WG-SUM: An automatic news summary generator alternative

Yifeng Ou(@Souler Ou) 、Qi Zheng(@Tidyzq)

Directory

Abstract

Create content summary for medium or lengthy text is a hotspot in information retrieval field in recent years.As for many science researchers - especially for those who majored in Machine Learning , Data Mining and Information Retrieval Analysis , Creating Summary is always a toughing job for them while the work seldom easy yet always time-consuming and energy-costing.What's the most terrible , such tiring work will never pays back equivalently.This article , which based on **TG-SUM:Building Tweet-based Multidocument summary dataset** , using Integer Linear Programming(ILP) , ROUGE Analysis Metrices and some other methods ,with the help of the weibo post database , scraping out the weibo posts which strongly relative to the

given news dataset and finally provide an effective way for automatic generating the text summary for news or other news-like articles.

Introduction

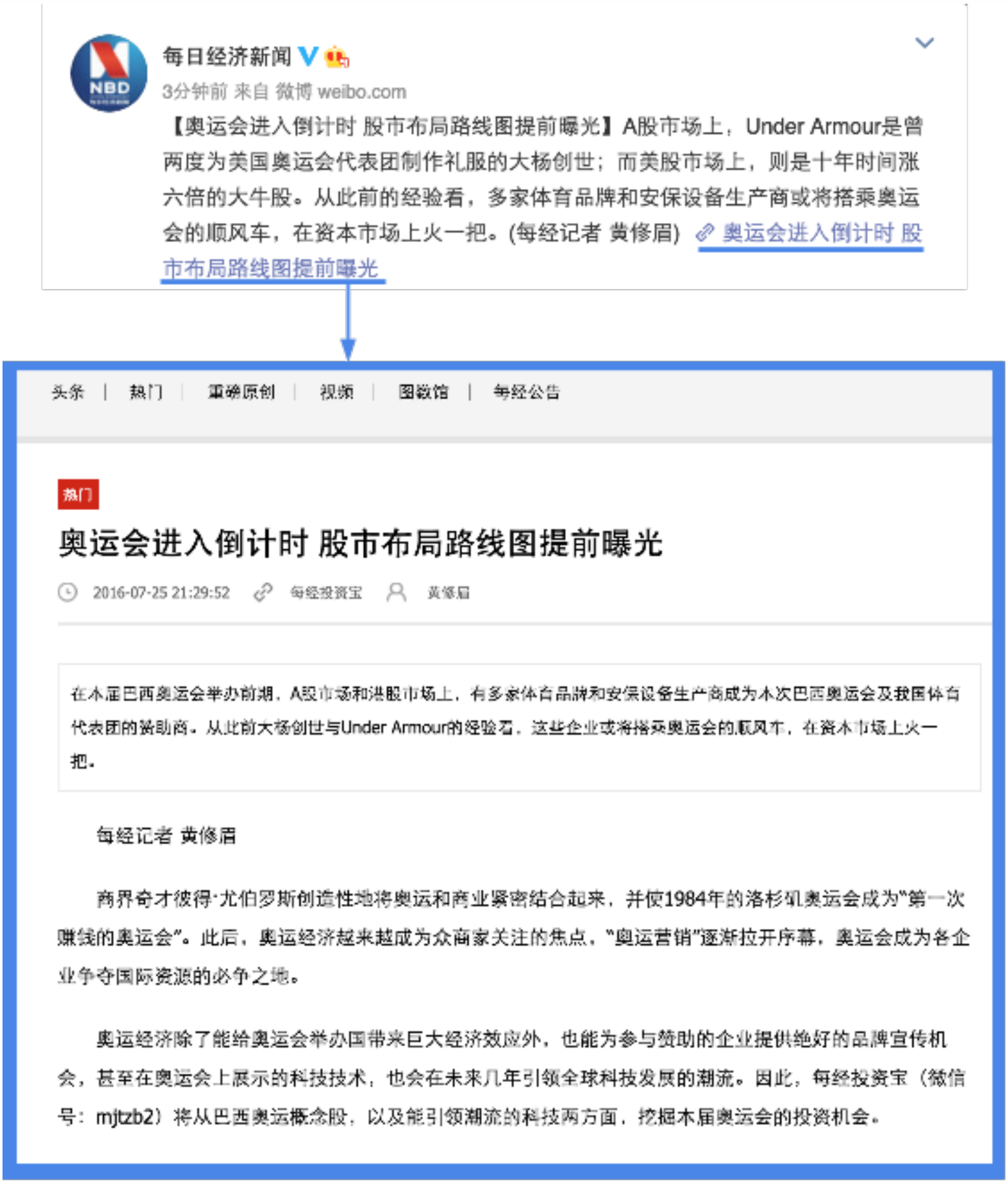
As is known to all, the biggest headache for the editor of the newspaper or other news press products is to write the summaries for the news reports. Since whether a news summary is brief yet clear is a boost to the acceptance and the read probability of the public to a certain related news , acquiring the capability of writing a nice summary is a must to those editors.As alongside with time went on , the RSS reading resources, as well as the e-based news are facing a remarkably boom in the upcoming brand-new Internet+ era. For that we all knows , human-writed edition summary not only *time-consuming and energy-costing* . but also *subjective*.Therefore , using human-handwrited edition of news summary seems not suitable for the new trend any more. As a result , auto-generator for summaries is becoming more and more necessary with coming demands.

