基于自然语言处理问政平台留言分析

摘要

本文根据网络平台的留言以及热点问题，运用机器学习构建了文本挖掘模型，从而解决了群众留言分类，建立了留言内容的一级标签分类模型；对热点问题进行了挖掘处理；并且从留言的答复相关性、完整性、可解释性等角度制定了一套评价方案。

针对问题一，利用朴素贝叶斯文本分类算法，对文本进行分词，然后去除停用词，再进行分析，由 F-Score 对分类方法做出了评价，得到评价值 F_1=0.8901432913590969；还可以利用支持向量机 SVM 文本分类算法，来进行分析评价，得到$F_2 = 0.8866695614415979$。

针对问题二，首先运用 SIF 加权平均词向量，得出词向量的相似度，然后利用 HDBSCAN 聚类算法，以及移词距离-WMD 来计算文档间的距离，通过文档表达的最小转移代价，衡量文档之间的距离。最后得出相应的热点问题评价结果。

针对问题三，利用余弦相似度算法计算文本答复的评价，首先进行分词分析，再列出所有词，然后对已分词进行编码排列，再利用余弦函数计算两个文本的相似度。

关键词：支持向量机 SVM 文本分类；朴素贝叶斯文本分类；SIF 加权平均词向量；移词距离-WMD；余弦相似度

Analysis of message on political platform based on natural language processing

Abstract

Based on the comments and hot issues on the network platform, this paper constructs a text mining model by means of machine learning, so as to solve the classification of people's comments and establish a level 1 label classification model of message contents. The hot issues are explored and dealt with. And from the comments of relevance, integrity, interpretability and other aspects of a set of evaluation program.

For question one, the Naive Bayes text classification algorithm is used to segment the text, then the stop words are removed, and then the analysis is carried out. The classification method is evaluated by f-score, and the evaluation value F_1= 0.8901432913590969.The SVM text classification algorithm can also be used for analysis and evaluation, F_2=0.8866695614415979.

For question two, SIF weighted average word vector is firstly used to obtain the similarity of word vector, then HDBSCAN clustering algorithm and word shift distance -WMD are used to calculate the distance between documents, and the distance between documents is measured by the minimum transfer cost expressed in documents. Finally, the corresponding evaluation results of hot issues are obtained.

For question three, the cosine similarity algorithm is used to calculate the evaluation of the text response. Firstly, word segmentation analysis is carried out, then all words are listed, then the word segmentation is coded and arranged, and then the cosine function is used to calculate the similarity of the two texts.

Keywords: SVM text classification by support vector machine; Naive Bayesian text classification; SIF weighted average word vector; Shift distance -WMD; Cosine similarity

目录

一、问题的重述

1.1 问题背景

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。请利用自然语言处理和文本挖掘的方法处理一下三个问题。

1.2 要解决得问题

1) 根据网络问政平台的群众留言，按照一定的划分体系对留言进行分类。根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。
2) 根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。
3) 针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量制定了一套评价方案。

二、模型假设

根据网络问政平台的群众留言，本文做出如下假设：

- 假设获得的数据是真实可靠。

三、问题一的分析与求解

3.1 数据的预处理

1) 首先利用 python 的工具包 pandas 删除空白数据和无用数据;

2) 处理文本，首先将文本分词;

3) 去除停用词，再进行分析。

3.2 基于朴素贝叶斯的文本分类算法

问题一要求我们根据网络问政平台的群众留言，按照一定的划分体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。根据 F-Score 对分类方法进行评价:

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{1}$$

P_i为第i类的查准率，R_i为第i类的查全率。

本题目的是根据平台的群众留言，建立关于留言内容的一级标签分类模型，所以本题根据朴素贝叶斯的文本分类算法进行分类研究。

3.2.1 贝叶斯原理

贝叶斯的定理是基于条件概率得出的，基本求解公式为:

$$\mathrm{P(A|B)} = \frac{P(AB)}{P(B)} \tag{2}$$

贝叶斯定理之所以有用，是因为我们在生活中经常遇到这种情况：我们可以很容易直接得出P(A|B)，P(B|A)则很难直接得出，但我们更关心P(B|A)，贝叶斯定理就为我们打通从P(A|B)获得P(B|A)的道路,即得出贝叶斯定理:

$$\mathrm{P(B|A)} = \frac{P(A|B)P(B)}{P(A)} \tag{3}$$

根据贝叶斯算法，计算过程如下:

第一阶段——准备工作阶段，这个阶段的任务是为朴素贝叶斯分类做必要的准备，主要工作是根据具体情况确定特征属性，并对每个特征属性进行适当划分，然后由人工对一部分待分类项进行分类，形成训练样本集合。这一阶段的输入是所有待分类数据，输出是特征属性和训练样本。这一阶段是整个朴素贝叶斯分类中唯一需要人工完成的阶段，其质量对整个过程将有重要影响，分类器的质量很大程度上由特征属性、特征属性划分及训练样本质量决定。

第二阶段——分类器训练阶段，这个阶段的任务就是生成分类器，主要工作是计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估 计，并将结果记录。其输入是特征属性和训练样本，输出是分类器。这一阶段是机械性阶段，根据前面讨论的公式可以由程序自动计算完成。

第三阶段——应用阶段，这个阶段的任务是使用分类器对待分类项进行分类，其输入是分类器和待分类项，输出是待分类项与类别的映射关系。这一阶段也是机械性阶段，由程序完成。

现在开始进入本文的主旨部分：如何将贝叶斯分类器应用到文本分类上来。

在文本分类中，假设我们有一个文档$d \in X$，X是文档向量空间(document space)，和一个固定的类集合$C = \{c_1, c_2, c_3, c_4 \cdots c_j\}$，类别又称为标签。显然，文档向量空间是一个高维度空间。我们把一堆打了标签的文档集合$< d, c >$作为训练样本，$< d, c > \in X * C$.例如：

$< d, c >=$ {Beijing joins the World Trade Organization, China}。

对于这个只有一句话的文档，我们把它归类到{China},即打上China标签。我们期望用某种训练算法，训练出一个函数γ，能够将文档映射到某一个类别

$$\gamma: X \to C \tag{4}$$

这种类型的学习方法叫做有监督学习，因为事先有一个监督者（我们事先给出了一堆打好标签的文档）像个老师一样监督着整个学习过程。

由 F-Score 对分类方法进行评价，当

1) F_1大于 0.8 则评价很好;

