**《告台湾同胞书》智能分析**

小组成员及份工：

张宗毓：LDA实现、大作业文档、说明文档

滕博文：词云实现、PPT完成

张海波：数据爬取、大作业文档

目录

[1. 项目简介 2](#_Toc30456)

[1.1目的意义 3](#_Toc29378)

[1.2实现方法 3](#_Toc15120)

[2. 分词 3](#_Toc27671)

[2.1概述 3](#_Toc17504)

[2.2基于统计分词 4](#_Toc7642)

[2.3基于统计的中文分词步骤 5](#_Toc15644)

[2.4 基于统计的分词方法的语言模型 5](#_Toc9266)

[3. 关键词提取 9](#_Toc2187)

[3.1概述 9](#_Toc5663)

[3.2基于主题模型的关键词提取 10](#_Toc26423)

[3.2.1 什么是LDA模型 10](#_Toc18788)

[3.2.2 LDA贝叶斯模型 10](#_Toc6578)

[3.2.3 多项分布 10](#_Toc7868)

[3.2.4共轭先验分布 12](#_Toc24177)

[3.2.5 LDA主题模型 12](#_Toc24776)

[4. 结果展示 13](#_Toc19339)

[4.1 词云展示 13](#_Toc21337)

[4.2词频统计： 15](#_Toc19293)

[5. 结果分析 15](#_Toc15863)

[6. 总结与展望 16](#_Toc31587)

# 项目简介

## 1.1目的意义

台湾问题一词及其相关概念，或者叫台海问题或者两岸问题，是指从1949年中华人民共和国成立与以蒋介石为首的国民党集团从中国大陆败退台湾，所衍生的维护国家主权和领土完整的问题。最早出现于1955年寓居美国的李宗仁之《对台湾问题的具体建议》一文，并因强力宣传而通行于中国民间以至国际中国台湾岛社会迄今。

从1949年12月国民党败退台湾开始，大陆一直奉行武力解放台湾的政策。在1950年6月25日，朝鲜战争爆发，美军进入朝鲜，美军的第七舰队驶向台湾海峡，致使大陆方面面临着来自东北和东南两方面的军事威胁，因此毛泽东和中共中央决定暂时推迟渡海攻台。扫清大陆地区残余的国民党势力以及国家周边环境的情况下，我国逐渐开始了社会主义改造，又进行了第一个五年计划的经济建设，国内的环境愈发变好。台湾方面，美国与蒋介石政府的合作出现了矛盾，并且国际形势缓和，呼唤和平的声音高涨，在这样的情况下，大陆又提出了和平解放台湾政策。

台湾问题向来是国内外争论了焦点话题，以往的分析都是台海问题的专家亲力亲为，对大陆和台湾方面的发言和政策进行分析得出结论。这样的方法在现在仍不过时，但是它需要大量的人力物力，并且效率相对来说比较低。在网络技术飞速发展的今天，使用网络技术分析相关问题会更加的省时省力，而且还能得到更加新颖更加准确的观点。

## 1.2实现方法

从网上下载了多篇大陆对台态度的发言稿和台湾地区领导人上台发言稿，其中包括了习近平主席在2019年1月2日在《告台湾同胞书》发表40周年纪念会上的讲话，从1950年到2009年大陆的领导人和中央部门关于《告台湾同胞书》的6次讲话文稿，还有目前台湾地区领导人蔡英文上台的完整发言稿，以及她双十讲话完整稿。

利用Pycharm和Jupyter Notebook工具运行代码实现相关功能。

PyCharm是一种Python IDE，带有一整套可以帮助用户在使用Python语言开发时提高其效率的工具，比如调试、语法高亮、Project管理、代码跳转、智能提示、自动完成、单元测试、版本控制。

Jupyter Notebook的本质是一个Web应用程序，便于创建和共享文学化程序文档，支持实时代码，数学方程，可视化和[markdown](https://baike.baidu.com/item/markdown/3245829" \t "https://baike.baidu.com/item/_blank)。用途包括：数据清理和转换，数值模拟，统计建模，机器学习等等。

# 分词

## 2.1概述

中文分词(Chinese Word Segmentation) 指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。

