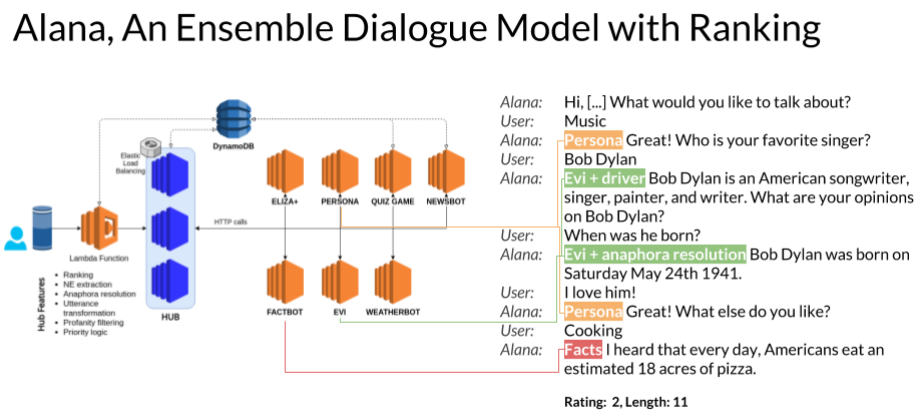
自然语言理解：Alana

1 Alana背景介绍

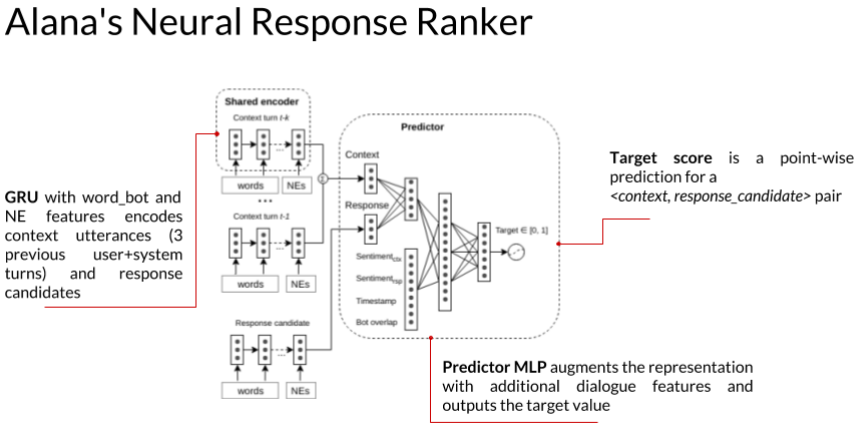
Alana AI 公司成立于 2015 年，在墨西哥和波多黎拥有办事处。今天，Alana 人工智能公司已经在巴西、美国和英国开展了相关业务。

Alana AI 已经实现了销售额 1000 万 R$ 大关，向可口可乐、Diageo、Polishop 和 Nivea 等大型企业提供其专用人工智能后，该公司的目标是未来 12 个月内在拉丁美洲创造 300 万美元的销售额。

自主机器人是高度配置的智能机器人，它们能够自主完成某些任务，而无需人工控制。例如，机器人可以说话、走路、打扫房间、开门等，可以像人类一样自行做出决定并相应地执行任务。Alana 机器人是一个自主的聊天机器人，利用 AI 技术来延续人类的表达方式，并使用机器学习算法像人类一样“学习”。Jim Al-khali 教授在 BBC 上首次介绍了 Alana，这是一个可以像人类一样增强、模拟和表演的机器，涵盖了机器学习、人工神经网络、NLP、NLU 等技术。下面将一一介绍Alana机器人所用到的上述技术，以及目前人工智能领域关于自然语言处理研究的现状、趋势以及遇到的一些瓶颈。



Papaioannou et al. Alana: Social Dialogue using an Ensemble Model and a Ranker trained on User Feedback, 2017



1.1 自然语言处理——NLP

1.1.1 定义

自然语言处理(Natural Language Processing, NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法、研究在人与人交互中以及在人机交互中的语言问题。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系，但又有重要的区别。自然语言处理并不是一般地研究自然语言，而在于研制能有效地实现自然语言通信的计算机系统，特别是其中的软件系统。因而它是计算机科学的一部分。

