# 基于短文本数据流的分类软件

# 使用说明书V1.0

# 目 录

[基于短文本数据流的分类软件 - 0 -](#_Toc6489057)

[使用说明书V1.0 - 0 -](#_Toc6489058)

[目 录 - 1 -](#_Toc6489059)

[第一章 软件概述 - 2 -](#_Toc6489060)

[1.1 软件简介 - 2 -](#_Toc6489061)

[1.2 功能简介 - 2 -](#_Toc6489062)

[1.2.1 有监督短文本数据流分类 - 2 -](#_Toc6489063)

[1.2.2 半监督短文本数据流分类 - 2 -](#_Toc6489064)

[1.3 设计原理简介 - 2 -](#_Toc6489065)

[第二章 软件组成及运行环境 - 4 -](#_Toc6489066)

[2.1 软件组成 - 4 -](#_Toc6489067)

[2.1.1 有监督短文本数据流分类模块 - 4 -](#_Toc6489068)

[2.1.1.1 文本扩展 - 5 -](#_Toc6489069)

[2.1.1.2 特征表示 - 7 -](#_Toc6489070)

[2.1.1.3 集成模型构建、更新与预测 - 8 -](#_Toc6489071)

[2.1.2 半监督短文本数据流分类模块 - 11 -](#_Toc6489072)

[2.1.2.1特征表示模块 - 11 -](#_Toc6489073)

[2.1.2.2集成模型构建、更新与预测模块 - 12 -](#_Toc6489074)

[2.1.3 图形化界面设计模块 - 16 -](#_Toc6489075)

[2.2 软件运行环境要求 - 17 -](#_Toc6489076)

[第三章 系统功能描述及界面预览 - 18 -](#_Toc6489077)

[第四章 软件使用规程 - 21 -](#_Toc6489078)

[4.1 概述 - 21 -](#_Toc6489079)

[4.2 软件操作流程示例 - 21 -](#_Toc6489080)

[4.2.1 搜索示例1 - 21 -](#_Toc6489081)

[4.2.2 搜索示例2 - 24 -](#_Toc6489082)

第一章 软件概述

1.1 软件简介

随着移动互联网和智能手机的普及，类似于微博、淘宝这样的社交购物平台，每天都会产生大量的短文本数据流。这些短文本数据流蕴含着丰富的商业价值与研究价值，有助于政府、企业决策以及个人学习。短文本数据流分类作为处理短文本数据的基础方法常常被广泛使用，因此，本项目设计并实现了基于短文本数据流分类的软件用以快速处理短文本分类问题。考虑到有标签短文本数据的缺少会导致分类模型效果不佳，本项目在设计一个有监督短文本数据流分类方法的同时也给出了一个半监督短文本数据流分类方法，从而使得在有标签短文本数据缺少的情况下，也能有效分类短文本数据流。

1.2 功能简介

### 1.2.1 有监督短文本数据流分类

用户可以通过输入有标签的短文本数据流文本来分类短文本数据。

### 1.2.2 半监督短文本数据流分类

用户可以通过输入部分有标签的短文本数据流文本来分类短文本数据。

1.3 设计原理简介

本项目的研究内容是由两部分组成：有监督短文本数据流分类和半监督短文本数据流分类。有监督短文本数据流分类是通过将有标签的短文本数据流作为输入，借助外部语料库和文本扩展的相关信息扩展短文本缓解数据的稀疏问题，然后通过online BTM主题模型将扩展后的短文本数据流中的每个短文本表示成一组主题概率，用以解决短文本数据流的高维问题，最后通过构建集成模型来分类短文本数据流，同时检测短文本数据流中的概念漂移来更新集成模型。半监督短文本数据流分类则是通过将部分有标签的短文本数据流作为输入，同样借助外部语料库扩展短文本同时将短文本进行特征向量表示，集成模型的构建是由有标签短文本构建的分类器和无标签短文本构建的聚类模型共同组成，从而能够有效的分类短文本数据流。

第二章 软件组成及运行环境

2.1 软件组成

图2.1给出了短文本数据流分类软件的系统框架示意图，该软件由两部分组成：有监督短文本数据流分类模块和半监督短文本数据流分类模块。



图2.1 系统框架示意图

有监督短文本数据流分类模块首先将有标签的短文本数据流作为输入，接着借助外部语料库扩展输入的短文本数据流，然后将扩展后的短文本数据流进行特征表示，最后构建集成模型并预测下一个到来的短文本数据。半监督短文本数据流分类算法则是将部分有标签的短文本数据流作为输入，经过外部语料库来对输入的短文本数据流进行特征表示，最后同样构建并更新集成模型从而更好的预测下一个到来的短文本数据。