2) F_1大于 0.6 小于 0.8 则评价一般;

3) F_1大于 0.4 小于 06 则评价较差。

所以根据贝叶斯的文本分类算法，我们得出:

$$F_1 = 0.8901432913590969 \tag{5}$$

则由 F-Score 得知对分类方法进行评价很好。

3.3 基于支持向量机 SVM 文本分类算法

3.3.1 支持向量机 SVM 原理

支持向量机（SVM）算法被认为是文本分类中效果较为优秀的一种方法，它是一种建立在统计学习理论基础上的机器学习方法。该算法基于结构风险最小化原理，将数据集合压缩到支持向量集合，学习得到分类决策函数。这种技术解决了以往需要无穷大样本数量的问题，它只需要将一定数量的文本通过计算抽象成向量化的训练文本数据，提高了分类的精确率。

支持向量机（SVM）算法是根据有限的样本信息，在模型的复杂性与学习能力之间寻求最佳折中，以求获得最好的推广能力支持向量机算法的主要优点有:

1）专门针对有限样本情况，其目标是得到现有信息下的最优解而不仅仅是样本数量趋于无穷大时的最优值;

（2）算法最终转化为一个二次型寻优问题，理论上得到的是全局最优点，解决了在神经网络方法中无法避免的局部极值问题;

（3）支持向量机算法能同时适用于稠密特征矢量与稀疏特征矢量两种情况，而其他一些文本分类算法不能同时满足两种情况。

（4）支持向量机算法能够找出包含重要分类信息的支持向量，是强有力的增量学习和主动学习工具，在文本分类中具有很大的应用潜力。

3.3.2 文本提取特征

目前，在对文本特征进行提取时，常采用特征独立性假设来简化特征选择的过程，达到计算时间和计算质量之间的折中。一般的方法是根据文本中词汇的特征向量，通过设置特征阀值的办法选择最佳特征作为文本特征子集，建立特征模型。（特征提取前，先分词，去停用词）。

本特征提取有很多方法，其中最常用的方法是通过词频选择特征。先通过词频计算出权重，按权重从大到小排序，然后剔除无用词，这些词通常是与主题无关的，任何类的文章中都有可能大量出现的，比如“的”“是”“在”一类的词，一般在停词表中已定义好，去除这些词以后，有一个新的序列排下来，然后可以按照实际需求选取权重最高的前 8 个，10 个或者更多词汇来代表该文本的核心内容。

综上所述，特征项的提取步骤可以总结为：

（1）对全部训练文档进行分词，由这些词作为向量的维数来表示文本；

（2）统计每一类内文档所有出现的词语及其频率，然后过滤，剔除停用词和单字词；

（3）统计每一类内出现词语的总词频，并取其中的若干个频率最高的词汇作为这一类别的特征词集；

（4）去除每一类别中都出现的词，合并所有类别的特征词集，形成总特征词集。最后所得到的特征词集就是我们用到的特征集合，再用该集合去筛选测试集中的特征。

3.3.3 计算方法

1) TF-IDF 公式来计算词的权值：

$$w_{ik} = \frac{tf_{ik} * \log\left(\frac{N}{n_k} + 0.01\right)}{\sqrt{\sum_{i=k}\left[tf_{ik} * \log\left(\frac{N}{n_k} + 0.01\right)\right]^2}} \tag{6}$$

其中tf_{ik}表示特诊次t_k在文档d_i中现的频率，N 为训练文档总数，n_k为在训练集中出现词t_k的文档数。由 TF-IDF 公式，一批文档中某词出现的频率越高，它的区分度则越小，权值也越低；而在一个文档中，某词出现的频率越高，区分度则越大，权重越大。

2) 归一化处理：

归一化就是要把需要处理的数据经过处理后（通过某种算法）限制在你需要的一定范围

内：

$$\frac{a-min}{max-min}=b \tag{7}$$

公式中 a 为关键词的词频，min 为该词在所有文本中的最小词频，max 为该词在所有文本中的最大词频。这一步就是归一化，当用词频进行比较时，容易发生较大的偏差，归一化能使文本分类更加精确。

由 F-Score 对分类方法进行评价，当

4) F_1大于 0.8 则评价很好；

5) F_1大于 0.6 小于 0.8 则评价一般；

6) F_1大于 0.4 小于 06 则评价较差。

所以根据贝叶斯的文本分类算法，我们得出：

$$F_2 = 0.8866695614415979 \tag{8}$$

则由 F-Score 得知对分类方法进行评价很好。

四、问题二的分析与求解

请根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

4.1 数据预处理

1) 首先利用 python 的 pandas 删除空白数据和无用数据；
2) 用 SIF 加权平均词向量求出距离；
3) 再用 HDBSCAN 进行聚类，聚类完成后用 WMD 相似度过滤数据，过滤之后对数据进行整理，再整理过程中通过聚类后的数量进行排名；
4) 在数量相同的情况下：通过点赞数排名，找出每种聚类的时间范围，再由 python 的中文工具包 FoolNLTK 找出地名和人名，最后将聚类后的所有数据汇总导出表格。

4.2 SIF 加权平均词向量

SIF 是一种基于无监督学习的 Sentence Embedding 方法，其效果超过了目前主流的 Sentence Embedding 方法，计算流程如下：

1. 计算 Word Embedding，通过对无标签的语料库；
2. 用加权词向量来表征一个句子；
3. 用 PCA/SVD 来改善它们。

Word Embedding 已经成为了自然语言处理和信息检索中的基石。最近的研究则主要是 Sentence Embedding，之前已经有很多的研究方法，如词向量的简单组合、CNN、RNN…… 2016 年 Wieting et al 在 PPDB 上对标准的 Word Embedding 进行修改，训练一个 word averaging model，但是若无修改的过程，直接对初始的词向量进行平均操作，效果并不好。本文提出的算法 SIF (smooth inverse frequency)

1. 计算词向量的加权平均值：$\mathrm{weight(w)} = \frac{a}{a+p(w)}$
2. common component removal: remove the projection of the average vectors on there first component