现有的分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法

基于字符串匹配的分词方法：这种方法又叫做机械分词方法，它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行配，若在词典中找到某个字符串，则匹配成功（识别出一个词）

1）正向最大匹配法（由左到右的方向）

2）逆向最大匹配法（由右到左的方向）:

3）最少切分（使每一句中切出的词数最小）

4）双向最大匹配法（进行由左到右、由右到左两次扫描）

基于理解的分词方法：这种分词方法是通过让计算机模拟人对句子的理解，达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断，即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段。

基于统计的分词方法：给出大量已经分词的文本，利用统计机器学习模型学习词语切分的规律（称为训练），从而实现对未知文本的切分。例如最大概率分词方法和最大熵分词方法等。随着大规模语料库的建立，统计机器学习方法的研究和发展，基于统计的中文分词方法渐渐成为了主流方法。

主要统计模型：N元文法模型（N-gram），隐马尔可夫模型（Hidden Markov Model ，HMM），最大熵模型（ME），条件随机场模型（Conditional Random Fields，CRF）等。

## 2.2基于统计分词

基于机械匹配的中文分词算法非常简单和方便，但是对于歧义词以及未登录词的分词效果不好，因为这些词语是词典里没有的。实际上，中文分词的主要瓶颈是切分排歧和未登录词识别。切分歧义和未登录词降低了自身正确切分的可能性，同时还干扰了其相邻词的正确处理。更糟糕的是，未登录词往往和切分歧义交织在一起，进一步增加了分词的难度。如：在 “克林顿对内塔尼亚胡说”中，“内塔尼亚胡”是一个词典中没有收录的译名，实际切分的时候，“对”与“内”，“胡” 与“说”往往会粘在一起，最终导致错误的切分结果：“克林顿/对内/塔尼亚/胡说/”。并且，基于机械匹配的中文分词算法没有考虑词语与词语之间的关系，也没有进行语法方面的考量。而基于统计的分词方法，优势就在于处理歧义，对于识别未登录词拥有较好的识别能力，分词精度比较大。现在基于统计的中文分词方法渐渐成为了主流方法。

基于机械匹配的方法是理性主义研究方法，基于统计的方法是经验主义研究方法。两种方法最明显的区别是：基于统计的分词方法摈弃了词典，在进行分词的时候不需要词典作为输入，而是需要输入各种各样的语料库，通过训练出的语言模型来分词。

基于统计的分词方法通过对大众语料进行统计、分析、提取和加工，得到用标识(tags)标记的标准化语料库供进一步的研究，如马尔可夫模型(Markov model)语法应用。这种基于统计分词方法有许多的应用：早期以真实语料为基础的语言学研究，有夸克和里奇等编写的《英语语法大全》。知名度高、影响力大的计算机语料库则有20世纪60年代由美国Brown大学整理的Brown语料库，约含100万个经过标注的词汇。第一本基于语料库语言学出版的词典是波士顿出版商霍顿一米福林(Houghton-Mifflin)出版的《美国传统词典》。继Brown美式英语言语料库产生之后，柯林斯(Collins)出版商出版了由英国挪威等国家的大学学者整理的COBUILD语料库，在此基础上研究的成果有诸如《柯林斯英语大辞典》等柯林斯英语学习系列词典300余种，可以说是语料库语言学在词典方面开发应用最广泛的一个例子。

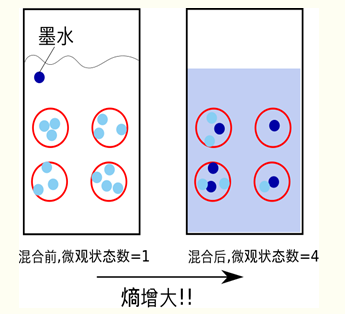
## 2.3基于统计的中文分词步骤

第一步，加载大规模中文语料集；这里需要输入中文语料集一般是包含大规模（十万数量级以上）的中文句子的文档。第二步，通过大规模中文语料集训练出相应的语言模型；这个“训练”的意思其实是，利用统计机器学习模型来学习词语切分的规律。第三步才是进行中文分词。