## 2.1.1 有监督短文本数据流分类模块

图2.2给出了有监督短文本数据流分类模块的整体流程图，首先，提取界面输入的参数信息，然后读取短文本数据流，接着进行文本扩展和特征向量化，最后集成模型被构建与更新从而分类短文本。



图2.2 有监督短文本数据流分类模块的整体流程图

下面本文将分为四个部分来介绍有监督短文本数据分类模块，分别为文本扩展、特征表示、集成模型构建与预测以及界面的设计

### 2.1.1.1 文本扩展

**（1）概述**

短文本本身的稀疏性问题导致借助外部语料库扩展变得尤为重要。该模块借助由训练好的LDA主题模型作为外部语料库来扩展输入的短文本数据流。

**（2）主要原理**

由LDA主题模型训练相关文本数据获得模型被视为外部语料库，首先利用该模型推断输入的短文本数据流中每个短文本的主题概率，然后选择概率值高的主题下的词来扩展短文本，其主题下词的扩展倍数由用户提供的区间信息文件决定，选择的主题数和主题下的词个数也由用户自己控制。

**（3）构成**

整个文本扩展模块由两个部分组成：LDA主题模型推断短文本数据流中每个短文本的主题概率和根据主题概率扩展短文本数据流。

图2.3给出了LDA主题模型推断短文本数据流中每个短文本的流程图，这里使用JGibbLDA的Java版本直接读取已经训练好的LDA主题模型，推断短文本数据流中每个短文本的文档-主题概率。



图2.3 LDA推断短文本数据流的主题概率流程图

根据LDA主题模型推断的每个短文本的主题概率扩展短文本，其扩展过程如图2.4所示，首先数据参数准备阶段：1）读取区间信息文件，获得概率区间和对应区间的扩展倍数；2）获取最高可扩展的主题个数和主题下词的个数；3）获取外部语料库（即LDA主题模型）的主题下词的数据；4）获取前一部分得到的短文本数据流中每个短文本的文档-主题概率，然后对每个短文本的文档主题概率进行降序排序从而能够更快的扩展短文本，最后根据概率区间信息、对应区间的扩展倍数、最高可扩展的主题个数以及主题下词的个数，将外部语料库的主题下的词扩展到短文本中从而丰富短文本的语义信息。



图2.5 短文本扩展流程图

表2.1给出了文本扩展所用到的类，其中ShortTextExpansion类为主要扩展类，包括调用LDA类获得推断后短文本数据流的文档-主题概率以及短文本的扩展；ShortTexts类用来存储短文本数据流的基本信息，包括原始短文本数据和扩展后的短文本数据以及后面提到的特征表示后的短文本数据等；LDA类是封装的一个调用JGibbLDA的接口，其余的类都是JGibbLDA中的类。

|  |  |
| --- | --- |
| 表2.1 文本扩展类构成 | |
| **类名称** | **作用** |
| ShortTextExpansion类 | 用于扩展短文本数据流 |
| ShortTexts类 | 用于存储短文本数据流信息 |
| LDA类 | 用于推断短文本数据流的主题概率 |
| Constants类 | JGibbLDA的常量类 |
| Conversion类 | JGibbLDA的中间文件名称设置类 |
| Dictionary类 | JGibbLDA的字典类 |
| Document类 | JGibbLDA的文本存储类 |
| Inferencer类 | JGibbLDA的主题推断类 |
| LDACmdOption类 | JGibbLDA的参数设置类 |
| LDADataset类 | JGibbLDA的数据集类 |
| Model类 | JGibbLDA的模型类 |
| Pair类 | JGibbLDA的工具类 |

### 2.1.1.2 特征表示

**（1）概述**

该模块借助在线BTM主题模型将扩展后的短文本数据流进行特征表示，将每个短文本用一组主题表示。

**（2）主要原理**

借助Java的本地接口（即JNI）将已有的C版本的online BTM算法封装成一个动态链接库从而使得Java程序能够直接调用online BTM程序。

**（3）构成**

扩展后的短文本数据流会被分割成一组数据块，在线BTM主题模型对这个数据块进行特征表示。图2.6给出了一个短文本数据块的特征表示流程图，首先获取online BTM的参数信息和扩展后的短文本数据流，其中参数信息包括由界面输入的两个Dirichlet的先验参数alpha和beta、衰退值、迭代数以及主题数，然后调用由JNI编译获得的online BTM的动态链接库，最后获取生成的短文本数据流的特征表示。



图2.6 特征表示流程图

### 2.1.1.3 集成模型构建、更新与预测

**（1）概述**

假设短文本数据流按照数据块的形式到来的，为了分类短文本数据流，该模块选择一组数据块分别构建SVM分类器作为基模型从而组成集成模型。为了适应短文本数据流，概念漂移检测会被用于判断当前的数据块与构建集成模型的数据块之间是否发生概念漂移，从而根据漂移检测的结果更新集成模型中的基分类器。根据被预测短文本与用于构建集成模型的数据块之间的语义距离作为集成模型中每个基分类器的权值来预测。