However it's very unfortunately that *auto-generated summary created by machine* have never show a performance better than none.As with the former researchers' research results , on the one hand , this is because that the human-handwritten summary are in a poor size to provide the training dataset for machine learning training. On the other hand , auto-summary-generators which based on machine learning is relatively shows a strong dependency on the feature selecting , which leads the unsupervised learning have a better result than supervised learning in many occasions.

As given in the **TG-SUM:Building Tweet-based Multidocument summary dataset** , we got a effective alternative for summarizers — Reference according to the *relative posts* on some influential social medias(Majorly is Weibo in China mainland) , create the multi document reference set , and then perform Integer Linear Programming methods on the sentences and words(usually bi-grams) in these sets and find out its optimal value,we will then get the summary text for a specified news.As for the former studies , many researchers have launched several NLP-based research on the social tags on the social medias, which is certainly of great help for us in the process of finding the local optimal.

As refer to the former researches:

```
> These are the useful findings according to the relations between the social media
posts and the news:
>
> * The majority of the social media posts can provide the *Key Points* to a
relative news report.
> * A lot of tags started by "#" and "@" will frequently appeared in the social
media post , which brings a lot of noises.
> * Posts often contains the summary of only one aspect or some specific content of
the news due to the the system size restriction of words(in weibo its 140 Chinese
Characters most).
```



每日经济新闻

3分钟前 来自 微博 weibo.com

【奥运会进入倒计时 股市布局路线图提前曝光】A股市场上，Under Armour是曾两度为美国奥运会代表团制作礼服的大杨创世；而美股市场上，则是十年时间涨六倍的大牛股。从此前的经验看，多家体育品牌和安保设备生产商或将搭乘奥运会的顺风车，在资本市场上火一把。(每经记者 黄修眉) 奥运会进入倒计时 股市布局路线图提前曝光

奥运会进入倒计时 股市布局路线图提前曝光

2016-07-25 21:29:52 每经投资宝 黄修眉

在本届巴西奥运会举办前期，A股市场和港股市场上，有多家体育品牌和安保设备生产商成为本次巴西奥运会及我国体育代表团的赞助商。从此前大杨创世与Under Armour的经验看，这些企业或将搭乘奥运会的顺风车，在资本市场上火一把。

每经记者 黄修眉

商界奇才彼得·尤伯罗斯创造性地将奥运和商业紧密结合起来，并使1984年的洛杉矶奥运会成为"第一次赚钱的奥运会"。此后，奥运经济越来越成为众商家关注的焦点，"奥运营销"逐渐拉开序幕，奥运会成为各企业争夺国际资源的必争之地。

奥运经济除了能给奥运会举办国带来巨大经济效应外，也能为参与赞助的企业提供绝好的品牌宣传机会，甚至在奥运会上展示的科技技术，也会在未来几年引领全球科技发展的潮流。因此，每经投资宝（微信号：mjtzb2）将从巴西奥运概念股，以及能引领潮流的科技两方面，挖掘本届奥运会的投资机会。

【A chart showing the news and linked weibos】

What should be focus on is that while comparing news and strongly-relative weibos, we will find out that the sentences in the news are more well-written than those in weibos.Based on that result , we chose the sentences both in the news and the strongly-relative weibos rather than from weibo directly , and managed to cover the news key points as much as possible.An alternative (Provided by Flippova and Alton) model uses the first sentence of the news and the key points from the social media posts to generate the summary, they uses the *key points* to remove the unnecessary factors which exist in the raw sentence.Another alternative , provided in **TG-SUM:Building Tweet-based Multidocument summary dataset** , is that using *ROUGE Analysis Metrices* to calculate the *Coverage Ratio* of the *Key points* to judge the performance of the summary.What's more , they uses the *Integer Linear Programming* method to find the *sentence set that reaches the upper bound of ROUGE* , which used as the output summary.However , we regard this a nice alternative for building a summary based on social media posts , therefore we will use this alternative in this article.