1.1.2 研究方向

信息抽取、文本生成、问答系统、对话系统、文本挖掘、语音识别与生成、信息过滤、网络舆情分析、信息检索、机器翻译等等。

1.1.3 发展历史

- 早期自然语言处理阶段

  第一阶段(60~80年代):基于规则来建立词汇、句法语义分析、问答、聊天和机器翻译系统。好处是规则可以利用人类的内省知识，不依赖数据，可以快速起步；问题是覆盖面不足，像个玩具系统，规则管理和可扩展一直没有解决。

- 统计自然语言处理阶段

  第二阶段(90年代开始)：基于统计的机器学习(ML)开始流行，很多NLP开始用基于统计的方法来做。主要思路是利用带标注的数据，基于人工定义的特征建立机器学习系统，并利用数据经过学习确定机器学习系统的参数。运行时利用这些学习得到的参数，对输入数据进行解码，得到输出。机器翻译、搜索引擎都是利用统计方法获得了成功。

- 神经网络自然语言处理阶段

  第三阶段(2008年至今)：深度学习开始在语音和图像发挥威力。随之，NLP研究者开始把目光转向深度学习。先是把深度学习用于特征计算或者建立一个新的特征，然后在原有的统计学习框架下体验效果。比如，搜索引擎加入了深度学习的检索词和文档的相似度计算，以提升搜索的相关度。自2014年以来，人们尝试直接通过深度学习建模，进

行端对端的训练。目前已在机器翻译、问答、阅读理解等领域取得了进展，出现了深度学习的热潮。

1.1.4 Alana中的自然语言处理

位于加利福尼亚州的 OpenAI实验室在今年推出了一种新的人工智能语言模型——GPT-3，（**Generative Pre-trained Transformer**）。新一代程序只不过是执行简单动作（自动完成）的文本预测器，但代表了 NLP 领域的一项重要技术进步。

Alana 使用了该模型，主要是因为该程序能够生成高度复杂的文本，并且可以通过人造内容，例如 机器为卫报撰写的这篇文章。另一方面，正如 Alana AI 的 CTO Marcellus Amadeus 在接受 Exame 采访时提到的那样，该程序非常强大，因为它具有读取大量信息的能力，但仍然不知道它们与现实之间的关系。

1.2 自然语言理解——NLU

1.2.1 三种实现方式

- 分布语义（Distributional semantics）
- 框架语义（Frame semantics）
- 模型论语义（Model-theoretic semantics）

目前采用的是框架语义表示的一种变形——采用领域（domain）、意图（intent）和属性槽（slots）来表示语义结果。

1.2.2 应用场景

NLU 是 NLP 的一个特定领域，是个人智能助理的基础模块和核心模块，我们目前的自然语言理解服务同时实现了无上下文的理解和有上下文的理解，无上下文的理解主要应用在搜素场景，有上下文的理解主要应用于对话场景。

1.2.3 Alana中的自然语言理解

Alana 使用 NLU 来理解上下文、检测情绪、理解语音模式，甚至回忆以前的对话。无论您选择如何具体地表达命令或短语， Alana 对话式 AI 都能够准确地解释您希望从对话中实现的目标。

1.3 Alana中的机器学习

机器学习允许 Alana 根据数据和经验的组合来选择行为和响应方式。它还允许对话式 AI 助手做出的决策和预测随着时间的推移逐渐变得更加准确。这种更加自动化和自主的数据使用消除了对每个响应进行专门编程的必要性，从而减少了所需的人工干预量。

1.4 Alana中的深度学习

一种更高级的“无监督”机器学习模型。深度学习技术不依赖于人类输入的规则，而是使用自己的推理来做出决策。这种逻辑是由多层算法提供的，这些算法创建了一个模仿人脑的人工神经网络。因此，基于深度学习的对话式 AI 需要更少的人类指导和纠正来提供更加人性化、智能化和更加准确的响应。

1.5 自然语言处理领域最新进展

1.5.1 现状与趋势

过去几年，由于数据越来越多，出现各种测试集；算法越来越复杂、越来越先进，包括神经网络的架构、预训练模型等等；计算能力越来越高，在这三大因素的作用下，自然语言处理得到了飞速的发展。