**（2）主要原理**

为了预测新到来的短文本数据块，该模块首先选择一组经过文本扩展和特征表示后的短文本数据块分别构建基分类器，然后计算新的数据块与构建基分类器的数据块之间的语义距离来获得每个基分类器的权值，最后根据每个基模型的预测结果乘以对应权值来累加获得该短文本数据块的预测结果。

随着时间推移短文本数据流的文本信息会发生潜在的漂移等现象，即概念漂移。该模块通过检测新的短文本数据块相对于构建基模型的数据块之间是否发生概念漂移来更新集成模型。更新方法为：当新短文本数据块相对于构建基模型的每个数据块均发生概念漂移，且集成模型中基模型个数少于给定个数，则由新数据块构建的分类器就会被加入到集成模型中，如果等于，则用该分类器替换集成模型中最老的基模型。否则，该分类器被用于替换与其语义距离最小的数据块构建的基分类器。

**（3）构成**

集成模型的更新涉及到新的短文本数据块是否发生概念漂移检测，因此，这里我们给出漂移检测方法的流程图，如图2.7所示。首先，将用于构建基模型的每个短文本数据块按照类标签划分簇；接着，计算新的短文本数据块中每个短文本与每个簇之间的语义距离，这里的语义距离计算则是通过累加短文本与簇中每个短文本语义距离之和的均值获得，语义距离为1减去两个短文本之间的余弦相似度；然后，选择新数据块中每个短文本与基模型数据块中簇的语义距离最小值作为该短文本与基模型数据块的语义距离，通过累加新数据块中每个短文本与基模型数据块的语义距离之后的均值获得新数据块与基模型数据块之间的语义距离；最后，根据阈值判断新数据块相对于构建基模型的数据块之间是否概念漂移。

第一个经特征表示后的短文本数据块会直接被用来构建SVM分类器作为集成模型的第一个基分类器，图2.8给出了新短文本数据块到来时集成模型的更新与预测。首先，计算新数据块中每个短文本与用于构建基分类器的数据块之间的权值，该权值用短文本与构建基分类器的数据块之间语义相似性（即1减去短文本与构建基分类器的数据块之间语义距离）乘以新数据块与构建基分类器的数据块之间语义相似性（即1减去新数据块与构建基分类器的数据块之间语义距离）来表示；接着，通过累加每个基模型预测新短文本数据块中短文本的类标签与对应权值的乘积来获得新数据块中每个短文本的类标签结果；然后，根据图2.7检测新数据块与用于集成模型中每个基模型的数据块之间的是否发生概念漂移；最后，根据漂移检测的结果用新短文本数据块构建的SVM分类器更新集成模型，具体更新过程如图2.9所示。



图2.7 漂移检测流程图



图2.8 集成模型更新与预测流程图



图2.9 集成模型的更新

表2.2给出了集成模型构建、更新以及预测的相关类说明。

|  |  |
| --- | --- |
| 表2.2 集成模型构建、更新以及预测类构成 | |
| **类名称** | **作用** |
| EnsembleClassification类 | 用于新数据块的预测和集成模型的更新 |
| DistanceCalculation类 | 用于计算语义距离 |
| svmModel类 | 用于存储SVM分类器模型 |
| svm\_train类 | 用于训练数据获得SVM模型 |
| SVMTools类 | 用于获得SVM相关参数和数据 |

## 2.1.2 半监督短文本数据流分类模块

图2.10给出了半监督短文本数据流分类的整体流程图，首先，提取界面的参数信息，然后读取短文本数据流和标签信息，接着对短文本数据流进行特征表示，最后集成模型被构建用于预测短文本数据流，同时根据被预测的短文本数据流更新集成模型。另外需要说明的是读取后的短文本数据流会直接被划分为一组数据块，每个数据块是否有标签会由输入的标签信息文件决定。



图2.10 半监督短文本数据流分类模块的整体流程图

### 2.1.2.1特征表示模块

**（1）概述**

短文本数据流本身的稀疏性问题会影响最后的分类效果，因此，我们考虑用词向量Word2Vec训练文档数据集获得的所有词的原始词向量集合作为外部语料库。通过拥有丰富语义信息的外部语料库中的词向量来表示短文本数据流中对应的词，扩展了短文本的语义信息，缓解短文本数据的稀疏问题。

