ 0 & , \quad else \end{cases}$$

$$distance = \sum_{i}^{inter_k} weight_i * tf_w[i]$$

此处距离越大，相似度越高。

CPLEX优化函数的编写

按照 **TG-SUM:基于Tweet的多文档概要数据集构建** 这篇参考论文来说，具体实现的函数主要有下面几个：

二词词数向量的构建

将文章的分词前后分别组合，获得n-1二词向量即可。

分句函数的实现

直接依据"。"、"？"、"！"、"....."四个符号对原始文本分割即可。

常数的计算

直接对词数向量求和取倒数即可。

构建摘要

构建数学模型

论文中给出了用于生成摘要的线性规划模型：

$$
\begin{aligned}
max \quad & \sum_{i=1}^{K}\left(\frac{\sum_b min\{n_{weibo_i}(b), n_{cnd}(b)\}}{\sum_b n_{weibo_i}(b)}\right) \\
s.t. \quad & n_{cnd}(b) = \sum_s z(s) \times n_s(b) \\
& \sum_s z(s) \times |s| \leq L \\
& z(s) \in \{0,1\}
\end{aligned}
$$

$n_{weibo_i}(b)$	第i条微博的$\mathbf{bi\text{-}gram}$系数
$n_{cnd}(b)$	正在生成的摘要的$\mathbf{bi\text{-}gram}$系数
$n_s(b)$	第s条候选句子的$\mathbf{bi\text{-}gram}$系数
$z(s)$	表示第s条候选句子是否选为摘要句子
L	构建摘要的最大长度

由于在微博内容确定的情况下，微博的bi-gram系数可以作为常数看待，因此可以设：

$$
w_{b,weibo_i} = \frac{1}{\sum_b n_{weibo_i}(b)}
$$

然后注意此式子：

$$
Gain_i(b) = min\{n_{weibo_i}(b), n_{cnd}(b)\}
$$

此式子中存在一个逻辑约束min，即取最小值。但是在CPLEX中，无法使用逻辑约束，因此，我们将此式子拆分成以下约束：

$$
\begin{aligned}
Gain_i(b) \leq n_{weibo_i}(b), \forall b, i \\
Gain_i(b) \leq n_{cnd}(b), \forall b, i
\end{aligned}
$$

由于我们的线性规划目标是取到式子的最大值，可以看出，当且仅当Gain为两数的最小值时，模型取到最大值。

将以上式子结合后我们就得到了最终的线性规划模型：

$$
\begin{aligned}
max \quad & \sum_{i=1}^{K}(w_{b,weibo_i}\sum_b Gain_i(b)) \\
s.t. \quad & n_{cnd}(b) = \sum_s z(s) \times n_s(b) \\
& \sum_s z(s) \times |s| \leq L \\
& z(s) \in \{0,1\} \\
& Gain_i(b) \leq n_{weibo_i}(b), \forall b, i \\
& Gain_i(b) \leq n_{cnd}(b), \forall b, i
\end{aligned}
$$

求解该模型的目标，即在约束内构建一个z向量，使得模型中的max函数取得最大值。

构建CPLEX模型

变量

首先，z是我们需要构建的向量，因此z需要作为变量给予CPLEX，让其对模型进行求解得到结果。

由于z的取值为0或1，表示候选句子是否被选取，因此在CPLEX中，将z定义为binary类型，即布尔类型。

其次，由于Gain的取值没有线性公式，依赖于上界约束，Gain也需要作为变量，根据模型的求解过程而改变。

通过计算我们可以得出，z是一维向量，长度为s（候选句子数）。Gain是二维矩阵，大小为 i*b （i为链接微博数，b为总bi-gram数目）。

因此我们一共需要 s + i*b 个变量。

目标

直接将此式子设定为最大化目标即可：

$$
\sum_{i=1}^{K}(w_{b,weibo_i}\sum_b Gain_i(b))
$$

约束

首先，对于约束：

$$
Gain_i(b) \leq n_{weibo_i}(b)
$$

等式的右边可以看做常数，因此我们可以将此约束作为Gain变量的upper_bound约束加入CPLEX中而不占用线性约束。

其次，对于以下约束：

$$
\begin{aligned}
& n_{cnd}(b) = \sum_s z(s) \times n_s(b) \\
& Gain_i(b) \leq n_{cnd}(b), \forall b, i
\end{aligned}
$$