## 2.4 基于统计的分词方法的语言模型

**（1）熵**

熵的概念最早来源于物理学，熵是对一个随机事件不确定性的衡量。简单来说，东西变得无序化的过程，就是熵增大的过程。熵越大，系统越混乱无序。例如，房间变乱的过程，熵在增大。与此类似的还有，一滴墨水均匀地分布到一杯清水的这个过程中，熵也在增大。因为杯子里的墨水和水逐渐混合到了一起，当它们混合均匀时，熵最大，因为均匀分布时熵最大的模型。

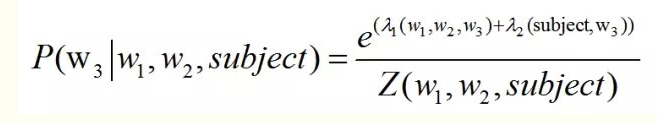


**（2）最大熵模型**

最大熵原理是统计学的原理，也是概率模型学习的一个准则。最大熵原理认为，学习概率模型时，在所有可能的概率模型中，熵最大的模型是最好的模型。

举个例子，一个色子有6个面，问人掷到每个面的概率分别是多少，大家一定会说每面都是1/6。为什么大家会说1/6，而不是说什么1是1/3，2是1/2这样子的组合呢？因为大家潜意识里面觉得这样子最“保险”。“最保险”其实就是因为均匀分布刚好是熵最大的模型。最大熵模型认为，对于那些未知的事件，认为他们等可能，是最好的，也满足最大熵的情况。在已知条件下选择一个合适的分布来预测可能出现的事件，最主要的思想是在只掌握未知分布的部分知识时，应该选取符合这些知识，但熵值最大的概率分布。不确定性越大，熵值就越大；若随机变量退化成定值，熵为0。

下面公式是根据上下文（前两个词）和主题预测下一个词的最大熵模型，其中 w3 是要预测的词w1 和 w2 是它的前两个字（，也就是其上下文的一个大致估计，subject 表示主题。



由于最大熵模型在数学上十分完美，对科学家们有很大的诱惑力，因此不少研究者试图把自己的问题用一个类似最大熵的近似模型去套。但近似之后最大熵模型就变得不完美了，所以实验出来的结果并不怎么好。于是很多人又放弃了这种方法。而第一个在实际信息处理应用中验证了最大熵模型的优势的人是一个叫拉纳帕提(Adwait Ratnaparkhi)的人，原IBM现微软的研究员。拉纳帕提的聪明之处在于他没有对最大熵模型进行近似，而是找到了几个最适合用最大熵模型、而计算量相对不太大的自然语言处理问题，比如词性标注和句法分析。拉纳帕提成功地将上下文信息、词性（名词、动词和形容词等）、句子成分（主谓宾）通过最大熵模型结合起来，做出了当时世界上最好的词性标识系统和句法分析器。拉纳帕提的论文发表后让人们耳目一新。拉纳帕提的词性标注系统，至今仍然是使用单一方法最好的系统，这个系统也使人们看到了用最大熵模型解决复杂的文字信息处理的希望。

**（3）HMM隐马尔可夫模型**

隐马模型是一种非常经典的统计模型，它用来描述一个含有隐含未知参数的马尔可夫过程。下面用一个简单的例子来解释HMM模型：

假设这里有三个不同的骰子。第一个骰子是最常见的6面的骰子（称这个骰子为D6），每个面（1，2，3，4，5，6）出现的概率是1/6。第二个骰子是个4面的骰子（称这个骰子为D4），每个面（1，2，3，4）出现的概率是1/4。第三个骰子有8面的骰子（称这个骰子为D8），每个面（1，2，3，4，5，6，7，8）出现的概率是1/8。