**（2）主要原理**

短文本数据流中的每个短文本的特征表示通过借助外部语料库中的词向量来累加对应短文本中每个词的词向量获得。

**（3）构成**

图2.11给出了半监督短文本数据流分类的特征表示方法，首先获取短文本数据块和外部语料库，即原始词向量集合；然后根据外部语料库查找数据块的每个短文本中对应词的词向量；最后通过累加短文本中每个词的词向量获得数据块中每个短文本的特征表示。类Doc2Vector完成了半监督短文本数据流分类的特征表示的主要算法。



图2.11 半监督短文本数据流分类的特征表示流程图

### 2.1.2.2集成模型构建、更新与预测模块

**（1）概述**

该模块对最新的一组有标签短文本数据块构建分类器，无标签短文本数据块构建聚类器，从而联合分类器和聚类器共同构成集成模型来预测新的短文本数据块，这里的分类器采用SVM，聚类器采用Kmeans。由于聚类器预测短文本获得的是聚类簇的ID而非真正的类标签，该模块采用由有标签数据向聚类簇传递标签的方式来为每个聚类模型的聚类簇分配类标签信息。

**（2）主要原理**

为了预测新的短文本数据块，该模块选择一组最新的有标签数据块分别构建SVM分类器，一组最新的无标签数据块分别构建Kmeans聚类模型。Kmeans聚类模型中每个簇的标签信息是通过有标签数据块传递获得，具体方法为：首先根据类标签信息将每个有标签数据块划分簇集合，然后计算聚类模型中聚类簇与所有有标签数据块的每个簇集合的语义相似性，最后累加标签相同的语义相似性并进行归一化操作，最终得到的该聚类簇与对应标签的语义相似性值即为该聚类簇预测短文本为对应标签的概率。集成模型中每个基模型的权值是通过计算新数据块的聚类簇与用于构建集成模型中基分类器和基聚类器的簇集合的语义相似性来获取。另外，新数据块与构建基模型的数据块之间的概念漂移结果会被用来判断该基模型是否参与预测任务。

**（3）构成**

集成模型是由一组最新的有标签数据块构建的SVM分类器和一组最新的无标签数据块构建的Kmeans聚类模型组成。为了使得聚类模型能够预测短文本的标签信息，该模块通过图2.12从有标签数据块向聚类簇传递标签信息，其中簇与簇之间的语义相似性是通过两个簇的半径之和与两个簇的中心距离的比值获得，聚类簇的类标签概率是通过累加标签相同的与簇集合的语义相似性并进行归一化操作获得。



图2.12 类标签传递

图2.13给出了集成模型如何预测新的短文本数据块，该集成模型是选择最新的一组有标签数据块构建分类器和一组无标签数据块构建聚类模型组成，同时聚类模型中每个聚类簇通过图2.12过程获得类标签概率，这里的分类器个数与聚类模型个数由用户自己设定。每个基模型的权值是通过计算新的数据块的聚类簇与基模型数据块的簇之间的语义相似性获得（分类器是按照类标签划分得到簇集合，聚类模型是聚类簇集合）。用集成模型预测新的短文本数块之前会先检测新的数据块相对于构建基模型的数据块之间是否发生概念漂移，如果发生则对应的基模型不参与新数据块的预测。



图2.13 集成模型的构建与预测

为了更新集成模型，如果新的数据块是有标签数据块则直接构建SVM分类器替换原集成模型中最老的基分类器，如果新的数据块是无标签数据块则构建聚类模型替换原集成模型中最老的基聚类模型，同时为聚类模型中每个聚类簇分配类标签信息。

图2.14给出了新的数据块与构建基模型的数据块的概念漂移计算方法，如果基模型是分类模型则对应的簇集合是按照类标签划分得的的，如果是聚类模型则是对应的聚类簇。新数据块中簇与基模型数据块之间的语义距离是通过选择新数据块中簇与基模型的簇之间最小的语义距离来获得，这里的语义距离是两个簇之间的语义相似性的倒数乘以对应基模型簇的短文本数占总的数据块的权值来获得。新的数据块与基模型数据块之间的语义距离时通过累加新数据块中簇与基模型数据块的语义距离的均值获得。



图2.14 概念漂移检测流程图

表2.3给出了该模块的主要类的构成

|  |  |
| --- | --- |
| 表2.3 集成模型构建、更新以及预测类构成 | |
| **类名称** | **作用** |
| SemiEnsembleClassifier类 | 半监督分类器 |
| InstanceConstruction类 | 用于构建聚类模型数据 |
| Classifier类 | 基分类器 |
| Cluster类 | 基聚类模型 |
| ClusterDistanceCalculation类 | 用于计算语义距离 |
| ClusterInfo类 | 存放簇信息 |

