

智文（ZIWEN） 多行业文本大数据分析平台

用户手册

v3.0

北京商智通信息技术有限公司

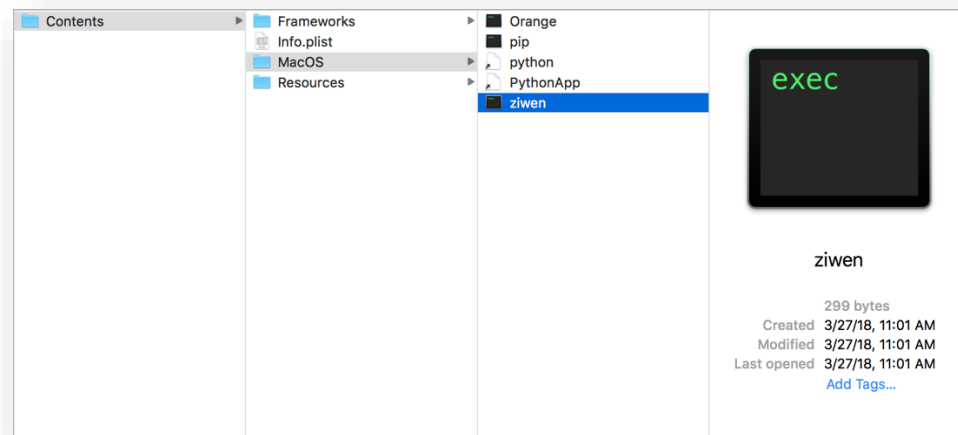
1、平台简介

本平台可以让用户从大量文本数据中挖掘出有价值的信息。这些信息可以帮助用户对市场动向，客户满意度，品牌影响力等一系列重要指标进行量化，从万千无头绪的数据中轻松得出结论。

2、安装说明

想要使用智文，只需要在您的计算机运行我们提供的安装包。对于 Windows 系统的用户，运行 exe 文件后可根据个人情况设置安装路径，安装完成后即可开始使用。如果使用 Mac OS 系统，您只需按照传统安装流程将安装内容拖入应用文件夹即可。

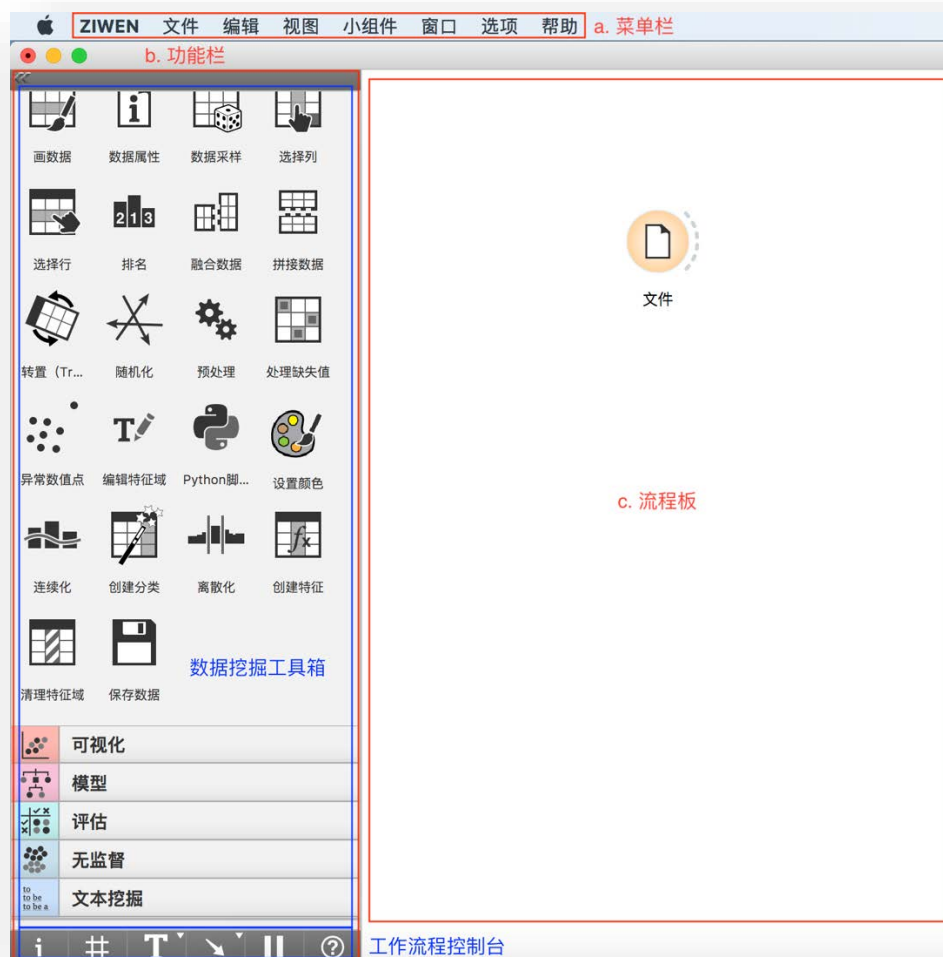
注意：一些 Mac OS 的用户在使用“数据->数据集”组件时系统会提示出错。如果您也遇到同样的问题，请通过路径/Applications/ziwen.app/Contents/MacOS/ziwen 找到文件，双击运行进入智文：