4.3 HDBSCAN 聚类

4.3.1 HDBSCAN 算法原理

- 根据密度/稀疏度对空间进行变换
- 建立距离加权图的最小生成树
- 构造连接组件的簇层次结构
- 根据最小的簇大小压缩簇层次结构
- 从压缩树中提取稳定的簇

4.3.2 变换空间

为了找到簇，我们希望在一片稀疏的噪音海洋中找到密度更高的孤岛。聚类算法核心是单链接聚类，它对噪声非常敏感： 一个位于错误位置的单个噪声数据点可以充当岛屿之间的桥梁，将它们粘合在一起。 显然，我们希望我们的算法对噪声是鲁棒的，所以我们需要找到一种方法，以帮助"降低海平面"之前运行一个单一的连接算法。

我们如何在不进行聚类的情况下描述“海洋”和“陆地”我们只要能够得到一个密度的估计，我们就可以把密度较低的点看作是“海洋”。这里的目标不是完全区分”海洋”和”陆地”，只是为了使我们的簇核心对噪音更加健壮。 因此，鉴于”海洋”的定义，我们希望降低海平面。就实际目的而言，这意味着使”海洋”中的点彼此之间和”陆地”之间的距离更远。

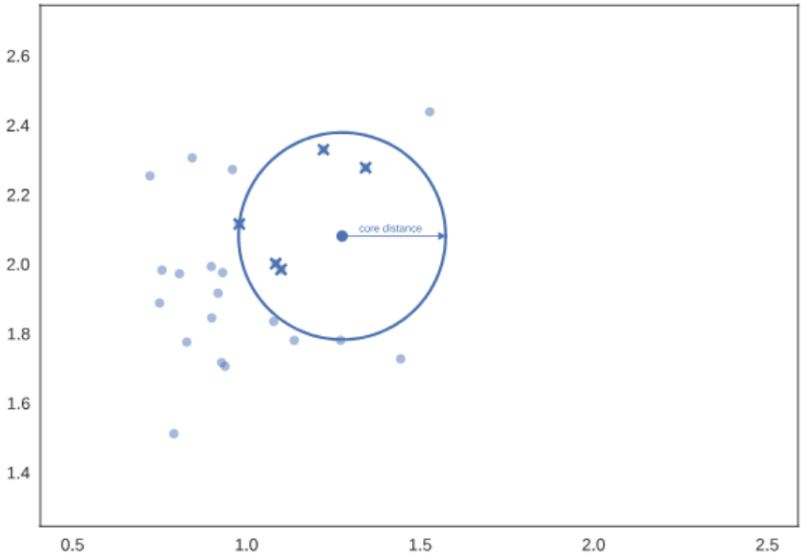
然而，这只是设想。它在实践中是如何工作的？我们需要一个非常低成本的密度估计，最简单的是到 kth 最近邻距离。将其称为为针对点 x 的参数 k 定义的核心距离（定义为当前点到其第 k 近的点的距离），并表示为$core_k(x)$:

$$core_k(x) = d\Big(x, N^k(x)\Big) \tag{9}$$

现在我们需要一种方法，以低密度(相应的高核心距离)分散点。要做到这一点，简单的方法是定义一个新的点之间的距离度量，我们将调用相互可达距离。 我们将相互可达距离定义如下:

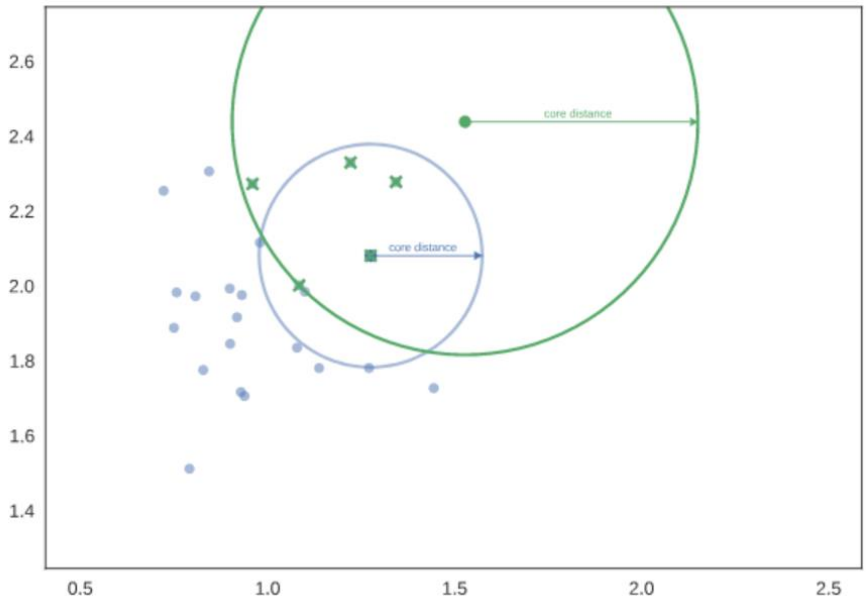
$$d_{mreach-k}(a,b) = max\{core_k(a), core_k(b), distance(a,b)\} \tag{10}$$

式中，d(a,b)是 a 与 b 的原始距离。在该式中密集点（核心距离较低）彼此保持相同的距离，但较稀疏的点被推开，以使其核心距离至少远离任何其他点。 这实际上”降低了海平面”，稀疏的”海洋”指向外界，而”陆地”则没有受到影响。这里需要注意的是，这显然取决于 k 的选择，较大的 k 值将更多的点解释为处于“海洋”中。所有这些用一张图片来说都比较容易理解，我们使用 k 值为 5，然后对于给定的一个点，我们可以画一个核心距离的圆，作为与第六个最近邻接触的圆(包括点本身)，如下所示:

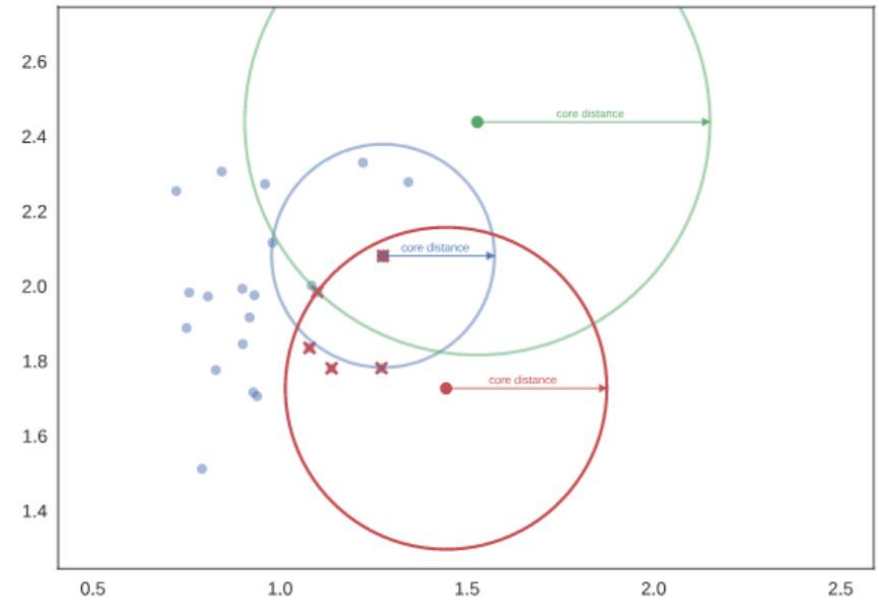


再选择另外一个点，我们可以做同样的事情，这一次用一组不同的邻居（其中一个甚至包含

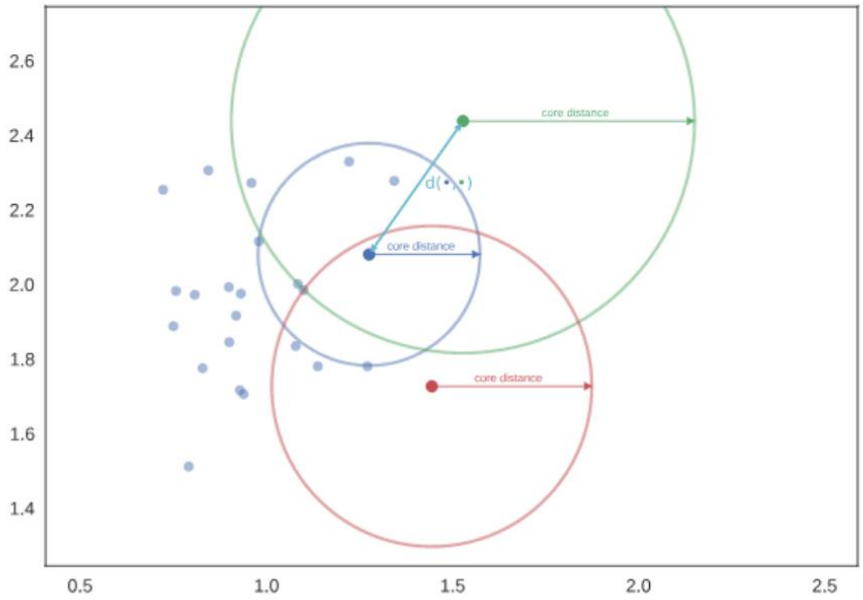
我们选择的第一个点)：



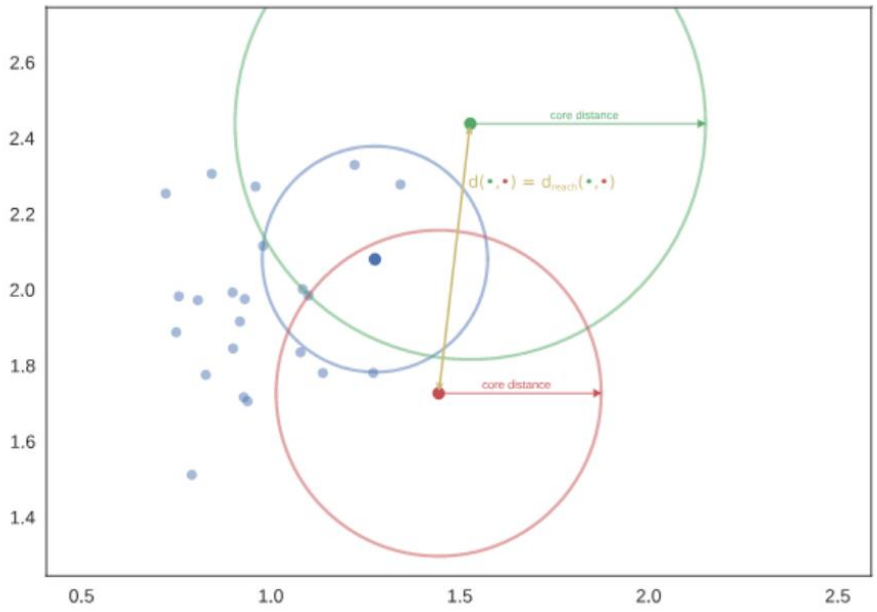
我们可以再用另一组六个最近邻，和另一个半径略有不同的圆：



现在，如果我们想知道蓝点和绿点之间的相互可达距离，我们可以先画一个箭头，给出绿点和蓝点之间的距离：



它穿过蓝色的圆圈，但不是绿色的圆圈——绿色的核心距离大于蓝色和绿色之间的距离。因此，我们需要将蓝色和绿色之间的相互可达距离标记为大于等于绿色圆的半径。另外，从红色到绿色的相互反应距离就是从红色到绿色的距离，因为这个距离大于两个核心距离：