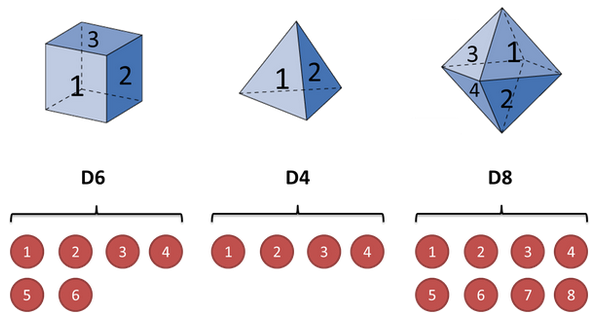
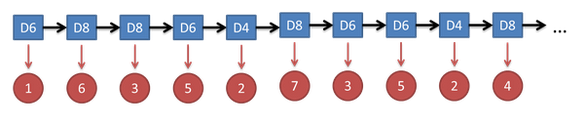


图3-4

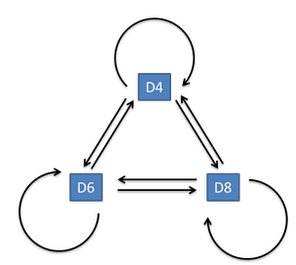
假设我们开始掷骰子，我们先从三个骰子里挑一个，挑到每一个骰子的概率都是1/3。然后我们掷骰子，得到一个数字，1，2，3，4，5，6，7，8中的一个。然后不停的重复上述过程，我们就会得到一串数字，每个数字都是1，2，3，4，5，6，7，8中的一个。假设我们掷骰子10次，然后得到这么一串数字：1 6 3 5 2 7 3 5 2 4。那么这串数字就是可见状态链。但是在隐马尔可夫模型中，我们不仅仅有这么一串可见状态链，还有一串隐含状态链。在这个例子里，这串隐含状态链就是所用的骰子的序列。比如，隐含状态链有可能是：D6 D8 D8 D6 D4 D8 D6 D6 D4 D8。

一般来说，HMM中说到的马尔可夫链其实是指隐含状态链，因为隐含状态（骰子）之间存在转换概率（transition probability）。在我们这个例子里，D6的下一个状态是D4，D6，D8的概率都是1/3。D4，D8的下一个状态是D4，D6，D8的转换概率也都一样是1/3。这个例子里设定的概率是相等的，但其实转换概率是可以随意设定的，比如，我们可以这样定义，D6后面不能接D4，D6后面是D6的概率是0.9，是D8的概率是0.1。这样就能够构成一个新的HMM。同样的，尽管可见状态之间没有转换概率，但是隐含状态和可见状态之间有一个概率叫做输出概率（emission probability）。

同样的，尽管可见状态之间没有转换概率，但是隐含状态和可见状态之间有一个概率叫做输出概率（emission probability）。就我们的例子来说，六面骰（D6）产生1的输出概率是1/6。产生2，3，4，5，6的概率也都是1/6。我们同样可以对输出概率进行其他定义。比如，我有一个被赌场动过手脚的六面骰子，掷出来是1的概率更大，是1/2，掷出来是2，3，4，5，6的概率是1/10。



隐含状态转换关系如下：



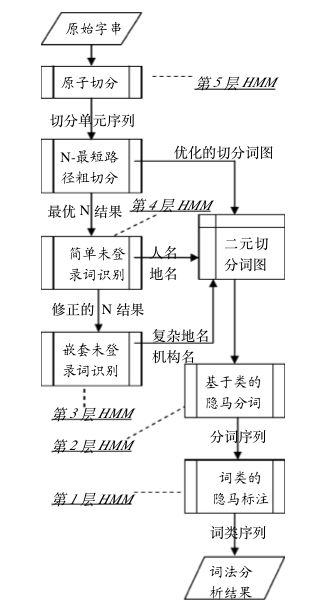
**（4）层叠隐马模型**

隐马模型是一种非常经典的统计方法，但是相对于复杂的自然语言现象来说，传统的隐马模型显得略为简单。所以，有时候需要采用多个层次的隐马模型，对汉语词法分析中遇到的不同情况，进行分别处理，即为层次隐马模型HHMM。层次隐马模型是一种不同于隐马模型的数学模型，它更为复杂，并且比隐马模型具有更强的表达能力，但使用起来的时空开销也比较大。层次隐马模型求解的时间复杂度是O(NT3)，而隐马模型只有O(NT)。