可以合并为一个约束，同时将变量全部移至式子的左边：

$$
Gain_i(b) - \sum_s z(s) \times n_s(b) \leq 0
$$

此式子容易用线性约束表示。此式子需要对每个Gain变量设定约束，因此一共需要使用 i*b 个约束。

最后，将此约束加入CPLEX：

$$\sum_{s} z(s) \times |s| \leq L$$

约束全部添加完毕，一共使用 i*b + 1个线性约束。

评估结果

以下为部分结果

新闻原文	生成摘要
从昨天（3日）开始，在上海，有一群人冒着雨，在上海影院门口等待着什么，就像这样：到了今早（4日）7：30分左右，这支队伍变得更长了：原来，第19届上海国际电影节于今天上午8时正式开票，他们都是前来买票的观众。排在队伍最前的是孙小姐，她从昨天下午2点就开始等待，她告诉东方网记者：今年最想看的是日本影片：《和母亲一起生活》、《暗杀教室》等，为了这些难得一见的电影，排18个小时的队也是值得的。还有一位70岁高龄的“铁杆影迷”方先生，他从第一届上海电影节开始每年都会第一时间来现场买票。今天，方先生赶乘地铁首班车早早到达现场，他告诉东方网记者：自己这几天一直都在作“功课”，列出了一份长长的电影节心愿片单，今年最想看的是好莱坞影片《乔布斯》。在网络上，另一群人也是早早定好8：00分的闹钟，打算通过淘票票APP来订票。每经小编（微信号：nbdnews）了解到，6月11日到19日，第19届上海国际电影节将举行。日前公布了此次展映的完整片单及排片表，这次参加展映的影片有近600部。不仅有在大银幕上极为罕见的大师之作、今年戛纳电影节的入围电影，还有年轻人喜欢的日韩片，《哈利·波特》系列八部连映、莎翁影展、迪斯尼·皮克斯电影周、007回顾等，被影迷称为“上影节史上最强片单”。其中，电影节推出的“安德烈·塔可夫斯基回顾展”非常重磅，《伊万的童年》等多部名作都将在电影节公映。这 些片子在上海的放映场次将超过1250场，这是部分将会在电影节期间放映的影片海	第19届上海国际电影节开幕。第19届上海国际电影节11日开幕，成龙、刘烨、黄晓明夫妇、杨洋、舒淇、宋茜、李敏镐等300多位中外艺人出

当天于在上海的放映场次将超过1250场。这是部分将会在电影节期间放映的影片海报：此时的上海影城，已经接近9点，百米长队迟迟得不到动弹。	大出席红毯仪式。在电影节主竞赛单元金爵奖评选中，共有2403部影片角逐个奖项。
新华社无锡6月4日体育专电(记者 王镜宇 王恒志)国家体育总局棋牌运动管理中心党委书记、国际围棋联盟事务总长杨俊安4日在这里透露，如果不出意外柯洁九段将在年内进行和“阿尔法狗”的围棋“终极人机大战”。在4日下午举行的第37届世界业余围棋锦标赛新闻发布会上，杨俊安透露了这一消息。据他介绍，中国围棋协会和“阿尔法狗”的团队就此事进行了接触和沟通，双方都有意向促成这项对抗。如果不出意外的话，这次比赛将安排在年内，但是具体时间和比赛地点等还“无从谈起”。今年3月进行的“阿尔法狗”和李世石的围棋人机大战引起了全世界的广泛关注。来自中国、韩国、欧洲和美国的围棋官员均表示，这次对抗极大提升了围棋在当地的关注度。在此间举行的国际围棋联盟全体代表大会上，还有人提议向“阿尔法狗”颁发“围棋推广特别贡献奖“。在那场举世瞩目的人机大战中，“阿尔法狗”以4：1战胜了韩国名将李世石九段。不过，中国等级分排名第一的柯洁九段当时就表示，虽然“阿尔法狗”战胜了李世石，但它赢不了自己。因此，有不少棋迷也期待看到柯洁和“阿尔法狗”的对决。据刚刚卸任的国际围棋联盟事务局长、韩国棋手李夏辰介绍，李世石和“阿尔法狗”的人机大战为围棋在韩国所赢得的关注是空前的。当时，包括KBS等重量级电视台在内的9家电视媒体对比赛进行了转播，收视率接近男足世界杯，李世石也成为国家英雄一样的人物	柯洁年内将战“阿尔法狗”，你看好谁?据新华社，国际围棋联盟事务总长杨俊安昨天透露，如果不出意外柯洁九段将在年内