Constructing News Summary

In our construction , we majorly divide the construction process into 4 parts:

- Data gathering
- Data preprocessing
- Summary generating
- Result evaluating

Data Gathering

To gather weibos and news, we simply used a python-based web crawler, the crawler uses Weibo API so it can automatically capture original weibos posted by certain users. While gathering, we only capture the text of content ignoring pictures.

Some long weibos (longer than 140 words) will be cut by API and give a url for full text. We can use a regex expression to recognize whether the weibo has been cut and capture it's full text according to the giving url.

In total, we selected 14 active news accounts and capture their original weibos between May to July in 2016, we collected 4379 weibos altogether.

After collecting weibos, we captured news text under the linked url in the content of weibo. But not all weibos have linked url, and some news was deleted before we can capture, we only collected 1489 articles of news.

Data Preprocessing

According to the storaging organization in *Data Gathering* step , we have got 2 csv files for separately storing weibo and news content.

The first and foremost problem that preprocessing faced is the *Clustering Problem*.We have to assign the news text its *strongly-related* weibos.To complete this , we refer to the classic solution to the text clustering problem in the Information Retrieval field and make a little modification on it in order to suit the problem here.

Solving Clustering problem

Chinese Word Segmentation problem

It's the first problem in clustering. Since Chinese paragraph and sentence structure is a lot different from English , we couldn't simply do the word segmentation job by simply using the space. Thankfully for the formers' job , we discovered a *Chinese Word Segmentation Generator* — **Jieba** , a Chinese segment generator *based on NLP* . With the help of this tool , we have generated news and weibo segments successfully.

Building word-count vectors

We simply use the Counter in Collection pack in python to help us solve the word count calculations. What's beyond that is we choose the Data structure of *Python Dictionary* to act as the role of word-count vector as long as it's more convenient to use in reading , storaging and analyzing.

Building stop-words list

A bundle of special characters in Chinese , including emojis , prepositions , quantifiers , relative and comparative adverbs , frequency adverbs and some other function words will cause a lot of noise to the clustering process and leads to , awfully , the machine put 2 unrelatively text together in the nearest neighbor of the specific news.Therefore , here we modified a stop-word list to throw away the useless function words (not so-relative to the text features) to provide better clustering results.

Calculating the similarity between 2 texts

Traditionally , we are used to compute the similarity with *Euclidean distance* , *Cosine distance* , or *Weighted Jaccard distance* based on the *TF-IDF regularized* word-count vectors.However , we find this classical method is not suitable here anymore after we analayzed the data.The problem is that the *dimension difference* for the news' word-count vector and weibos' word-count vector is extremely large although we have already removed the not-so-really-useful function words.Base on the assumption , we will face the tremendous precision error and many other problems when calculating the word-vec difference distance on vector calculation because we need to raise the dimension of the low-dimension one.Therefore we here construct a new formula for calculating distances. We can regularize the vector distance effectively by reassgining weights and eventually solve the faults on finding nearest neighbors.

Finding nearest neighbors for news

Here we simply use the classical methods , therefore this part is omitted for a more brief article.

Calculating Similarity between news and weibo

The strengths of TF-IDF algorithm is that it's *Easy* , *Speedy* , *Efficient*. Its weakness , however in our case more remarkable , is that it's not comprehensive enough while only judge a word's importance by *term frequencies* for that the important words (So called as *key points*) usually don't appears much.Here we discovered a new method for handle this problem by placing different weights on different words.

The step are as follows：

- Step 1：Calculate all weibos' and all news' TF-IDF word-count vectors.
- Step 2：Construct a *big word-count vec* on all news report documents, and sum each term by each news TF-IDF word-count vectors.