3、界面介绍

运行智文平台后您将看到的是平台的整个界面。此界面是智文唯一的主界面，用户将在这个主页面上完成文本挖掘工作流程全部的设计和搭建工作。界面分为三部分：

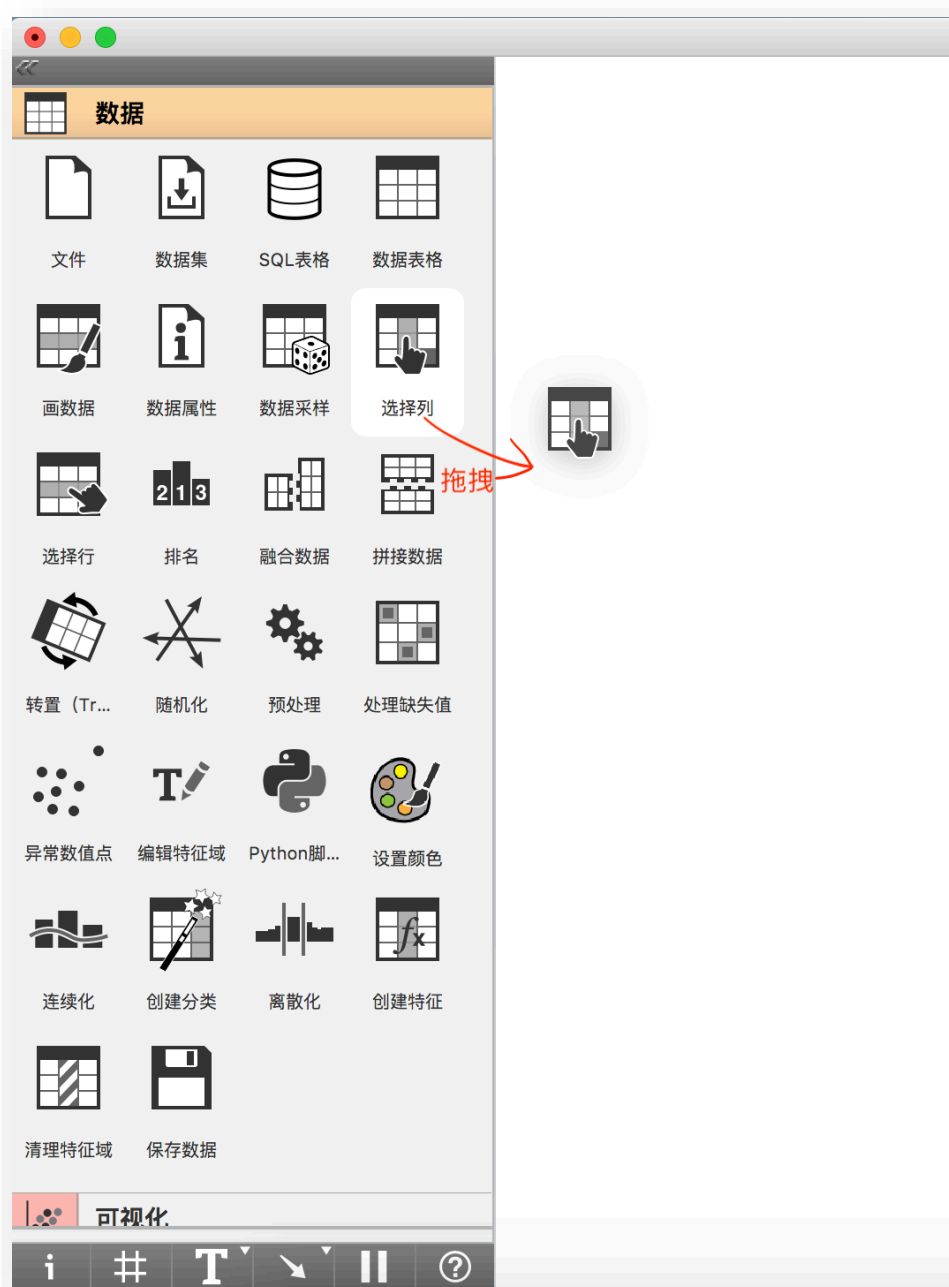
- a) 菜单栏：提供一切众多文本挖掘具体工序以外的功能，如保存/读取工作流程，浏览帮助案例等。
- b) 功能栏：展示并提供全部文本挖掘功能。功能栏可分为两部分：占据绝大部分空间的“数据挖掘工具箱”，和功能栏最下方的“工作流程控制台”。智文为用户提供六大数据挖掘工具箱，近百个功能组件，帮助用户应对数据挖掘工作中的几乎所有数据处理需求。“工作流程控制台”则负责对流程板上的数据挖掘工作流程进行控制，检查，标注，暂停等操作。用户可以根据需求点击功能栏左上角箭头标志，展开或收起功能栏。
- c) 流程板：用户在流程板上画出整个数据挖掘工作流程，完成数据挖掘任务。