一般来说，有潜在的理论来证明，相互可达距离作为一种变换，可以很好地允许单链接聚类更接近水平集的层次结构，无论我们采样的点的实际密度分布是什么。

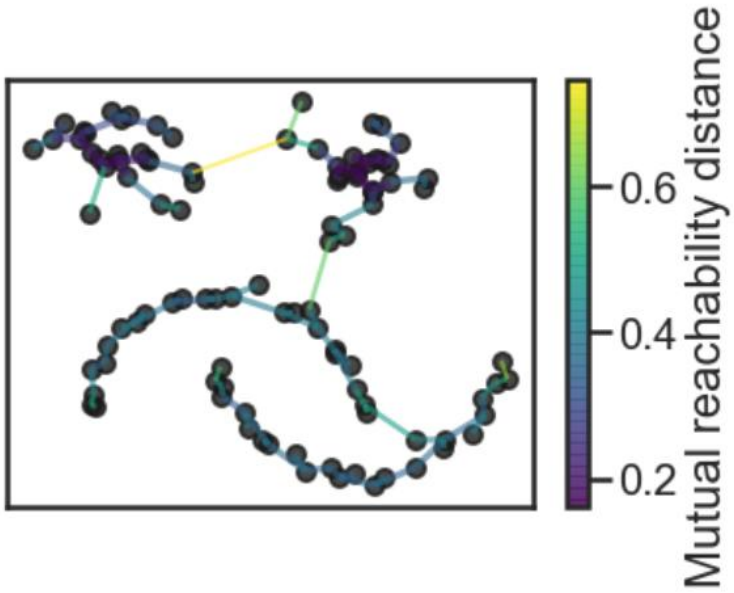
4.3.3 建立最小生成树

现在我们在数据上有了一个新的相互可达性度量，我们希望开始在稠密数据上寻找孤岛。当然，密集区域是相对的，不同的岛屿可能有不同的密度。 从概念上讲，我们将要做的是：

将数据看作一个加权图，其中数据点为顶点，任意两点之间的边的权重等于这些点之间的相互可达距离。

现在考虑一个阈值，从高开始，逐步降低。 删除任何重量超过该阈值的边。 当我们删除边时，我们将开始断开图形的连接组件。 最终，我们将在不同的阈值水平上得到一个连接组件的层次结构(从完全连接到完全不连接)。在实践中，这是非常低效的：我们有 n2 个边，并且不期望连接的组件算法运算那么多次。正确的做法是找到一个最小的边集合，这样从集合中删除任何边都会导致组件断开。幸运的是，图论为我们提供了这样一个东西:的最小生成树。

我们可以通过 Prim 算法非常有效地构建最小生成树树-我们一次构建一条边，总是添加最小的权重边，将当前的树连接到树中还没有的顶点。您可以看到下面构造的 HDBSCAN 树。注意这是相互可达距离的最小生成树，它不同于图中的纯距离。在这个例子中，k 值为 5。



4.3.4 提取簇

直观地说，我们希望选择的簇能够持续存在并且有更长的生命周期；短命的簇可能仅仅是单链接方法的产物。在前面的图中，我们可以说，我们要选择那些簇有最大面积的情节油墨。 为了创建一个平面集群，我们需要添加一个进一步的要求，如果您选择了一个簇，那么您就不能选择它的后代的任何簇。事实上，关于应该做什么的直观概念正是 HDBSCAN 所

做的。当然，我们需要把事情形式化，使之成为一个具体的算法。

首先，我们需要一种不同于距离的度量方法来考虑簇的持久性；作为替换我们使用λ = $\frac{1}{distance}$。针对给定的簇，我们可以定义值λ_{birth}和λ_{death}为当簇分离并成为它自己的簇时的λ值，以及当簇分别拆分为较小的簇时的λ值（如果有）。反过来，对于给定的集群，对于集群中的每个点 P，我们可以将值λ_p定义为“从簇中掉出来”的λ值，该值介于λ_{birth}和λ_{death}之间。现在，对于每个簇，稳定性计算如下：

$$s_{cluster} = \sum_{p \in cluster} \left(\lambda_p - \lambda_{birth}\right) \tag{11}$$

其中：

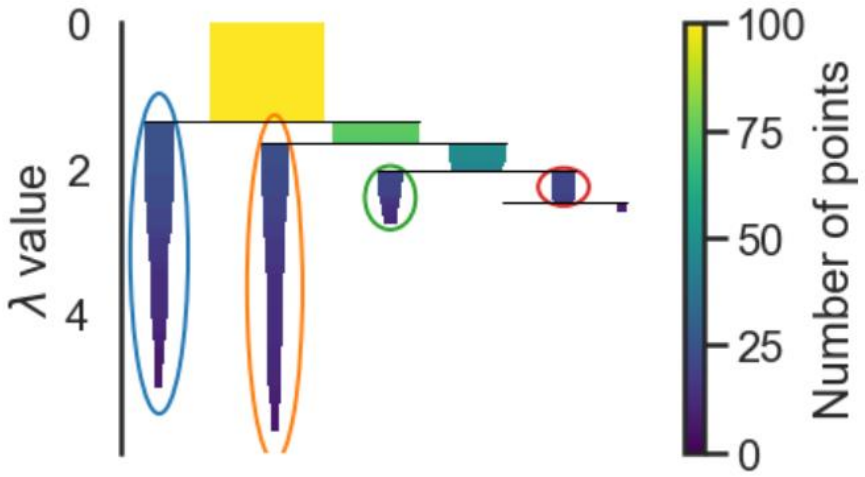
λ_{birth}：团簇形成时的 λ 值；

λ_{death}：团簇分裂为两个子团簇时的 λ 值；

λ_p：从团簇中分离出去时的 λ 值。

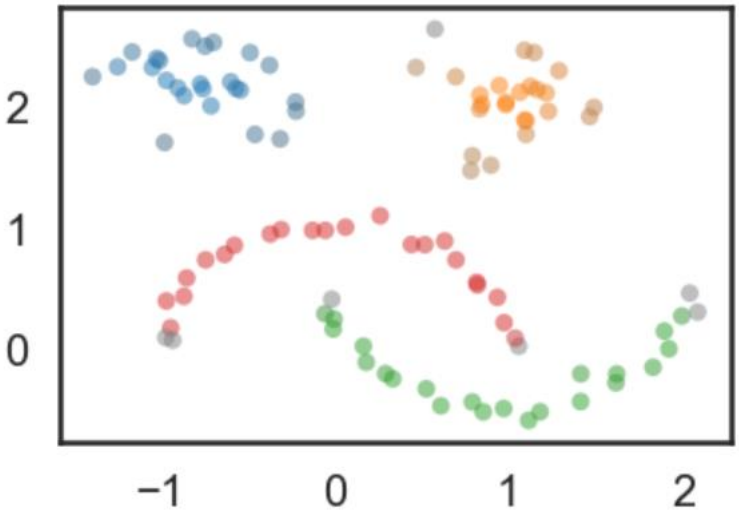
将所有叶节点声明为选定的簇。现在通过遍历树（反向拓扑排序顺序）。如果子簇的稳定性之和大于簇的稳定性，那么我们将簇稳定性设置为子簇稳定性之和。另一方面，如果簇的稳定性大于其子簇的总和，那么我们将簇声明为选定簇，并取消选择其所有子簇。当我们到达根节点时，我们将当前选定的簇集合称为平面簇并返回它。