而层叠隐马模型实际上是若干个简单隐马模型的组合，各层隐马模型形成一种紧密的耦合关系。整个系统的时间复杂度与隐马模型相同，为O(NT)。

基于层叠隐马模型的中文分词算法步骤：第一，在预处理的阶段，采取N-最短路径粗分方法，快速得到能覆盖歧义的最佳N个粗分结果；第二，在粗分结果集上，采用底层隐马模型识别出普通无嵌套的人名、地名，并依次采取高层隐马模型识别出嵌套了人名、地名的复杂地名和机构名；第三，未登录词与歧义词都不作为特例，而是与普通词一起参与各种候选结果的竞争；第四，在全局最优的分词结果上进行词性的隐马标注。

该方法已经应用到中科院计算所汉语词法分析系统ICTCLAS中，这个系统是目前最好的汉语词法分析系统之一。下图为应用的基于CHMM的汉语词法分析框架：



# 关键词提取

## 3.1概述

在自然语言处理领域，处理海量的文本文件最关键的是要把用户最关心的问题提取出来。而无论是对于长文本还是短文本，往往可以通过几个关键词窥探整个文本的主题思想。与此同时，不管是基于文本的推荐还是基于文本的搜索，对于文本关键词的依赖也很大，关键词提取的准确程度直接关系到推荐系统或者搜索系统的最终效果。因此，关键词提取在文本挖掘领域是一个很重要的部分。

关于文本的关键词提取方法分为有监督、半监督和无监督三种：

1、有监督的关键词抽取算法：它是建关键词抽取算法看作是二分类问题，判断文档中的词或者短语是或者不是关键词。既然是分类问题，就需要提供已经标注好的训练预料，利用训练语料训练关键词提取模型，根据模型对需要抽取关键词的文档进行关键词抽取。

2、半监督的关键词提取算法：只需要少量的训练数据，利用这些训练数据构建关键词抽取模型，然后使用模型对新的文本进行关键词提取，对于这些关键词进行人工过滤，将过滤得到的关键词加入训练集，重新训练模型。

3、无监督的方法：有监督的文本关键词提取算法需要高昂的人工成本，因此现有的文本关键词提取主要采用适用性较强的无监督关键词抽取。其文本关键词抽取流程如下：

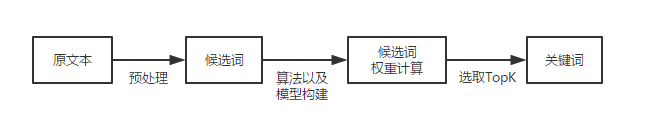


图1 无监督文本关键词抽取流程图

无监督关键词抽取算法可以分为三大类，基于统计特征的关键词抽取、基于词图模型的关键词抽取和基于主题模型的关键词抽取。

## 3.2基于主题模型的关键词提取

### 3.2.1 什么是LDA模型

LDA(Latent Dirichlet Allocation)是一种文档生成模型。它认为一篇文章是有多个主题的，而每个主题又对应着不同的词。一篇文章的构造过程，首先是以一定的概率选择某个主题，然后再在这个主题下以一定的概率选出某一个词，这样就生成了这篇文章的第一个词。不断重复这个过程，就生成了整片文章。当然这里假定词与词之间是没顺序的。

LDA的使用是上述文档生成的逆过程，它将根据一篇得到的文章，去寻找出这篇文章的主题，以及这些主题对应的词。

LDA是非监督的机器学习模型，并且使用了词袋模型。一篇文章将会用词袋模型构造成词向量。

### 3.2.2 LDA贝叶斯模型

LDA是基于贝叶斯模型的，涉及到贝叶斯模型离不开“先验分布”，“数据（似然）”和"后验分布"三块。在贝叶斯学派这里：

先验分布 + 数据（似然）= 后验分布

比如你对好人和坏人的认知，先验分布为：100个好人和100个的坏人，即你认为好人坏人各占一半，现在你被2个好人（数据）帮助了和1个坏人骗了，于是你得到了新的后验分布为：102个好人和101个的坏人。现在你的后验分布里面认为好人比坏人多了。这个后验分布接着又变成你的新的先验分布，当你被1个好人（数据）帮助了和3个坏人（数据）骗了后，你又更新了你的后验分布为：103个好人和104个的坏人。依次继续更新下去。