  (tf_n[i] stands for news i 's TF-IDF word-count vector , news-whole stands for the big word-count vector for all news documents)

$$\underset{tf_n[i]}{\rightarrow} = [tf - idf_k]$$

$$\overrightarrow{}_{news-whole} = \sum_{i}^{i<size(news)} \overrightarrow{}_{tf_n[i]}$$

- Step 3：Divide the big vector by the *largest term among all dimensions* , and we gained the whole_r , the regularized big vector.

$$\overrightarrow{}_{news-whole_r} = \overrightarrow{}_{news-whole} / max(\overrightarrow{}_{news-whole})$$

- Step 4：Compute the intersect set on *news i* and *weibo j*

$$inter_s = \overrightarrow{}_{news_i} \bigcap \overrightarrow{}_{weibo_j}$$

- Step 5：Give each term exists in the intersection different weights by different TF-IDFs in the given news , given weibo and big-vector on news.Sum up the weight * tf_value and we finally got it.

$$weight_k = \begin{cases} news-whole[k] & , \quad (news-whole_r[k] < 0.1, tf_w[k] > 0, tf_n[k] > 0) \\ news-whole_r[k] & , \quad (news-whole_r[k] > 0.1, tf_w[k] > 0, tf_n[k] > 0) \\ 0 & , \quad (tf_n[k] <= 0) \\ 0 & , \quad else \end{cases}$$

$$distance = \sum_{k}^{inter_s} weight_k * tf_w[k]$$

In our definition , the similarity is larger the better.

Coding on functions for CPLEX

According to the **TG-SUM:Building Tweet-based Multidocument summary dataset** , we need to completed several function for providing accurate values in running CPLEX optimizer.

Building bi-gram-count vectors

Simply combine with the next word. Here we not-remove any function word at all and again use simple count vector instead of the TF-IDF versions.

Get sentence vectors

Here we only split original raw document by this four Chinese punctuations： "。"、"? "、"! "、"......".

Calculate const terms

We simply sum up bi-gram count vector on its rows and get reciprocal of the result.

Construct the Summery

Construct mathematical model

The article **TG-SUM:Building Tweet-based Multidocument summary dataset** already gives the ILP modal of the problem:

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

$n_{weibo_i}(b)$	the bi-gram value of the i th weibo
$n_{cnd}(b)$	the bi-gram value of the candidate summary
$n_s(b)$	the bi-gram value of the s th sentence
$z(s)$	whether the s th sentence has been selected
L	maximum length of summary

For a certain weibo, we can consider it's bi-gram value as a constant, so we have:

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

Then we notice:

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

There is a logical constraint **min** in this formula, but we can't use any logical constraints in *CPLEX*, thus we rewrite this formula in another form:

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

Because our objective of the *ILP* problem is to reach the maximum value, it's easy to see that **Gain** will reach the minimum of the two values.

Combine the formulas above, we get our final *ILP* modal:

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

The goal of the modal is to construct a vector **z** which indicate the sentence of summary to maximum the formula with **max** label .

Construct CPLEX model

Variables

First, **z** is the vector we want to construct, so **z** must be a variable to *CPLEX*, so *CPLEX* will give us the solution.

Because **z** indicate whether the candidate sentence is selected or not, we need to define **z** as *binary* type for *CPLEX*, which is boolean type.

Second, **Gain** has only upper bound constraint rather than linear formula to it's value, **Gain** also need to be set as variable.

By calculation, we can get that **z** is an one-dimensional vector with the length of **s** (number of candidate sentences). **Gain** is a two-dimensional matrix with the size of **i * b** (**i** is the number of linked weibos, **b** is the total number of *bi-gram* values).

In total, we need **s + i * b** variables altogether.

Objective

The objective is to maximum this formula:

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

Constraints

First, for the given constraint:

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

The right side of the formula can be regard as constant, so we can simply treat this constraint as *upper bound constraint* instead of *linear constraint*.

Second, for these constraint:

$$n_{cnd}(b) = \sum_{s} z(s) \times n_s(b)$$
$$Gain_i(b) \le n_{cnd}(b), \forall b, i$$

We can merge these two formula as one constraint and move all the variables to the left side:

$$Gain_i(b) - \sum_{s} z(s) \times n_s(b) \le 0$$

This formula is easy to be represent using *linear constraint*. But it requires one constraint for a single **Gain**, so we need **i * b** *linear constraint*.

Finally, add the last constraint into *CPLEX*:

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

Now we have **i * b + 1** *linear constraint* in total.

Evaluate the result

There are some summary result below:

Original News	Generated Summary
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。	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	

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。	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。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。

Conclusion

Through the summary generated, we can conclude that summary based on weibo not only contain the majority of information in original news text, but also ignore the unnecessary detail or useless message. This new method can also operate in a few cases of data as well as huge amount of information.

Comparing to the Machine Learning based way of generating news summary, this technique avoid the common problem of Machine Learning such as training and parameter adjusting. By contrast, our new method is far more concise with a greater potential and prospects.