好的，这很复杂,但它实际上只是在执行我们的“选择图片中最大面积的簇”，这要受我们前面解释的后代约束。我们前面解释过。我们可以通过这个算法选择压缩树图中的聚类，你会得到你想要的结果：



既然我们有了簇，那么根据 sklearn api 将其转化为簇标签就足够简单了。任何不在所

选簇中的点只是一个噪声点(并被分配为-1)。不过，我们可以做得更多：对于每个簇，我们都有该簇中每个点 P 的λ_p；如果我们简单地规范化这些值（因此它们的范围从 0 到 1），那么我们就可以度量集群中每个点的簇成员资格的强度。HDBSCAN 库将此作为 cluster 对象的概率属性返回。因此，有了标签和成员强度，我们就可以制作标准图，基于簇标签为点选择颜色，并根据成员强度对颜色进行去饱和（并使未聚集的点纯灰色）。



这就是 HDBSCAN 的工作原理。这可能看起来有点复杂，但最终每个部分实际上是非常简单的，可以很好地优化。

4.4 移词距离-WMD

WMD 是计算文档间距离的算法，它通过文档表达的最小转移代价衡量文档之间的距离。这其中词间转移代价和权重转移量是文档表达转移的关键。词间转移代价用 Word2vec 嵌入空间的欧氏距离度量，由此引入词间语义相似度。每个词可转移的总权重转移量为词的权重，由此引入词在文档中的贡献信息。算法中权重用词频度量。

令D和D′为分别有m和n个词的两个文本文档的标准词袋向量表达，文档D和D′之间的距离定义如下：

$$\text{distance}(\mathrm{D},\mathrm{D}') = min_{T\geq 0}\sum_{i=1}^{m}\sum_{j=1}^{n} T_{ij}c(i,j) \tag{12}$$

$$\sum_{j=1}^{n} T_{ij} = d_i \forall i \in \{1,2,\cdots m\} \tag{13}$$

$$\sum_{i=1}^{m} T_{ij} = d_j' \forall j \in \{1,2,\cdots n\} \tag{14}$$

其中T为流量矩阵，T_{ij}表示文档D中第i个词转移到文档D′中第j个词的权重转移量。为了完全转移D到D′，从文档 D 中第i个词转移出的所有权重转移量必须等于它在 D 中的权重：词频di。同理，转入文档 D'中第j个词的所有权重转移量必须等于d_j' $c(i,j)$为词间转移代价，即文档D中第i个词转移到文档D′中第j个词之间转移需要的代价。

算法通过最小化文档表达D转移到文档表达D′的累积代价得到文档之间的距离。最小化文档表达转移代价是一个双向的过程，即要求D到D′的转移代价最小，又要求D′到D的转移代价最小。

本文的词间转移代价用 1 减去归一化 Cosine 相似度计算。

则由上述的算法，且根据附件 3，对某一时段内反映特定地点或特定人群问题的留言进行归类，定义了合理的热度评价指标，给出评价结果，如下图所示：

index	热度指数	时间范围	地点/人群	问题描述	热度排名	问题ID
0	26	2018-11-15至2019-12-24	A市	人才购房补贴申请是否与单位注册地有关?	1	1
1	18	2019-01-07至2019-11-20	西地省	A市蜂投网涉嫌诈骗	2	2
2	16	2019-01-08至2019-12-29	A6区丁字中学	周末补课收费	3	3
3	13	2019-02-10至2019-10-08	A市	对公交线路的建议	4	4
4	12	2019-01-15至2019-12-15	A7县星湖湾	反映（洋房二期）小区违规建设医院问题	5	5
5	11	2019-01-07至2019-09-29	A7县	捂盘问题何时解决?	6	6
6	11	2019-01-23至2019-12-13	A3区云栖路云栖谷小区	辅道好多乱停车的	7	7
7	11	2017-06-08至2019-11-27	A市经济学院	寒假过年期间组织学生去工厂工作	8	8
8	11	2019-07-07至2019-08-24	伊景园滨河苑	关于捆绑销售车位的维权投诉	9	9
9	11	2019-03-15至2019-12-30	A市麓谷小镇	投诉物业停车费收取不合理	10	10
10	10	2019-04-28至2019-11-26	A市北辰三角洲E5区	因房屋品质问题导致外墙渗水严重	11	11
11	10	2019-01-06至2019-09-12	西湖街道茶场村五组	A3区什么时候能启动征地拆迁	12	12
12	10	2019-02-14至2019-09-15	A7县星沙四区凉塘路旧城	改造要待何时	13	13
13	10	2019-01-08至2019-11-01	A市	请加快国家区域医疗中心建设	14	14
14	9	2019-01-09至2020-01-03	枫华府邸小区	A3区停水7天了！	15	15
15	9	2019-03-06至2020-01-05	A市	咨询转业士官异地安置问题	16	16
16	8	2019-01-06至2019-05-22	A1区辉煌国际城	.期居民楼下商铺没有配套专用烟道，非法开饭	17	17
17	8	2019-01-12至2019-12-31	A7县	:民自留地征收标准及村组集体用地征收费用划	18	18
18	8	2019-04-17至2019-12-26	A市丽发新城	违建搅拌站，彻夜施工扰民污染环境	19	19