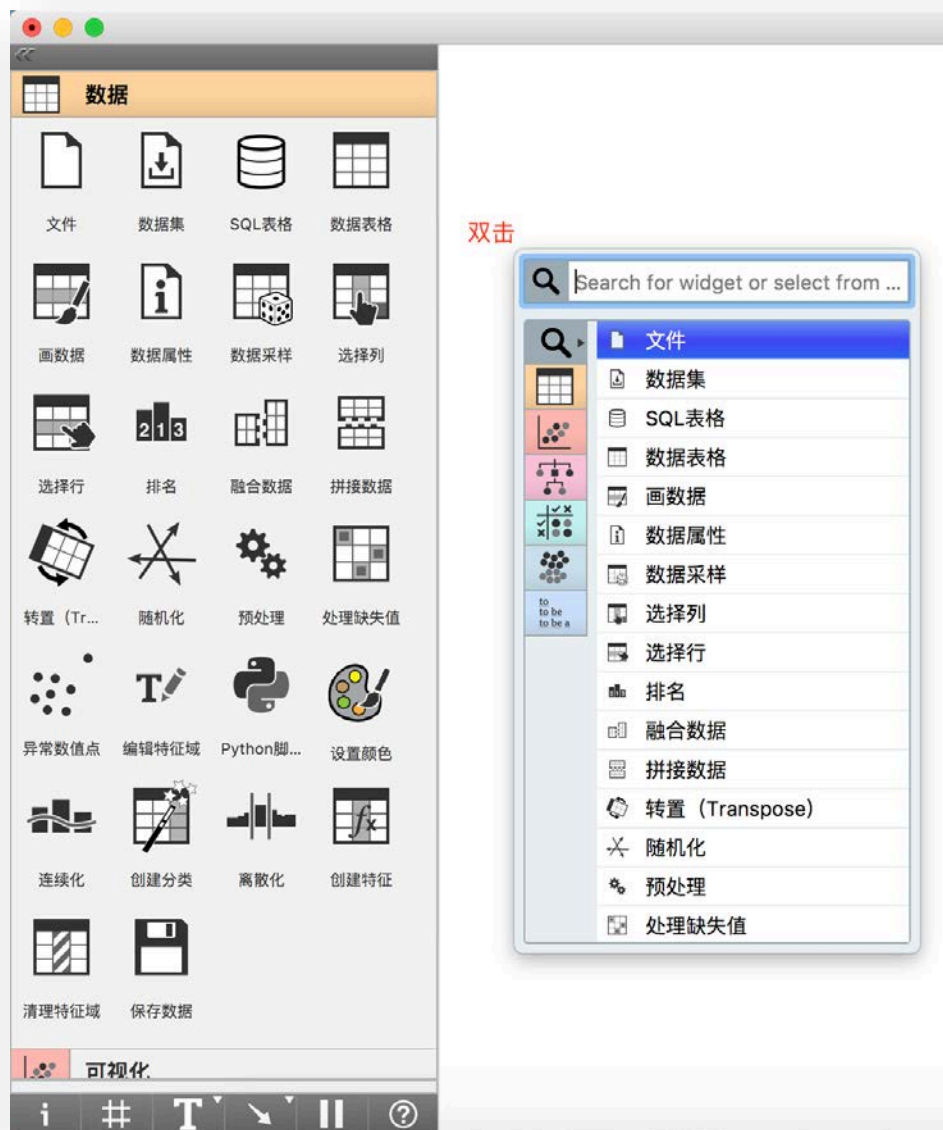


4、基本操作

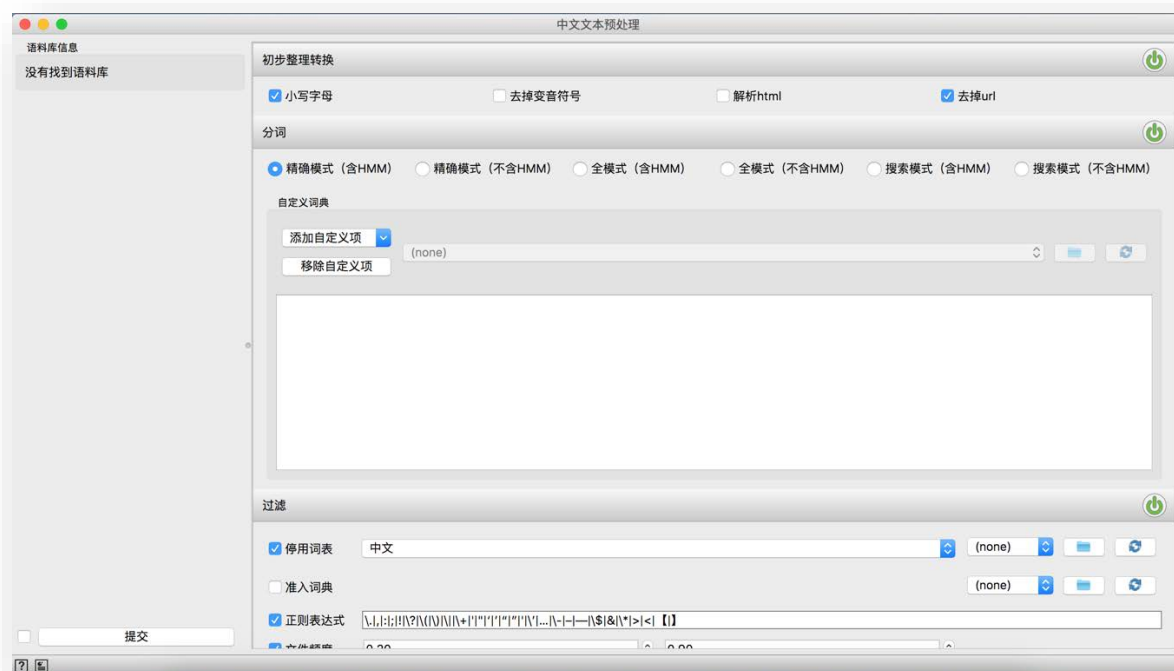
想要在智文平台上解决数据挖掘工作，用户的核心目标是在流程板上构建数据挖掘工作流程，构建工作流程需要三种基本操作：添加功能节点，设置功能节点参数，连接功能节点（一个核心目标，三个基本操作）。下面逐一介绍：