目前随着感知智能的大幅度进步，人们的焦点逐渐转向了认知智能。其中语言智能，也就是自然语言理解，则被认为是皇冠上的明珠。一旦有突破，则会大幅度推动认知智能，并提高人工智能的技术，并促进在很多重要场景落地。

未来，自然语言处理技术将沿着致力于实现智能化、人性化的搜索推荐、语音交互、语义理解的道路继续前行。相信不仅会有更多技术难题被攻克，也会有越来越多类似于“莎曼萨”的产品问世。

随着大数据技术的不断发展，大规模语料样本数据以惊人的数量不断积累以及自然语言处理在深度学习方面的不断深耕，目前业界已经开始使用上万小时的样本进行模型训练。不难预测，不久，自然语言处理技术发展将很快进入10万小时数据样本训练阶段，只有这样，才能覆盖千差万别的用户口音差异、多领域歧义语料数据以及复杂的语法规则。再考虑环境变化的影响，未来训练语料量可能会突破100万小时。未来，基于统计学的语义分析方法研究将会继续深化，会随着大规模语料样本数据的不断积累以及大数据挖掘技术、深度模型算法的不断发展呈现质的飞跃。

随着训练数据量的迅速增加，如何实现大规模LSTM（长短时记忆模型）建模和CTC（连接时序分类）的有效训练，会成为一个核心的技术难题。未来语音识别领域的深度学习将进入数百GPU并行训练的状态，理论创新和算法技术创新都将围绕大数据展开。语音识别技术的研发方法，相对于现在必将发生深刻的变革。此外，CTC建模技术进一步降低了语音识别应用的解码成本，随着适合深度模型计算的专业硬件的大量涌现，语音识别云服务的成本将大量降低，从而推动语言处理与语音交互技术的更大范围的普及。

1.5.2 需求与挑战

随着智能硬件技术与移动技术的蓬勃爆发，自然语言处理技术的应用趋势也发生了变化。一方面用户要求自然语言处理技术可以精准地理解自己的需求，而且直接给出最匹配的答案，而非简单地给出网址让用户自己去找答案（起码目前代表业内较高水平的小度机器人还是这样做的）。另一方面是需要自然语言处理技术可以与用户进行对话式搜索与智能交互。在这样的需求下，对于自然语言处理技术的未来发展提出了很大的挑战。它要求未来的自然语言处理技术能够做到：

1. 需求识别。通过用户提出了多种多样的、复杂的、基于情感式的、语意模糊的需求进行深刻分析，精确地理解用户的需求。
2. 知识挖掘。经过海量的网络数据与知识的挖掘分析，将各种结构化、非结构化、半结构化的知识进行组织与梳理，最终以结构化、清晰化的知识形式完整地呈现给用户。
3. 用户引导。这与对话式智能交互相关，不仅根据用户的需求来提供“建议”，还能“猜测”用户可能会有什么未想到、未提出的需求，从而“先人一步”为用户提供相关的扩展信息。
4. 结果组织和展现。由于用户更加青睐直接的答案，答案的形式可以是唯一答案、聚合答案、图片、多媒体的形式，这就要求自然语言处理技术能够将挖掘出的信息进行有效地

组织与整理，以条理化、简洁化、直接化的形式呈现给用户。

2 自然语言处理的流程

2.1 语料获取

语料，即语言材料，是构成语料库的基本单元。 所以，人们简单地用文本作为替代，并把文本中的上下文关系作为现实世界中语言的上下文关系的替代品。我们把一个文本集合称为语料库（Corpus），当有几个这样的文本集合的时候，我们称之为语料库集合(Corpora)。按语料来源，我们将语料分为以下两种：