按照表 2 的格式，得出了相应热点问题对应的留言信息，如下图所示：

留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	留言时间_new	热度排名
225657	A00051791	关于A市人才购房补贴的疑问	2019/2/25 14:43:15	，政策鼓励其在A市安家发展，而对于先	2	6	19-02-25 14:43:	1
189180	A000106515	A市人才购房补贴申请是否与单位注册地有关？	2019/6/18 9:51:36	现在长工作，单位出具的在职证明也写明	0	3	19-06-18 09:51:	1
265577	A0004787	咨询A市人才购房补贴通知问题	2019/12/2 11:57:49	并在A市域内工作，年龄35岁以下全日制	0	1	19-12-02 11:57:	1
282104	A00090921	于高级技师申报A市人才新政购房补贴的相关问题咨	2019/7/29 10:42:38	职工社会保险参保材料；是否要求社保必	0	1	19-07-29 10:42:	1
199435	A00024070	咨询A市人才购房补助发放问题	2019/7/2 13:11:31	产，计划7月4号领证，为了确定在补助发	0	0	19-07-02 13:11:	1
203760	A000111954	咨询A市人才购房补贴政策	2019/6/25 15:43:23	义购买了社保，社保信息齐全，资料齐全	0	0	19-06-25 15:43:	1
219026	A00083974	在A市购买二手房能享受人才新政购房补贴吗？	2019/9/27 9:17:14	所政落户A市，想买套二手房，请问购买二	0	0	19-09-27 09:17:	1
220015	A00073641	咨询A市贷款购房与购房资格的相关问题	2019/8/9 10:03:41	别对待，公路部门就要遭到驱赶？总公司	0	0	19-08-09 10:03:	1
221141	A000103845	咨询A市缴纳社保及购房等问题	2019/4/8 1:12:00	上保。 在2017年8月~2018年8月 从公司	0	0	19-04-08 01:12:	1
221153	A00013310	关于在A市工作想买房社保不能补交的咨询	2019/1/16 8:30:59	我就没管可是就是当月没扣2018年12月份	0	0	19-01-16 08:30:	1
224042	A00014225	咨询A市人才购房及购房补贴实施办法等相关问题	2019/1/16 11:58:48	，因为换了工作单位，2018年08月27日在	0	0	19-01-16 11:58:	1
229963	A00010465	咨询A市人才新政住房补贴申请的问题	2019/4/24 23:19:34	和附三院签署的是培训合同3年（也就是	0	0	19-04-24 23:19:	1
232575	A0001717	咨询A市购房资格审核问题	2019/6/27 11:51:34	套房能够安家在A市已经倾尽所有家当，	0	0	19-06-27 11:51:	1
236378	A0007118	询问A市住房公积金贷款的相关问题	2019/3/22 16:44:55	行吗？需不需A市户口，本人A市商贸旅游	0	0	19-03-22 16:44:	1
244951	A00026039	反映A市人才租房购房补贴问题	2019/7/7 19:12:32	位和高校工作就没有补贴了呢？难道这些	0	0	19-07-07 19:12:	1
249463	A00049304	咨询A市社保缴纳问题	2019/7/19 12:25:21	证件？2、后续想自助在手机银行缴费，	0	0	19-07-19 12:25:	1
253923	A00019917	咨询A市住房公积金贷款问题	2019/7/26 20:01:09	还了给他们。我和我父母户籍也没有一起	0	0	19-07-26 20:01:	1

五、问题三的分析与求解

5.1 数据预处理

1) 利用 python 的工具包 pandas 删除空白数据和无用数据；
2) 向量对齐，由于在实际应用中，表征文本特征的两个向量的长度是不同的，因此必然需要对上述向量进行处理；
3) 对文本进行预处理：去停用词（分词，介词，代词等）以及非文本符号；
4) 归并向量，并根据原向量是否在新向量（归并后的向量）存在，若存在则以该词汇的词频来表征，若不存在则该节点置为 0。

5.2 基于余弦相似度算法计算文本答复评价

题目要求：针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。所以我们利用余弦相似度算法来计算文本的相似度，从而得知文本答复意见的评价。

5.2.1 余弦相似度算法基本原理

向量$\vec{a}$用坐标(x_1, y_1)表示，向量$\vec{b}$用坐标(x_2, y_2)表示，向量$\vec{a}$和向量$\vec{b}$在直角坐标中的长

度为$a=\sqrt{x_1^2+y_1^2}, b=\sqrt{x_2^2+y_2^2}$向量$\vec{a}$和向量$\vec{b}$之间的距离我们用向量$\vec{c}$表示，那么向量$\vec{c}$在直角坐标系中的长度为，$c=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$,将$\vec{a}$，$\vec{b}$，$\vec{c}$带入三角函数的公式中得到如下的公式:

$$\cos(\theta)=\frac{a^2+b^2-c^2}{2ab}=\frac{x_1^2+y_1^2+x_2^2+y_2^2-(x_2-x_1)^2-(y_2-y_1)^2}{2\sqrt{x_1^2+y_1^2}*\sqrt{x_2^2+y_2^2}}=\frac{x_1*x_2+y_1*y_2}{\sqrt{x_1^2+y_1^2}*\sqrt{x_2^2+y_2^2}}$$

这是二维空间中余弦函数的公式，则多维空间余弦函数的公式为:

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}}=$$

$$\frac{a\cdot b}{||a||\times||b||} \tag{15}$$

- 当两个向量夹角越大，距离越远，最大距离就是两个向量夹角 180°;
- 夹角越小，距离越近，最小距离就是两个向量夹角 0°，完全重合

即相似度越小，距离越大；相似度越大，距离越小。

知道了向量的夹角余弦相似度计算方法，则只需要将文本变成向量就可以了。

文本是由词组成的，我们一般通过计算词频来构造文本向量——词频向量。

5.2.2 余弦相似度算法:

一个向量空间中两个向量夹角间的余弦值作为衡量两个个体之间差异的大小，余弦值接近 1，夹角趋于 0，表明两个向量越相似，余弦值接近于 0，夹角趋于 90 度，表明两个向量越不相似。

下面我们介绍使用余弦相似度计算两段文本的相似度。

思路:

1) 分词;
2) 列出所有词;
3) 分词编码;
4) 词频向量化;
5) 套用余弦函数计量两个句子的相似度。

句子 A：这只皮靴号码大了。那只号码合适。

句子 B：这只皮靴号码不小，那只更合适。

1、分词：

使用分词对上面两个句子分词后，分别得到两个列表：

listA=[‘这‘,‘只‘,‘皮靴‘,‘号码‘,‘大‘,‘了‘,‘那‘,‘只‘,‘号码‘,‘合适‘]

listB=[‘这‘,‘只‘,‘皮靴‘,‘号码‘,‘不小‘,‘那‘,‘只‘,‘更合‘,‘合适‘]

2、列出所有词，将 listA 和 listB 放在一个 set 中，得到：

set={'不小', '了', '合适', '那', '只', '皮靴', '更合', '号码', '这', '大'}

将上述 set 转换为 dict，key 为 set 中的词，value 为 set 中词出现的位置，即‘这’:1 这样的形式。

dict1={'不小': 0, '了': 1, '合适': 2, '那': 3, '只': 4, '皮靴': 5, '更合': 6, '号码': 7, '这': 8, '大': 9}，可以看出“不小”这个词在 set 中排第 1，下标为 0。