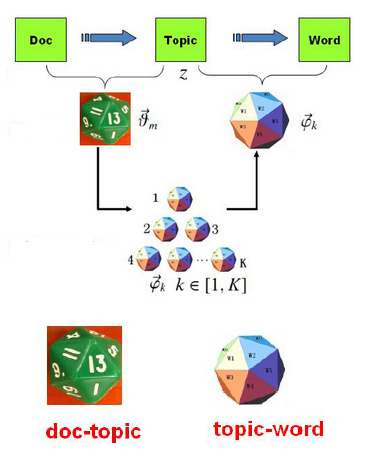
### 3.2.3 多项分布

多项式分布（multinomial）是二项分布在两点分布上的延伸。在两点分布中，一次实验只有两种可能性，p以及 （1-p）。例如抛一枚硬币，不是正面就是反面。在多项式分布中，这种可能的情况得到了扩展。例如抛一个骰子，一共有6种可能，而不是2种。

设某随机实验如果有k个可能情况 A1、A2、…、Ak，分别将他们的出现次数记为随机变量X1、X2、…、Xk，它们的概率分布分别是p1，p2，…，pk，那么在n次采样的总结果中，A1出现n1次、A2出现n2次、…、Ak出现nk次的这种事件的出现概率P有下面公式：



主题和词是符合多项分布的，这样便可以用骰子形象地表达一篇文章的生成的过程：



有两类骰子，一种是文章-主题（doc-topic）骰子，骰子的每面代表一种主题。这里设一共有K个主题，则K面。骰子的各个面的概率记为ϑ⃗=(p1,p2,p3,...,pk)。各个面的概率即为这个多项式分布的参数。

另一种骰子为主题-词(topic-word)骰子，一共有K个，从1~K编号，分别对应着不同的主题。骰子的一个面代表一个单词。由于有K个骰子，把不同主题-词骰子各个面的概率分别记为φ1,φ2,...φk。对于一个主题-词骰子，他的各个面的概率即为这个多项式分布的参数。

那么一篇文章的生成过程可以表示为：

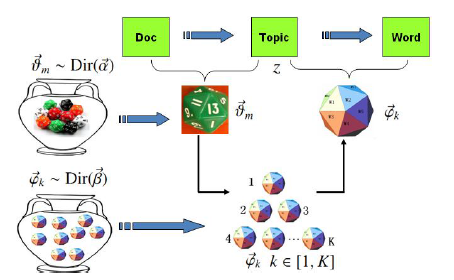
1.抛掷这个doc-topic骰子，得到主题编号z；

2.选择编号为z的topic-word骰子，得到词w；

3.不断重复步骤1以及步骤2。

### 3.2.4共轭先验分布

通过贝叶斯学派的观点对plsa模型进行改造。参数也是服从一定的分布的。



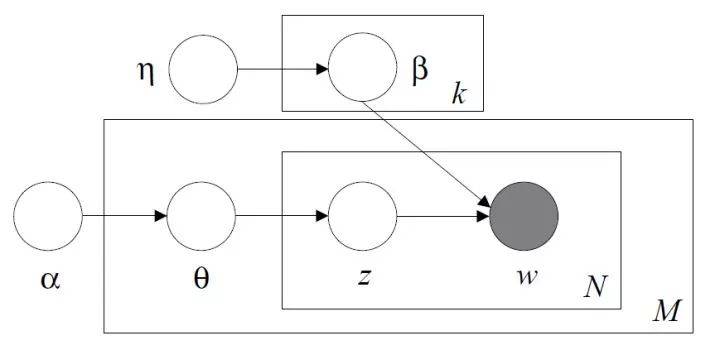
原本绿色骰子（doc-topic分布）是P(θ)先验概率，现在成了P(θ|x)后验概率。词分布也是同样的。我们已经知道，主题分布以及词分布式服从多项式分布的。那么他们的参数要服从什么分布呢？我们知道，P(ϑ|x)∝P(x|ϑ)P(ϑ)，而P(θ|x)服从多项式分布，因此P(θ)P(x|θ)归一化后也要服从多项式分布。