1) 添加功能节点：用户可以通过三种方式在流程板上添加功能节点，效果一样：从左侧数据挖掘工具箱中将功能节点拖拽进流程板，或右键点击流程板空白处后从弹出菜单中选择，或单击工具箱中的功能节点，其会自动出现在流程板上。

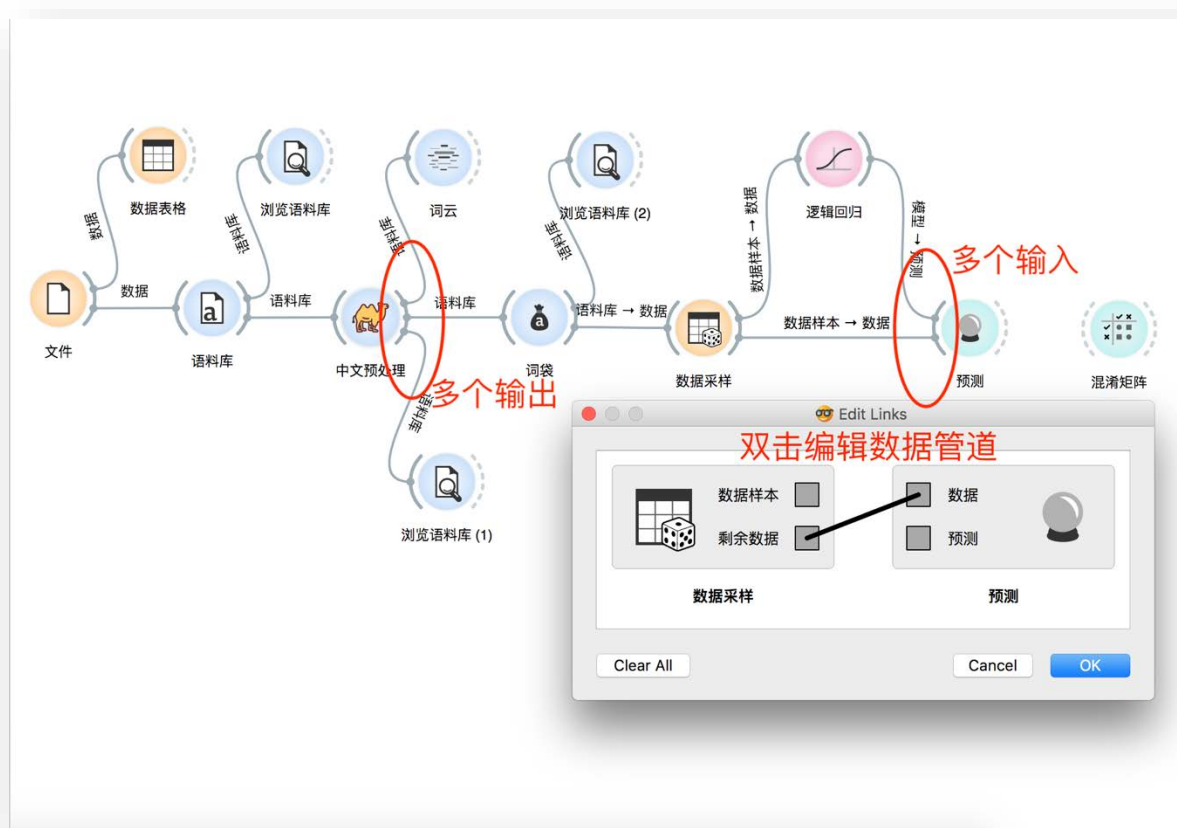




2) 设置功能节点参数：工具箱中的每一个功能节点都对应一个工序，为了不同的数据分析目的，用户需要设计每一道工序的具体实现。双击流程板上的任意功能节点即可查看修改其设置。以下图为例：双击“中文文本预处理”功能节点，用户可以定义一整套应对中文文本的预处理流程，单击“提交”或选中“自动提交”时，被更新的设置将进入整个工作流程。



3) 连接功能节点：流程板上的大部分功能节点都有左右两个弧线，左侧的弧线是功能节点的从上游工序接收数据的入口，右侧的弧线是向下游工序输出数据的出口。当功能节点并没有被连接时，输入输出弧线是虚线。在一个功能节点的输入输出弧线处点击并拖动至另一功能节点的输出输入弧线，数据管道就生成了。数据管道是实线时则说明数据已被输送，虚线则表明管道中并无数据被输送，管道是红色时则意味着发送数据的数据结构与接受数据不一致，管道不成立。一个功能节点可以接收多个数据管道，也可发送多个数据管道，用户可以双击数据管道，查看并修改输入输出接口，保证链接正确。



下图是一个完整的应用逻辑回归方法分类文本数据的工作流程，数据从左往右流动。
“文件”读取源文件，“中文预处理”“词袋”将文本向量化，“数据采样”将数据随机等分两份，一份被加入到“逻辑回归”中训练模型，另一份用来验证模型效果，最终通过“混淆矩阵”查看分类正确率和混淆情况。