3、将 listA 和 listB 进行编码，将每个字转换为出现在 set 中的位置，转换后为：

listAcode = [8, 4, 5, 7, 9, 1, 3, 4, 7, 2]

listBcode = [8, 4, 5, 7, 0, 3, 4, 6, 2]

我们来分析 listAcode，结合 dict1，可以看到 8 对应的字是“这”，4 对应的字是“只”，9 对应的字是“大”，就是句子 A 和句子 B 转换为用数字来表示。

4、对 listAcode 和 listBcode 进行 oneHot 编码，就是计算每个分词出现的次数。oneHot 编号后得到的结果如下：

listAcodeOneHot = [0, 1, 1, 1, 2, 1, 0, 2, 1, 1]

listBcodeOneHot = [1, 0, 1, 1, 2, 1, 1, 1, 1, 0]

5、得出两个句子的词频向量之后，就变成了计算两个向量之间夹角的余弦值，值越大相似度越高。

listAcodeOneHot = [0, 1, 1, 1, 2, 1, 0, 2, 1, 1]

listBcodeOneHot = [1, 0, 1, 1, 2, 1, 1, 1, 1, 0]

则得出

$$\cos(\theta) = \frac{10}{\sqrt{14} * \sqrt{11}} = 0.81 \qquad (16)$$

根据余弦相似度，句子 A 和句子 B 相似度很高

则根据附件 4 对留言的答复意见的评价，我们有：

1) 当cos(θ)的值大于 0.8 则答复的很好。

2) 当cos(θ)的值大于 0.6 小于 0.8 则答复的一般。

3) 当cos(θ)的值大于 0.4 小于 06 则答复的较差。

则根据余弦相似度算法，得部分文本答复意见评价，如下图:

留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	Cos	答复效果
A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	公司却以交20万保证金，车管理费，在业主大会结束后		19/5/10 14:56:5	0.647643984	答复一般
A00023583	A3区潇楚南路洋湖段怎么还没修好？	2019/4/24 16:03:40	的生意带来很大影响，换填，且换填后还有三趟雨		019/5/9 9:49:1	0.831652539	答复很好
A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	时更是加大了教师的工伹聘任教职工要依法签订劳动		019/5/9 9:49:1	0.874974578	答复很好
A000110735	在A市买公寓能享受人才新政购房补贴吗？	2019/4/24 15:07:30	户A市，想买套公寓，请周岁以下（含），首次购房后		019/5/9 9:49:4	0.84725215	答复很好
A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	马坡岭小学”，原“马坊坡岭”的问题。公交站点的		019/5/9 9:51:3	0.870602298	答复很好
A00077538	A3区含浦镇马路卫生很差	2019-04-08 08:37:20	巴泥巴冲到右边，越是上面中没有说明卫生较差的具体		019/5/9 10:02:0	0.857364279	答复很好
A000100804	A3区教师村小区盼望早日安装电梯	2019/3/29 11:53:23	老社区惠民装电梯的规人民政府办公室下发了《关		019/5/9 10:18:5	0.926126409	答复很好
UU00812	反映A5区东澜湾社区居民的集体民生诉求	2018/12/31 22:21:59	远，天寒地冻的跑好远，准备及设施设备采购等工作。		19/1/29 10:53:0	0.7158392	答复一般
UU008792	反映A市美麓阳光住宅楼无故停工以及质量问题	2018/12/31 9:55:00	得到相关准确开工信息分户检查后，西地省楚江新		19/1/16 15:29:4	0.673819803	答复一般
UU008687	A市洋湖新城和顺路洋湖壹号小区路段公共绿化带的	2018/12/31 9:45:59	桥等地方做立体绿化，按规划要求完成了建设，其中		19/1/16 15:31:0	0.888482589	答复很好
UU0082204	反映A2区大托街道大托新村违建问题	2018/12/30 22:30:30	划局审批通过《温室养殖一笔耕地征收补偿款给原大托		19/3/11 16:06:1	0.475597186	答复较差
UU008829	A5区鄱阳村D区安置房人防工程的咨询	2018/12/29 23:27:51	安置房地下室近两万平米人防发[2014]7号文件要求		19/1/29 10:52:0	0.841889746	答复很好
UU00877	4区万国城小区段请求修建一座人行天桥或者地下通	2018/12/29 11:55:34	大量从小区开车出去的合进行具体选址，招标（邀标		19/1/14 14:34:5	0.93410136	答复很好
UU0081480	举报A市芒果金融平台涉嫌诈骗	2018/12/28 17:18:45	相关政府部门的大力支警情，已由银盆岭派出所立		019/1/3 14:03:0	0	答复差
UU0081227	建议增开A市261路公交车	2018/12/28 7:53:25	以上！天寒地冻，其他于驾驶员工作时间长，劳动		19/1/14 14:33:1	0.759763456	答复一般
UU008444	于A市新开铺路与披塘路交叉路口通行安全问题的建	2018/12/27 15:18:07	：https://baidu.com/。塘路路口两端各拆除20米中		019/3/6 10:26:1	0.716653223	答复一般
UU0081194	投诉A3区桐梓坡路益丰大药房以次充好	2018/12/27 1:55:21	以各种理由拒绝退货，提供的信息进行投诉信息的		019/1/3 14:02:4	0.746088085	答复一般

则由余弦相似度，得出相应的答复意见的评价。

参考文献

[1] T M Mitchell. Machine Learning [M]. New York: McGraw-Hill, 1997.

[2] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2013.

[3] 林晓明. 华泰人工智能系列之三: 人工智能选股之支持向量机模型[R]. 广东: 华泰证券研究所, 2017.

[4] 周渐，基于SVM算法的多因子选股模型实证研究[D]，浙江工商大学硕士论文，2017.

[5] A Geron. 机器学习实战[M]. 北京: 机械工业出版社, 2018.

[6] M Banko, E Brill. Mitigating the Paucity-of-Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for Natural Language Processing [J]. Proceedings of the First International Conference on Human Language Technology Research, 2001: 1-5.

[7] A Halevy, P Norvig, F Pereira. The Unreasonable Effectiveness of Data [J]. IEEE Intelligent Systems, 2009, 24(2): 8-12.

[8] A Geron. 机器学习实战[M]. 北京: 机械工业出版社, 2018.

[9]]Aurelien Geron，机器学习实战：基于Scikit-Learn 和TensorFlow，机械工业出版社，2018.9.