在贝叶斯概率理论中，有这么一种定义，如果后验概率P(θ|x)和先验概率p(θ)满足同样的分布律，那么先验分布和后验分布被叫做共轭分布，同时，先验分布叫做似然函数的共轭先验分布。

那么，我们就可以选择似然函数P(x|θ)的共轭先验作为P(θ)的分布。而Dirichlet分布正是多项式分布的共轭先验概率分布。

### 3.2.5 LDA主题模型

文档的生成模型可以用如下图模型表示：



其中α和η为先验分布的超参数，β为第k个主题下的所有单词的分布，θ为文档的主题分布，w为文档的词，z为w所对应的主题。

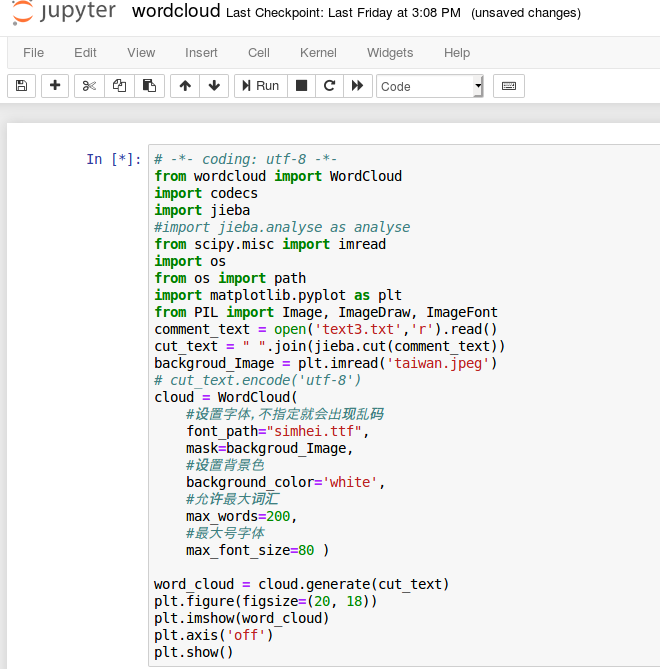
LDA一般有两种方法求解我们想要的每一篇文档的主题分布和每一个主题中词的分布，第一种是基于Gibbs采样算法求解，第二种是基于变分推断EM算法求解。