转播，收视率接近男足世界杯，李世石也成为国家英雄一样的人物。	年内进行和“阿尔法狗”的围棋“终极人机大战”。你看好谁?
毛坦厂是安徽的一座僻静小镇，周围是沟壑丛生的山峦。可这座小镇却拥有一座大名鼎鼎的中学，它被称作“亚洲最大高考工厂”，拥有超高的升学率，每年都有很多家长慕名而来，将孩子送进这里学习，这里就是毛坦厂中学。每年高考前，这所学校都将会为高三考生举办出征仪式，近万名考生乘坐大巴浩浩荡荡地驶出校门，好不壮观。今年的出征仪式，新浪新闻将与您共同见证！人群已经散去，部分家长在路边合影留念。一年一度的万人送考大会，也缓缓落下帷幕。祝考生们在高考中发挥出自己应有的水平，因为考场外还有人在默默地为他们祈祷和祝愿。直播结束。谢谢大家的收看。左边这位家长的女儿今年就要参加高考。她从6年前就随孩子过来陪读，那时孩子还在读初中。当被问到孩子想考什么学校时，阿姨笑容满面：“当然是好学校呀。”一个考生，牵动的是一个家庭的心脏；一场考试，书写的是一个家庭的未来。愿考生们今年都能顺利发挥！大巴已经全部开出。现在跟在后面的，是一些送考的私家车。这所中学成为整个毛坦厂镇跳动的心脏。每个寒暑假，在没有学生的日子里，这个小镇安静的吓人，商店歇业，居民盖起的三层大楼大门紧闭空空荡荡。而一旦开学，这里便又是一片沸腾。考生们一手拿手机拍照，一手向送考家长们挥手。镇上没有KTV、网吧等容易让学生分心的娱乐场所，据说曾经的一家网吧被家长们抵制，赶走了。第16辆大巴开出，速度已经快出了许多。可以近距离地看到车上的考生。有位女生笑容灿烂，向车外的人群挥手致意，似乎看不出对即将到来的“独木桥”有丝毫紧张。因为聚集的送考家长很多，现场有保安在维持秩序。车上有位考生露出了“迷之微笑”，是否是因为胸有成竹呢？愿你考试顺利。许多家长专程从外地赶来。陪读的日子也是家长们的煎熬，每到高考时，都有一大波陪读家长庆幸终于熬到头了。但也有家长在孩子考上大学后依然留在这里做一些简单的工作维持生计。数据显示，2015年高考中，毛坦厂	毛坦厂中学考生出征高考 场面壮观。这座小镇因一所被称作“亚洲最大高考工厂”的中学闻名。每年高考前，毛坦厂中学都会为高三考生举办出征仪式

大学后依然留在这里做一些简单的工作维持生计。数据显示，2015年高考中，毛坦厂中学参考人数13000人，其中达一本分数线3106人，二本人数4896人，本科达线近11000人。应届一本达线率为41.01%，应届本科达线率为85.94%。每辆大巴车的副驾驶都坐了一位学校的老师，会向家长挥手致意。第五辆大巴车驶出，速度已经比之前几辆快了很多。以前有讲究送考车的头车司机要姓马或者属马，寓意马到成功。但在近两年，似乎并没有这么严苛。第四辆大巴开出。今年出征的大巴车比往年少了很多，据毛坦厂中学工作人员称，以前送考车最多达到70多辆大巴车，今年很多外地考生都坐着私家车先走了。

式。今年毛坦厂中学用30辆大巴运送考生，头车尾号666，司机属马，寓意马道成功。

结论

通过生成的摘要，我们可以看到自动生成的摘要对新闻原文的信息有着良好的覆盖。通过微博内容生成的摘要，即能够表达出大部分原文中的信息，又能去除原文中不必要的繁琐细节。而且此技术可以在数据量很少的情况下正常运作，在数据量大的情况下也能轻松应对。

与传统的基于机器学习的自动摘要生成方法相比，通过微博生成摘要的技术能够在很大程度上代替旧方法方法。相比以往的方法，此技术避免了机器学习都有的训练困难，调参困难等问题。该技术更加简便，同时有着更大的潜力与发展前景。