1. 已有语料纸质或者电子文本资料->电子化->语料库。
2. 网上下载、抓取语料

   国内外标准开放数据集或通过爬虫获取。

2.2 语料预处理

1. 语料清洗

   对于原始文本提取标题、摘要、正文等信息，对于爬虫，去除广告、标签、HTML、JS等代码和注释。
2. 分词

   将短文本和长文本处理为最小单位粒度是词或词语的过程。
3. 词性标注

   对每个词或词语打词类标签，是一个经典的序列标注问题。

   词性标注表：

词性编码	词性名称	注 解
Ag	形语素	形容词性语素。形容词代码为 a，语素代码ｇ前面置以A。
a	形容词	取英语形容词 adjective 的第1个字母。
ad	副形词	直接作状语的形容词。形容词代码 a和副词代码d并在一起。
an	名形词	具有名词功能的形容词。形容词代码 a和名词代码n并在一起。
b	区别词	取汉字“别”的声母。
c	连词	取英语连词 conjunction 的第1个字母。
dg	副语素	副词性语素。副词代码为 d，语素代码ｇ前面置以D。
d	副词	取 adverb 的第2个字母，因其第1个字母已用于形容词。
e	叹词	取英语叹词 exclamation 的第1个字母。
f	方位词	取汉字“方”
g	语素	绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母。
h	前接成分	取英语 head的第1个字母。
i	成语	取英语成语 idiom 的第1个字母。
j	简称略语	取汉字“简”的声母。
k	后接成分	
l	习用语	习用语尚未成为成语，有点“临时性”，取“临”的声母。
m	数词	取英语 numeral 的第3个字母，n，u已有他用。
Ng	名语素	名词性语素。名词代码为 n，语素代码ｇ前面置以N。
n	名词	取英语名词 noun的第1个字母。
nr	人名	名词代码 n和“人(ren)”的声母并在一起。
ns	地名	名词代码 n和处所词代码s并在一起。
nt	机构团体	“团”的声母为 t，名词代码n和t并在一起。
nz	其他专名	“专”的声母的第 1个字母为z，名词代码n和z并在一起。
o	拟声词	取英语拟声词 onomatopoeia 的第1个字母。
p	介词	取英语介词 prepositional 的第1个字母。
q	量词	取英语 quantity 的第1个字母。
r	代词	取英语代词 pronoun 的第2个字母,因p已用于介词。
s	处所词	取英语 space的第1个字母。
tg	时语素	时间词性语素。时间词代码为 t,在语素的代码g前面置以T。
t	时间词	取英语 time的第1个字母。
u	助词	取英语助词 auxiliary
vg	动语素	动词性语素。动词代码为 v。在语素的代码g前面置以V。
v	动词	取英语动词 verb的第一个字母。
vd	副动词	直接作状语的动词。动词和副词的代码并在一起。
vn	名动词	指具有名词功能的动词。动词和名词的代码并在一起。
w	标点符号	
x	非语素字	非语素字只是一个符号，字母 x通常用于代表未知数、符号。
y	语气词	取汉字“语”的声母。
z	状态词	取汉字“状”的声母的前一个字母。
un	未知词	不可识别词及用户自定义词组。取英文Unkonwn首两个字母。(非北大标准，CSW分词中定义)

4. 去停用词

去掉对文本特征没有任何贡献的字词，如标点符号、语气、人称等。

2.3 特征工程

把分词之后的字和词语表示成计算机能够计算的类型。

常用两种表示模型：词袋模型和词向量。

2.3.1 词袋模型

不考虑词语原本在句子中的顺序，直接将每一个词语或者符号统一放置在一个集合（如 list），然后按照计数的方式对出现的次数进行统计。统计词频这只是最基本的方式，TF-IDF 是词袋模型的一个经典用法。

2.3.2 词向量

将字、词语转换为向量矩阵的计算模型。常用的词表示方法如下：

- One-Hot：把每个词表示为一个很长的向量。这个向量的维度是词表大小，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词。eg: [0 0 0 0 0 0 0 0 1 0 0 0 0 ... 0]
- Word2Vec：其主要包含两个模型：跳字模型（Skip-Gram）和连续词袋模型（Continuous Bag of Words，简称 CBOW），以及两种高效训练的方法：负采样（Negative Sampling）和层序 Softmax（Hierarchical Softmax）。值得一提的是，Word2Vec 词向量可以较好地表达不同词之间的相似和类比关系。
- Doc2Vec
- WordRank
- FastText

