摘要：

本文利用MUcut算法，进行特朗普推特的文本主题分析，分类得出其主题和对应的tweet,。可以看出

本文提出了MuNuct算法，以特朗普的tweet为载体，分析互联网媒体短文本，以特朗普的tweet为例，进行语义分析（归纳出与主题对应的tweet）和主题分析（归纳出主题），并论证了其可行性。

稳定性

此方法对除推特以外数据集的应用结果

背景：

自然语言处理的发展历程

自然语言处理有基于规则的和基于概率统计的模型，基于规则的模型就是根据现有的语法规则通过对单词的组合进行语义分析，通过规则的分析方法，建立符号处理系统，目前的方法有基于规则库、词典标准、推到算法设计等，具体实现方法有、形式语言、语法理论、词法理论、推理方法等，在历史上此模型最先出现，但由于随着这种模型的发展，语法规则库逐渐庞大，其效率到了发展的瓶颈期，至今为止，完善的规则系统还没有开始建立。而随着大数据时代的出现和相关技术如统计数据、存储于检索的发展，为基于概率统计模型的自然语言处理技术提供了巨大的样本和技术保障，所以现在自然语言的统计处理是当前最流行的处理得方式。但是基于统计的生成和处理规则只保留了语言文本的数量分布特征，丧失了其空间分布特征，所以还是存在效率低或者丧失原始文本信息的缺陷。因此，考虑基于规则和基于统计的模型的优缺点，我们从文本内容空间分布的特征方面出发，提出了新的基于文本部分空间分布特征的模型。

并且现有基于统计的自然语言处理模型，主题分析，语义分析都依靠概率统计模型，需要极大的样本和长文本。而对tweet等短文本或者样本数较少的研究对象不合适。

但是到目前为止概率统计模型还是具有不可比拟的优点，因此我们也采用了一些常用模型来帮助建立我们的模型和验证我们的模型。

我们模型主要是通过聚类来完成

我们的论文为研究自然语言处理提供了一个新的方向，虽然存在一定的问题，但以后肯定会日趋完善。

文本处理：

编码转换

我们主要针对的分析互联网短文本内容分析，如我们所用的tweet文本主要来自于网页，所以不可避免的就会带有表情等特殊符号，编码格式主要为utf-8编码，此外特朗普的推特之中难免还有一些用英文中文发表的内容，由于我们建模的对象为英文单词和文本，因此这些表情符号等首先需要过滤掉，我们通过过滤和替换非正常字符编码范围字符的方式进行内容的初步清洗。

正则表达式处理：

文本内容中有I’m you’re clintun’s fake news 等简写或者谁的东西等特殊形式，还有一些连接如http等形式的内容，我们采用正则表达式的方式进行查找替换，如下所示：

[.@或者@删除一整行](mailto:.@或者@删除一整行)

'(\.|")@.\*'

删除掉@的人@[A-Za-z-.']+

@[a-zA-Z0-9.-\_']+

RT删除一整行

删除#hashtag

删除&amp

删除http等

删除除了字母，空格以外的其他字符

找到例如I，he等代词

找到带有’s的单词

还原 带有not 的缩写形式

还原 would的缩写形式

还原will的缩写形式

还原 am的缩写形式

还原 are的缩写形式

还原have的缩写形式

然后进行替换，经过此项步骤，文本中就只剩下纯净的单词形式

词性还原：

文本内容中会存在一些如动词的三单形式和名词的复数形式或者形容词副词的最高级和比较级等，这些单词具有相同的词根，如like，likes，job，jobs等，因此我们需要进行词干提取和词形还原。

在本片论文的处理过程中，我们利用了斯坦福大学开发的NLTk进行词干提取和词形还原。

去除停用词：

在英文语法中有许多系动词，情态动词，助动词等对主题无关的词，因此需要去除

转化为小写：

由于大部分编程语言是区分大小写的，并且大小写不同一般不会影响词的意思，所以我们把全部文本转化为小写

单词推特网络的：

我们主要是为了弥补基于统计模型的自然语言处理得不足，刻画文本的空间分布特征并从空间分布特征的角度进行文本分析，因此我们采用复杂网络的形式，即建立一种分层的网络，第一层为单词网络，第二层为tweet网络，两层网络之间相互连接，既可以刻画单词的空间分布信息，又可以刻画tweet的空间分布信息，还可以刻画单词和推特的空间分布信息，有效的将文本与单词结合起来，这种方法适用于像tweet这中的互联网短文媒体的内容分析。下面我们来介绍一下单词网络和tweet网络以及如何的构建方法。

单词网络建立：

由于我们针对的是短文本内容的主题分析，每条文本中最多含有100多个单词，而概率统计模型最基本的就是基于条件概率分布，如：，这种短文本用基于概率统计的分布模型误差是非常大的，而从每个单词相对出现位置的角度去考虑，就可以避免掉因为短文本带来的问题。

建立方法：

我们把所有tweet中出现的单词构建一个网络，每个单词代表一个节点，当两个单词出现在同一tweet中时则连接两个对应的节点（以后还以当出现在两篇tweet中时才建立推特网络）。