## 2.1.3 图形化界面设计模块

**（1）概述**

本项目的图形化界面主要有两个子界面构成：有监督短文本数据流分类子界面和半监督短文本数据流分类子界面。

**（2）主要原理**

使用Java的GUI编程和BeautyEye外观框架。

**（3）构成**

如表2.4给出图形化界面设计模块的主要构成。

|  |  |
| --- | --- |
| 表2.4 图形化界面设计模块的主要构成 | |
| **类名称** | **作用** |
| SSTSCJPanel | 显示有监督短文本数据流分类界面 |
| SemiSSTSCJPanel | 显示半监督短文本数据流分类界面 |

图2.15给出了SSTSCJPanel界面的参数传递和数据显示流程图：



图2.15 SSTSCJPanel流程图

图2.15给出了SemiSSTSCJPanel界面的参数传递和数据显示流程图：



图2.15 SemiSSTSCJPanel流程图

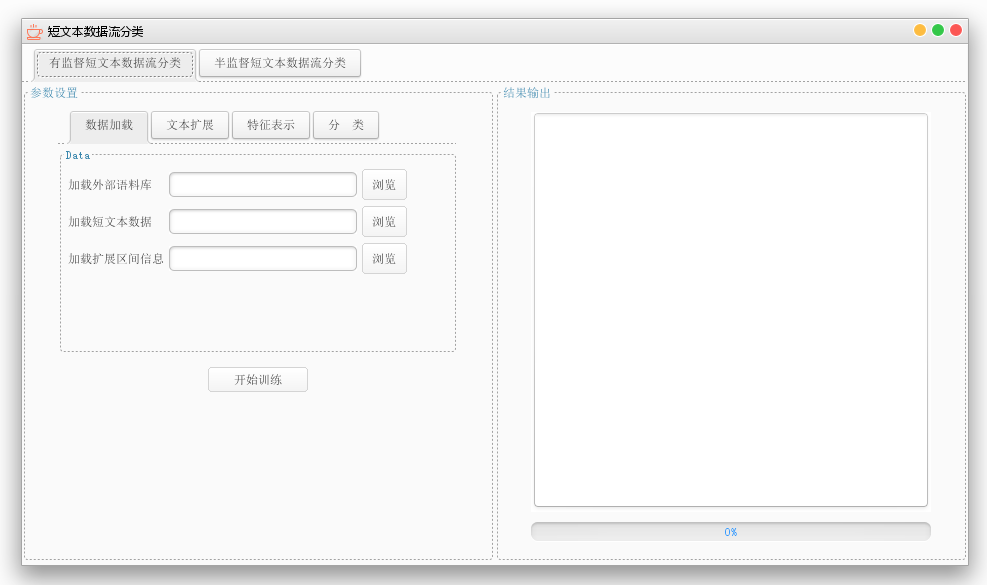
2.2 软件运行环境要求

Java JDK1.8以上，Eclipse Neon。

第三章 系统功能描述及界面预览

本章将描述系统的功能，同时进行界面的展示。

图3.1展示了有监督短文本数据流分类页面，左边为参数设置界面，包括相关文件加载、文本扩展参数、特征表示参数以及最后的分类参数，当参数信息输入完成后，点击开始训练按钮，整个分类过程就开始了。右边为运行有监督短文本数据流分类时相关信息的输出界面，下方是分类的进度展示。



**图3.1 有监督短文本数据流分类页面**

图3.2详细展示了有监督短文本数据流分类页面的参数设置部分，下面将分别描述这四个参数设置部分：

（1）图（a）给出的是相关文件的加载，包括需要分类的短文本数据流文件、外部语料库文件以及文本扩展需要用到的区间信息文件。

（2）图（b）给出文本扩展时需要给定的每条短文本最多可扩展的主题数和每个主题下的单词数，这里为了避免用户输入的主题值和单词值过大导致文本扩展引入太多的噪音，本项目在这里设置了选择框，给定两个参数的选择范围。

（3）图（c）给出特征表示过程中用到的算法online BTM主题模型的相关参数设置。为了在保证主题模型训练的参数合理化，本项目将前三个参数输入为0到1之间的浮点数，迭代数和主题数也提供了选值。

（4）图（d）给出分类的参数设置。为防止用户输入的基分类器个数过高导致有监督短文本数据流分类算法的训练时间太长，本项目将分类器个数设置为小于等于50的整数，本项目采用SVM作为基分类器，后三个参数皆为SVM的相关参数设置。