2.4 特征选择

选择合适的、表达能力强的特征。常用方法：**DF**、 **MI**、 **IG**、 **CHI**、**WLLR**、**WFO** 。

2.5 模型训练

不同的应用需求要使用不同的模型。

- 传统的有监督和无监督等机器学习模型： KNN、SVM、Naive Bayes、决策树、GBDT、K-means 等模型；
- 深度学习模型： CNN、RNN、LSTM、 Seq2Seq、FastText、TextCNN 等。

3 针对文本分类系统的算法改进

文本分类可以为文本提供有序的组织，网络信息的增长使文本分类对信息处理的意义变得更加重要。 本次公案中，我们小组选择了该模块进行学习与探讨，针对需要考虑的因素、系统模型的建立、大致思路的再设计三个方面，整理相关论文的算法思路，并提出一些可能的方案。

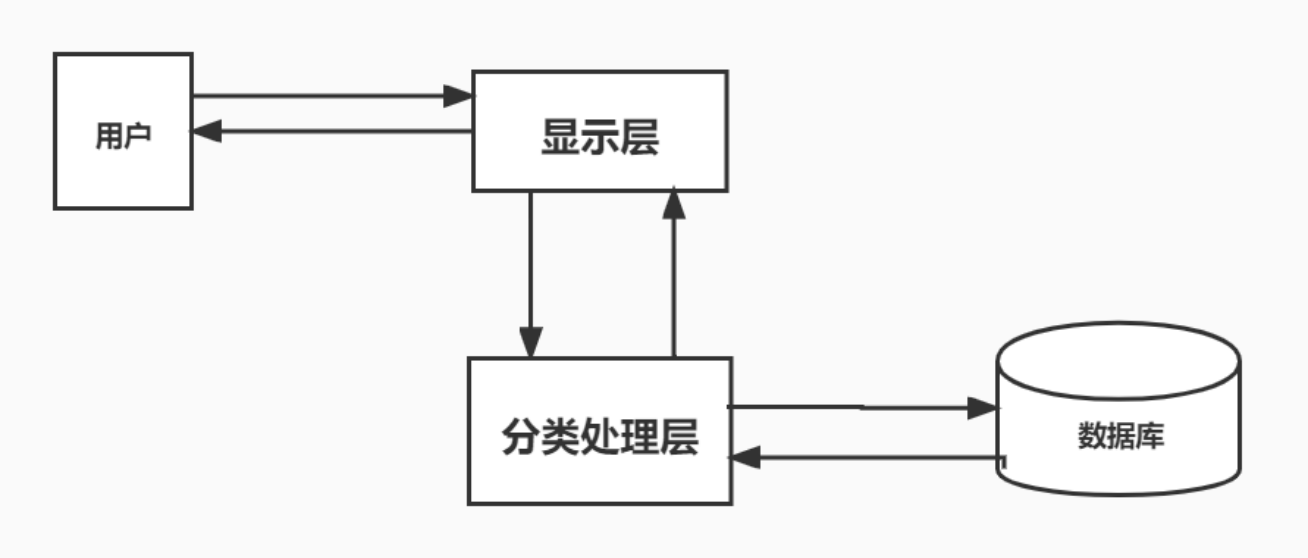
3.1 考虑因素

1. 运行效率
   训练文本分类器通常需要大量的训练样本，对文本进行分词后所得到词的数量很大，如何快速地对这些词进行处理是关系着整个文本分类过程关键问题。
2. 内存消耗
   文本分类涉及的文本数量很大，大量的数据放在内存中容易产生内存溢出,需要采用适当的数据结构存储和处理方案。
3. 可扩展性
   文本分类的每个阶段都有许多算法，这些算法各有优缺点，新的算法不断涌现，为了下

一步的研究工作,需要使系统具有良好的可扩展性。

4. 通用性
   文本分类还处于一种研究的阶段，对每个阶段的算法和技术的评价都能独立于平台和应用环境，即特殊的技术与算法适应于特殊的应用需求，无法一言以蔽之。因此一个通用的文本分类系统应该注意对各种具体实物的抽象。
5. 数据的重用性和可理解性
   文本分类系统在以大量数据为基础的训练过程中，积累了丰富的特征，这些特征有必要进行保存，为其他类似的系统提供有价值的参考，并在实验结束时，对训练情况进行分析，并及时进行可能的修正。也就是说，系统应该提供合适的方法，保存训练等运算的中间结果，为自身的完善和他类似系统的构造提供方便。
6. 友好的图形用户界面
   文本分类系统需要设置许多参数，有些参数需要不断的测试才能找到个最佳值，一个友好的图形用户界面是方便用户进行实验的首要条件。