采用此法建立的意义：

之所以采用这种方式是因为考虑到单词之间的出现的相关性，如特朗普tweet中很多make America great again等tweet，那么这几个单词出现在一起额概率就非常大，实际上的理解也是如此，说到中国，就会联系到北京，说到好莱坞就会联系到电影，当特朗普谈到希拉里克林顿一般是bad 等形容词。

未尽：

词袋方式和词向量方式

Tweet网络建立：

建立方法：

我们把所有搜集到的tweet建立一个网络，每条tweet作为一个节点，若两个tweet之间有相同的单词（大于等于一个单词，以后还可以有大于等于两个单词等），则连接对应的两个节点

如此建立的意义：

建立的Tweet网络的直观反映是为了体现每条tweet之间的关系，而两条tweet是否相关，仅从单词层面上的直接想法便是如果两条tweet之间存在相同的单词，则两条tweet所表达主题之间便可能存在着一定的关联。

未尽：

两条Tweet之间有相同单词的数量，有相同单词且位置相对相同的数量，有没有否定双重否定的单词等

单词tweet网络的建立：

建立方法：

在上述建立的两个网络中，如果某个单词存在与某条tweet中，则将单词网络和tweet网络中对应的节点想连接

意义：

经过上述两个部分我们分别建立起两个单层的网络，分别是单词网络和tweet网络。而实际上这两个网络是存在着一定的关系的，即单词网络中的单词都来自于tweet网络中的一条或者多条tweet，每个单词有自己独特的含义，含有若干个单词的tweet也有自己独特的含义，这两个网络中单词包含在tweet中，tweet中包含着若干单词。

MuNcut算法：

算法：

层内算法：

首先考虑单词网络，构建的单词网络以表示，若处为则代表第个单词和第个单词有连接。把其作为相似矩阵，代表权重矩阵，表达了两个单词之间的相关性，下面我们介绍一下算法：



其中：



代表类内距离。



代表类间距离。

因此理论上最好的分类结果就是类内距离越小越好，类间距离越大越好，值越小越好。因此，得到最优聚类等同于找到的最小值

同上，我们同样可以定义tweet网络的计算方法。两个网络有自己的权重矩阵，我们分别定义为，总的来说，单层聚类的计算方式为:



其中代表单词层的聚类值，代表tweet层的聚类值

代表了单层网络的聚类值

层间Ncut算法：

单词网络层和tweet网络层节点间的互相连接构成了新的网络，也可以用同样的方法进行聚类计算，我们以代表层间聚类。此时的tweet-单词网络定义为，注意此处的权重矩阵为非对称矩阵

MuNcut算法：

此时，便可以定义算法：



其中，为权重调节参数

模拟退火算法：

为了得到最优的聚类结果，即得到上述算法的最小值，我们采用模拟退火算法。在第t次迭代中，我们以作为第t次的聚类的结果，代表第t次的聚类目标函数值。另外设定B为最大的迭代次数，B值的设定非常重要，需要足够大。令温度函数为。其中L为调节参数，需要根据结果设定。有关参数的详细确定方法在本片论文中被省略掉。下面是迭代过程：

1. 设定K值，即想得到的类数值，然后随机把tweet和word分为K组，即然后根据得到的tweet分组和word分组确定两个网络连接网络的分组情况。在我们的模拟实验中，初始分组对最终的结果没有影响
2. 令，设定作为产生新分组的变化程度，随着迭代次数的进行，

递减。

1. 随机的从中拿出i，然后令，。重复次。
2. 如果的话，最好的分类结果任然为，如果的，令以概率仍然作为最好结果，否则。
3. 重复第2-4步直到

在模拟退火算法中，我们随着温度的降低，随机变化形成新类的随机次数减少，有效的保留了前面生成新类的比较好的信息，使着迭代朝着负梯度方向进行。

交叉验证确定参数：

稳定性验证：进行若干次验证

MuNcut算法后主题分析：

得到的单词进行主题对应

得到的推特进行主题对应

按时间段，可以看出主题的变化：

整体看，可以看出主要表达的主题

同一类中（单词，tweet）相关性分析：大多数单词存在与tweet中，反映了同样的主题

MuNcut算法的意义：

本身算法设计说明：

特朗普tweet验证说明：

背景：

模型：

结果：

关键词：

背景：

研究推特治国的背景：

文本处理的背景：

模型：

MuNcut：

计算：

结果：

分析：

在分类得到的每一组，按照文本中出现频率的多少进行排序可以看出十个主题词基本为：

1. fake news
2. make amerian great again
3. 克林顿
4. Whitehouse
5. Health care
6. China
7. Job
8. Refugee
9. Mexico
10. Terror

同理，得到的推文内容典型性代表也如：

同样通过tweet内容，可以看出其主题相关

三个阶段主要主题归结变化：

第一阶段主要主题：

第二阶段主要主题：

第三阶段主要主题：

讨论：

稳定性验证：

为了验证其稳定性，我们设计了稳定性验证的一般方法：

即衡量每次（word，tweet）分组的相似性，可以看出每次模拟退火的相似性稳定在90%以上。

结论：

参考文献：