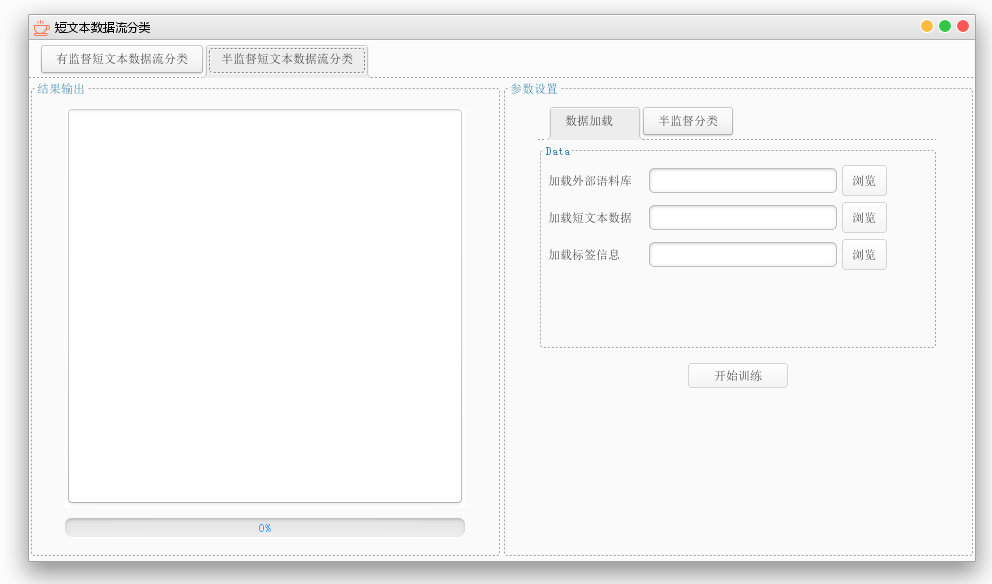
**图（a） 图（b）**



**图（c） 图（d）**

**图3.2 有监督短文本数据流分类的参数设置界面**

图3.3展示了半监督短文本数据流分类页面，与有监督短文本数据流分类界面类似，左边为运行半监督短文本数据流分类时相关信息输出，下面为运行时的进度条，右边为相关参数设置，包括数据加载和分类相关参数信息。



**图3.3 半监督短文本数据流分类页面**

图3.4详细介绍了半监督短文本数据流分类的参数设置部分，下面将分别描述数据加载和半监督分类这两个部分的参数设置：

（1）图（a）给出了相关文件的加载，包括需要分类的短文本数据流文件、外部语料库文件以及用于模拟半监督短文本数据流中哪些短文本有标签，哪些短文本无标签的标签信息文件。

（2）图（b）给出了半监督分类的参数信息，包括分类器个数和聚类器个数设置，以及SVM的相关参数。



**图（a） 图（b）**

**图3.4 半监督短文本数据流分类的参数设置界面**

第四章 软件使用规程

## 概述

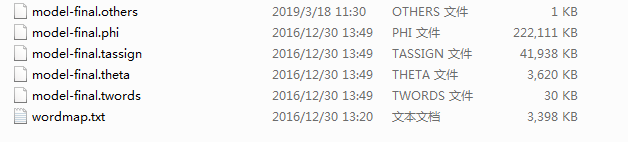
在有监督短文本数据流分类页面/半监督短文本数据流分类页面输入参数信息，点击开始训练按钮会首先检测参数输入格式是否正确，如果正确即可在结果输出部分输出运行时的相关信息以及每一阶段的分类准确率，否则提示错误参数。

## 软件操作流程示例

### 搜索示例1

期望分类有标签的短文本数据流。首先讲述外部语料库文件、短文本数据流文件以及扩展区间文件的格式要求，然后给出操作有监督短文本数据流分类页面的流程示例。

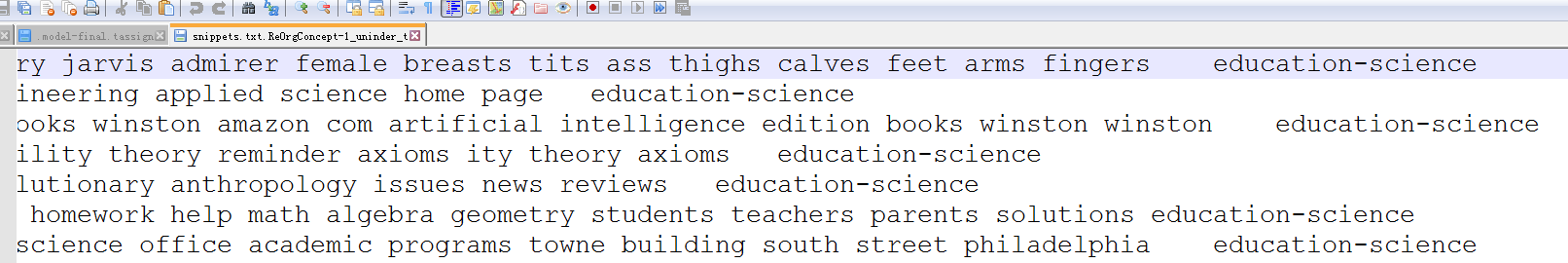
外部语料库是由Java版本的JGibbLDA训练与被分类的短文本数据流具有主题一致性的文本数据集获得的，文件格式如图4.1所示：