3.2 系统模型



此分类模型采用模块化的设计思想，模型主要包括训练文本统计模块、特征选择模块、性能评估模块、数据库接入模块和人机交互模块。

3.3 设计思路

1. 训练文本统计模块的设计
   训练文本统计模块的主要负责根据选择的训练文本集合建立特征项典，针对IDF 进行统计计算并保存计算结果于数据库中。首先清空数据库中的相关表，然后根据训练文本的特征项构造特征项表，最后根据训练文本提供的信息进行统计量的计算并保存结果。
2. 特征选择模块的设计
   特征选择模块主要的作用是降维，该模块按照不同的住值参数C进行特征选择。特征选择的具体步骤为:
   步骤1．选择参数;
   步骤2．检查输入阈值，是否符合标准(数字或字符);
   步骤3．根据特征选择方法(互信息差值)对特征进行计算、融合、排序，然后保存符合的特征;
   步骤4．标识类别中符合的特征。
3. 分类算法模块的设计
   分类算法模块提供三个分类算法:朴素贝叶斯、KNN和改进贝叶斯。针对不同的文本表示模型的类型，系统会根据不同文本表示模型使用不同的文本分类算法实现。分类算法

模块的功能如下:
①用户根据需要选择分类算法
②通过算法实现文本自动分类
③提供多种分类算法供比较测试

4. 性能评估模块的设计
   性能评估模块负责计算分类器的性能指标。本系统采用的性能评估函数为查准率、查全率、F测值。主要完成以下功能:
   ①根据分类结果更新性能评估参数 TP、FP、FN和TN;
   ②计算各个类别的查准率、查全率、F测值;
   ③分类算法完成文本的自动分类后，性能评估模块显示性能评估结果。
5. 数据库连接模块的设计
   数据库连接模块负责完成业务处理模块与数据库的交互。当系统需要对数据库进行读写操作的时候，就调用该模块。数据库连接模块主要完成如下的功能:
   ①打开与关闭数据库连接;
   ②数据库的查询与修改;
   ③调用数据库端存储过程;
   ④对数据库修改等操作。
   数据库连接模块封装了对数据库的连接和查询、修改等常用操作。对查询修改实现参数化控制，操作数据库时只需调用相应查询或修改方法并提供所需参数即可，这样在不了解SQL语句的情况下仍然可以实现对数据库的操作，因此该模块既方便用户操作，同时也减少写SQL语法错误的概率。
6. 人机交互模块的设计
   人机交互模块一方面用户向系统提供信息提出任务要求;另一方面系统向用户提供分析结果，以及出错提示信息等。进入操作界面后，首先确定是否重新设定训练文本集与测试文本集，如果是,那么在设定后重新计算特征统计量并根据训练文本集合让分类器重新学习。如果否，则进入文本分类技术选择部分，即特征选择部分，选择完毕后在后台进行文本自动分类，然后将分类结果和性能评估结果以表格的形式显示给用户。

3.4 结论观点

我们经过资料搜集、方法讨论，最终提出使用朴素贝叶斯模型的思想进行文本分类。

3.4.1 贝叶斯公式

贝叶斯公式：

> $P(Y|X)=frac{P(X|Y)P(Y)}{P(X)}$

它由联合概率公式推导得出：

> $P(Y,X) = P(Y|X)P(X)=P(X|Y)P(Y)$

其中P(Y)叫做先验概率，P(Y|X)叫做后验概率，P(Y,X)叫做联合概率。
在机器学习的视角下，我们把X理解成“具有某个特征”，把Y理解成“类别标签”。在最简单的二分类问题(判断是与否)下，我们将Y理解成“属于某类”。于是贝叶斯公式就变形成了下面的样子:

> *P(“属于某类”|“具有某特征”)=frac{P(“具有某特征”|“属于某类”)P(“属于某类”)}{P(“具有某特征”)}*

3.4.2 朴素贝叶斯方法

加上条件独立假设的贝叶斯方法就是朴素贝叶斯方法（Naive Bayes）。

朴素贝叶斯失去了词语之间的顺序信息。这就相当于把所有的词汇扔进到一个袋子里随便搅和，贝叶斯都认为它们一样。因此这种情况也称作词袋子模型(bag of words)。有些独立假设在各个分类之间的分布都是均匀的所以对于似然的相对大小不产生影响；即便不是如此，也有很大的可能性各个独立假设所产生的消极影响或积极影响互相抵消，最终导致结果受到的影响不大。

3.4.2 朴素贝叶斯方法进行文本分类的思路

先区分好训练集与测试集，对文本集合进行分词、去除标点符号等特征预处理的操作，然后使用条件独立假设，将原概率转换成词概率乘积，再进行后续的处理。

一句话总结：贝叶斯公式 + 条件独立假设 = 朴素贝叶斯方法。

4 项目分工与初步进展

4.1 项目分工

- 确定选题与关于算法改进的讨论：所有小组成员共同完成
- 文献查找与资料整理：苏波、张原溥
- 算法分析与代码调试：熊康
- 报告撰写与修改完善：刘佳明

4.2 初步进展

我们搜索自然语言理解、ALANA智能机器人的相关论文、百科、博客，了解了NLP、NLU的发展历史以及未来发展趋势，大致了解了自然语言处理的基本原理，并准备复现文本分类算法来深入学习自然语言理解，并提出了改进的想法：

首先对训练文本集中的每篇文本提取出原始的特征词,经过去除停用词、词义消歧的处理后,在类的内部利用信息差值来表达特征项之间的相关性,对相关性高的特征采取适当融合的方法来对特征向量进行局部降维。得出的向量与降维前相比,低频特征词的数目大为减少,高频特征词数目增多,且高频特征词的频度得到加强,特征词总的数目减少,向量的维数降低,对于所属类别具有更强的关联性和较好的表示效力,很好地达到了降维的目的 。 运用局部降维的思想提出了一种用互信息差值来表达特征项之间的相关性,对相关性高的特征采取适当融合来达到向量空间降维的方法 ，从而使文本分类效果得到提升。

5 分析讨论

1. 机器人ALANA真的能听懂吗？

   是，机器人 Alana 可以理解并可以进行十分钟或更长时间的对话。可以像人一样互动，也可以像人一样回复消息；它还是一种社交人工智能，可以与用户聊天，因为它可以连接到互联网，还可以获取新闻；它有自己最喜欢的收藏，如音乐、电影、书籍、体育，并谈论这些相关的事情。

2. NLP的局限性是什么？

   （1）口音。例如一个来自俄罗斯的人说英语，他（她）的口音和单词的发音与正常的英国人不同。NLP 有时很难处理俄语英语口音。然而，来自不同地域的人都有或多或少的口音，这给 NLP 带来了巨大挑战，是未来需要解决的问题之一。

   （2）机器翻译也是 NLP 的局限之一。在两种或更多种自然语言之间进行翻译，将会产生许多涉及歧义、文化差异、褒贬不一、一词多义的问题，这些是NLP需要解决的问题。通过深度学习、改进AI算法、构建语言知识库，有助于提升机器翻译的准确度。

3. 未来有哪些具有挑战性的问题？

   （1）人工智能是一项最新的新兴技术，未来最大的挑战是缺乏专家和训练有素的人才。

   （2）法律问题也是未来需要考虑的问题，如果人工智能收集个人敏感数据，我们需要建立适当的规则来规范这些敏感数据。

   （3）未来还可能会面临收集和利用相关数据的问题。未来，我们需要人工智能的强大计算能力和相关资源，我们必须构建必要的计算资源、利用深度学习等技术来操作大量数据。

4. 未来的正确道路是什么？

   人工智能已经逐渐影响我们生产生活的各个领域。诸如交通、医疗、教育、媒体、客户服务等。人工智能是物联网、大数据、分析等新兴技术的主要技术，使用人工智能，人类将能够用他们的语言相互交流选择。人工智能可以取代繁琐、单一化的人类劳动，并且比人类工作得更快。人工智能还有望接管化工厂、深海以及采矿、太空探索等领域的危险工作。