5、工具箱与功能节点列表

1. 数据	文件	3. 模型	支持向量机（SVM）
	数据集		逻辑回归
	SQL 表格		线性回归
	数据表格		朴素贝叶斯
	画数据		AdaBoost
	数据属性		神经网络
	数据采样		随机梯度下降
	选择列		保存模型
	选择行		读取模型
	排名	4. 评估	测试和评分
	融合数据		预测
	拼接数据		混淆矩阵
	转置		ROC 分析
	随机化		lift 曲线
	预处理		校准图
	处理缺失值	5. 无监督	距离文件
	异常数值点		距离矩阵
	编辑特征域		距离地图
	Python 脚本		层次聚类
	设置颜色		k 均值聚类
	连续化		流形学习
	创建分类		主成分分析（PCA）
	离散化		对应分析
	创建特征		距离
	清理特征域		距离转换
	保存数据		多维尺度（MDS）
			保存距离
2. 可视化	决策树查看	6. 文本挖掘	语料库
	箱线图		导入文档
	分布		卫报
	散点图		纽约时报
	筛图		Pubmed
	马赛克展示		推特
	FreeViz		维基百科
	线性预测		中文文本预处理
	Radviz		文本预处理
	热力图		词袋
	文氏图		相似度计算
	轮廓图		情感分析
	毕达哥拉斯树		

	毕达哥拉斯森林		推特事件探查器
	CN2 规则查看		主题模型化
	列线图		语料库查看
3. 模型	常数		词云
	CN2 规则归纳		一致性
	k-最邻近（kNN）		地图
	决策树		词扩展
	随机森林		重复检测