**图4.1 有监督短文本数据流分类的外部语料库文件格式**

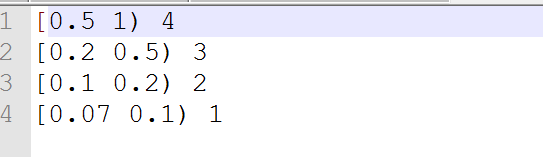
.other文件给出了JGibbLDA训练的参数信息，.phi文件给出了词-主题的多项式分布，.theta文件给出了短文本-主题的多项式分布，.words文件展示了每个主题下的topN个词以及词对应额概率，wordmap.txt文件给出了原始文本数据集的词以及词对应的编号，.tassign文件给出了用编号表示原始文本数据集中每个文本

短文本数据流文件中每一行由短文本和对应标签组成，短文本中每个词用空格分割，短文本与标签之间用制表符分割，如图4.2所示



**图4.2 短文本数据流文件格式**

扩展区间文件格式如图4.3所示，每一行的前一部分表示概率区间，第二部分表示每条短文本在概率区间下的主题扩展的倍数，两部分用空格分割。



**图4.3 区间信息文件格式**

有监督短文本数据流分类流程：

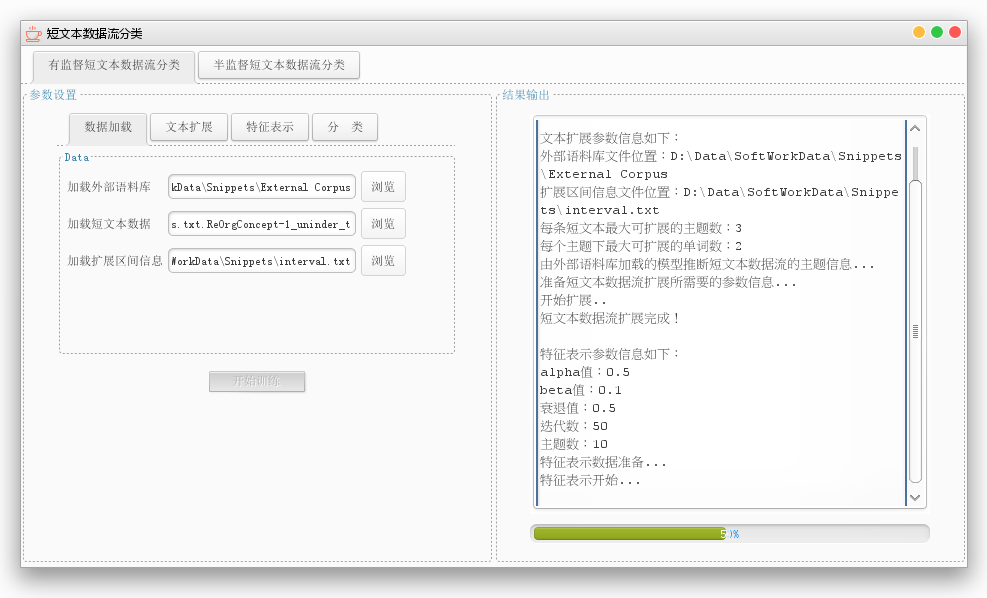
第一步：在有监督短文本数据流分类页面输入参数，如图4.4所示。



**图4.4 有监督短文本数据流分类的参数设置**

第二步：点击“开始训练”按钮，如果输入参数不为空且参数格式正确，则分类短文本数据流，其分类过程中的相关信息输出在右侧的结果输出框中，分类进程有页面的进度条显示。同时，“开始训练”按钮变为灰色。如图4.5所示。否则提示参数输入错误，请求重新输入，如图4.6所示。