# 结果展示

## 4.1 词云展示

首先使用基于HMM的分词方法将文本数据进行分词，之后在jupyter上使用词云的方式展示关键词。

jupyter程序如下所示：



展示结果如下：

下面为2019年1月2日习大大在《告台湾同胞书》发表40周年纪念会上的讲话词云图：



下面为从1950年到2009年大陆的领导人和中央部门关于《告台湾同胞书》的6次讲

话文稿词云图：

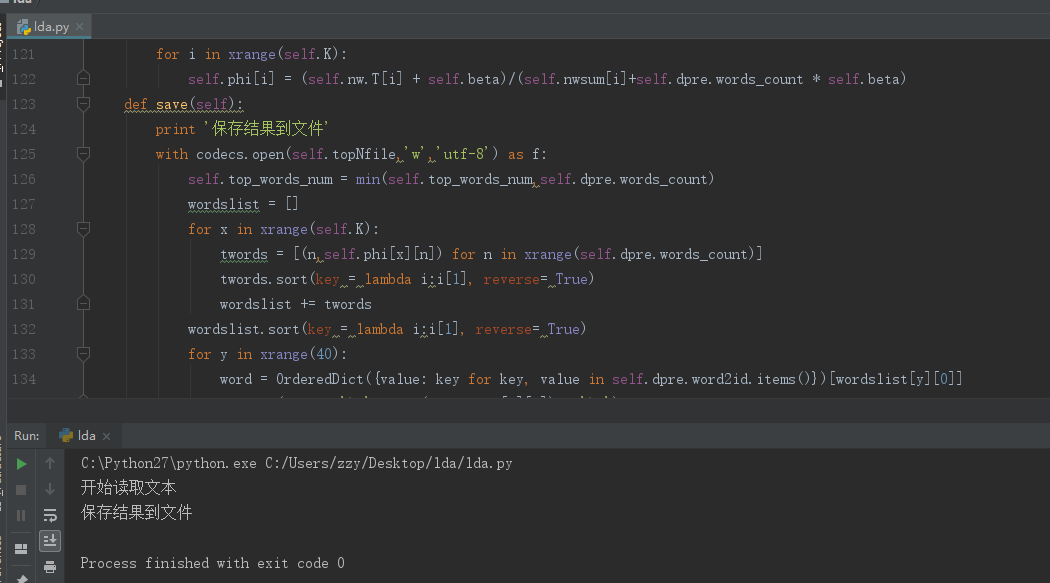


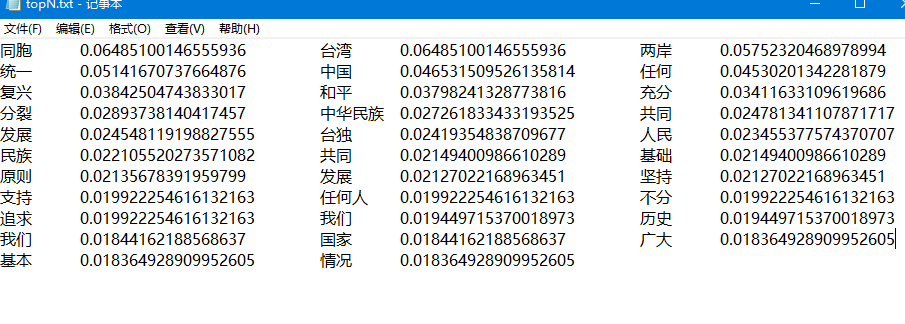
下面为目前台湾地区领导人蔡英文上台的完整发言稿，以及她双十讲话完整稿词云图：



## 4.2词频统计：

使用LDA主题模型分析得到主题下各关键词统计，以习大大《告台湾同胞书》为例运行结果如下：





对于习大大的《告台湾同胞书》位于前列的关键词是：“台湾同胞”、“统一”、“两岸”、“中国”、“任何”、“台湾”、“统一”、“复兴”、“民族”、“中华民族”。

对于以往《告台湾同胞书》位于前列的是：“台湾”、“中国”、“两岸关系”、“中华民族”、“伟大复兴”、“发展”、“合作”、“人民”、“和平”、“历史”。

而在蔡英文的发言中，词频位于前列的关键词是：“台湾”、“经济”、“国家”、“台湾人民”、“国际”、“政府”、“两岸”、“两岸关系”、“社会”、“国人”。

# 结果分析

结合程序的运行结果，我们进行了初步的分析，得出如下结论：

大陆方面向来重视两岸关系的推进，期待和平统一，并且随着时间的推移，这种期待愈发的强烈，渴望中华民族的伟大复兴。

台湾方面的发言，并没有体现出强烈的统一意愿，他们重视台湾地区的经济、社会发展，希望两岸关系的友好。

# 6. 总结与展望

大陆方面向来重视两岸关系的推进，期待和平统一，并且随着时间的推移，这种期待愈发的强烈，渴望中华民族的伟大复兴。台湾方面的发言，并没有体现出强烈的统一意愿，他们重视台湾地区的经济、社会发展，希望两岸关系的友好。

关键词提取算法和分词算法融合进行分析传统文本分析，可以将新世纪科技与传统文本分析结合起来，分析结果更准确，表现方式清晰易懂，能够提高准确率，节省人力物力，提高效率，增加软件的复用性，以及降低成本。

日后能继续发展，比如自动爬去各国政府官方网站的新闻动态，包括对中国方面的新闻但不仅仅限于中国方面的，然后再自动进行分析，最后呈现给国家领导人，这样就会节省了巨大的人力物力，然后还能及时的让我国对其他国家的政策作出反应，以消除或减少不必要的损失。而且还可以用来分析国际上持续存在并且难以解决的事件有关的官方报道或者声明，比如说南海问题、朝鲜核危机、伊朗核危机等其他事件，从而更好的把握某个事件，并及时作出应对措施。