第三步：当训练完成后，“开始训练”按钮重新变为可点击状态，表示可重新选择有标签的短文本数据流进行分类。如图4.7所示。

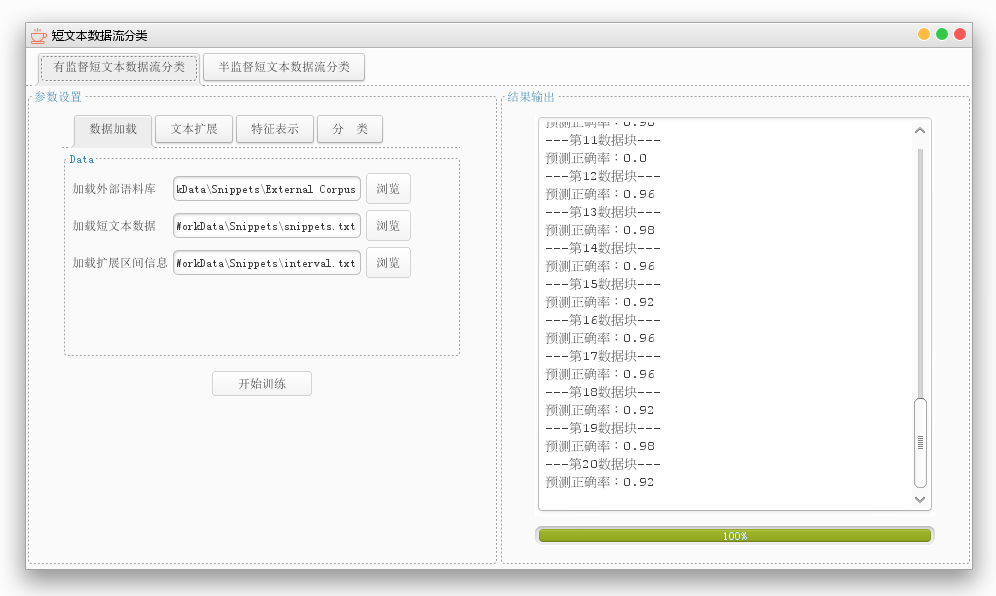


**图4.5 点击按钮后训练并分类短文本数据流**



**图4.6 点击按钮后参数检测错误**

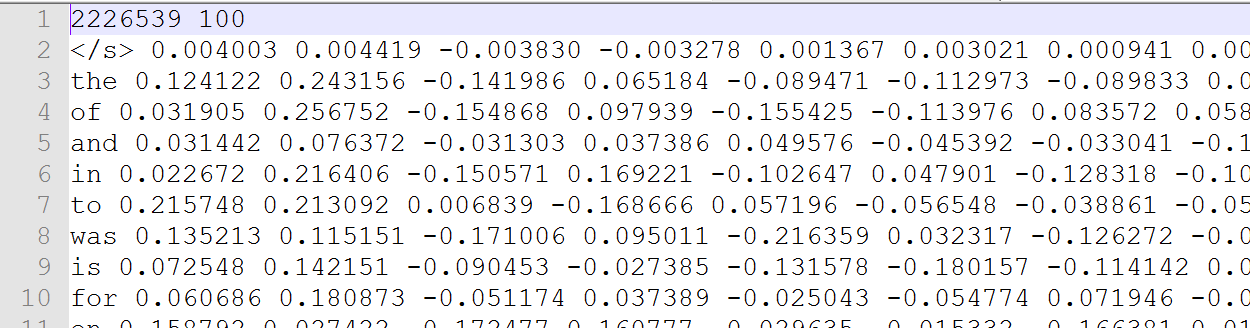
**图4.7 有监督短文本数据流分类完成**



### 搜索示例2

期望分类部分有标签的短文本数据流。首先讲述外部语料库文件、短文本数据流文件以及标签信息文件的格式要求，然后给出操作半监督短文本数据流分类页面的流程示例。

外部语料库是由词向量Word2Vec训练文档数据集获得的所有词的原始词向量集合，这里训练的文档数据集要求包含被分类的短文本数据流中95%的词。外部语料库的文件格式如图4.8所示：

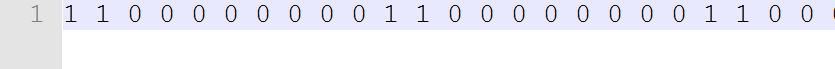


**图4.8半监督短文本数据流分类的外部语料库文件格式**

第一行表示的是词的总数和每个词的词向量表示维度，以后的每一行都是由词和该词对应的词向量表示。

这里的短文本数据流文件与有监督短文本数据流分类中的短文本数据流格式相同，示例1已经介绍过，这里不再赘述。

标签信息区间文件如图4.7所示，根据程序设定将短文本数据流划分为数据块，每个数据块对应一个标签信息，其中1表示有标签，0表示无标签。



**图4.9标签信息文件格式**

半监督短文本数据流分类流程：

第一步：在半监督短文本数据流分类页面输入参数，如图4.10所示。



**图4.10半监督短文本数据流分类的参数设置**

第二步：点击“开始训练”按钮，如果输入参数不为空且参数格式正确，则分类短文本数据流，其分类过程中的相关信息输出在右侧的结果输出框中，分类进程有页面的进度条显示。同时，“开始训练”按钮变为灰色。如图4.11所示。否则，提示参数输入错误，请求重新输入，如图4.12所示。



**图4.11点击按钮后训练并分类短文本数据流**

第三步：当训练完成后，“开始训练”按钮重新变为可点击状态，表示可重新选择有标签的短文本数据流进行分类。如图4.13所示。



**图4.12点击按钮后参数检测错误**



**图4.13半监督短文本数据流分类完成**