**基于Twitter的网络结构和社会群体演化**

**实验报告**

****

**学号：201600301079**

**姓名：崔玉峰**

**班级：2016级**

## 问题描述

1. **问题题目**

掌握从社交网络数据抽取用户关系并进行数据分析的方法。

1. **解题思路**
2. 地震前后网络构建根据附件 1 和附件 2 给出的推文数据，提取出用户发文时间、用户昵称、发文内容等关键字段，并根据这些信息构建地震前后的关系网络。以@作为每个用户之间的关系。

2. 对构建好的网络进行度分析为了对比地震对用户关系带来的影响，分别对去的英文使用者网络和日文使用者网络进行累积度分布变化统计、单独节点层面上的度变化统计并做对比分析。

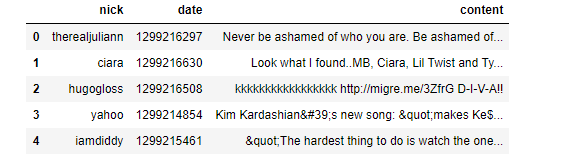
3. 通过Gephi对网络节点进行可视化

## 具体实现：

1. **加载数据**
2. 将所给的json文件数据读进来
3. 主要选择content(发文内容)，nick（用户昵称），data(发文时间)三个json字段进行分析

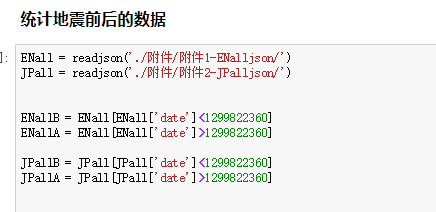


1. 都进来json数据之后转为DataFrame格式方便以后运算



1. **数据预处理：**
2. 分出根据data分出地震前后英文/日文发帖的数据。

按照地震时间（2011年3月11日 13：46）



1. 对数据进行去重处理
2. **构建关系网络**

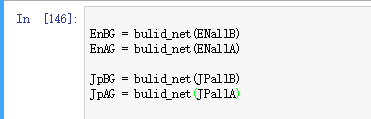
NetworkX 是一个用 Python 语言开发的图论与复杂网络建模工具，内置了常用的图与复杂网络分析算法，可以方便的进行复杂网络数据分析、仿真建模等工作。

1. 网络中每个节点是一个用户。
2. 如果用户A在发文中@了用户B，则认为用户A与用户B有关系，则在节点A与节点B中增加一条边。

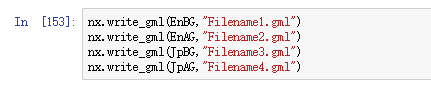
根据上述网络构建规则，对地震前后，日语英语使用者建立关系网络



地震前后，日语英语使用者建立关系网络：



将构建的网络存入gml文件

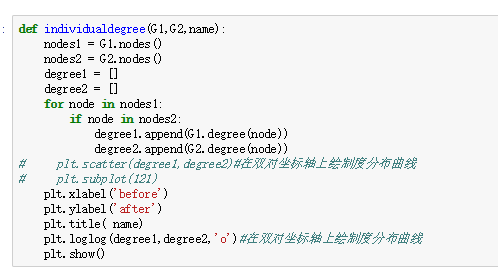


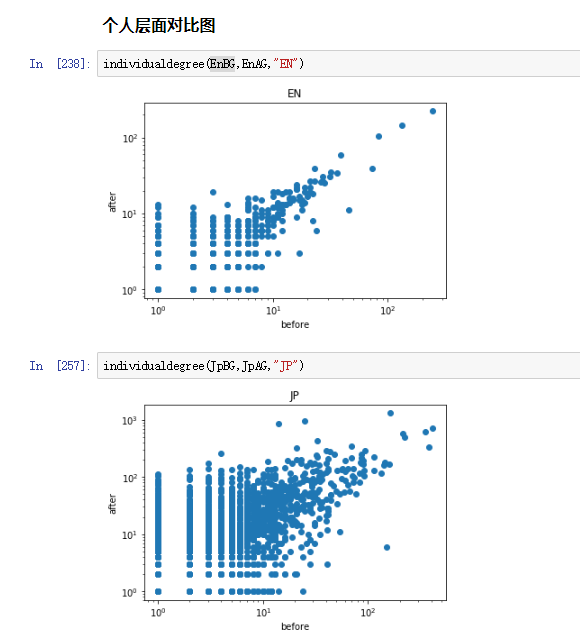
1. **对构建的网络进行分析**
2. **求出网络的基本特性（平均度、最大联通片、群聚系数）**





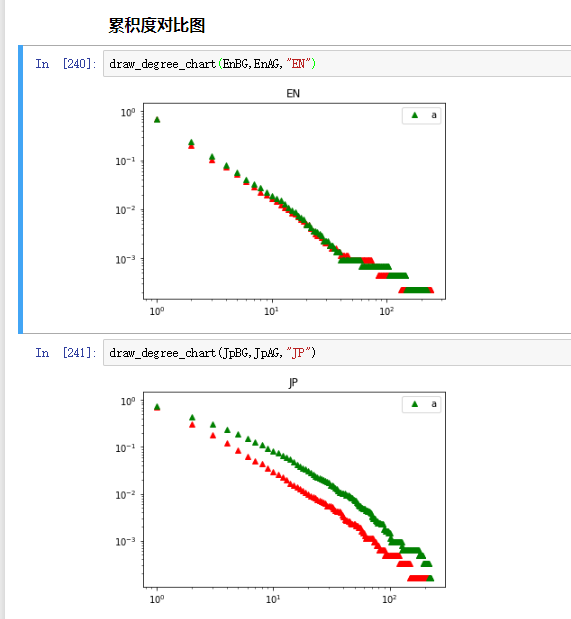
1. **个人层面分析**





1. **累积度分析**

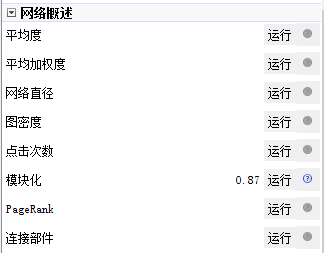




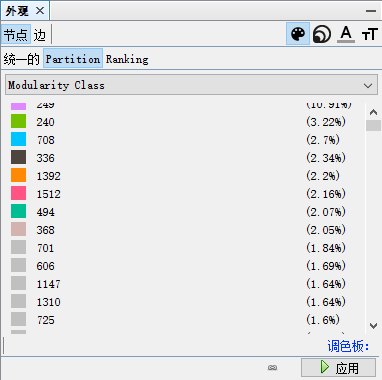
1. **通过Gephi软件对关系网络进行可视化**
   * + 1. 下载好Gephi软件：



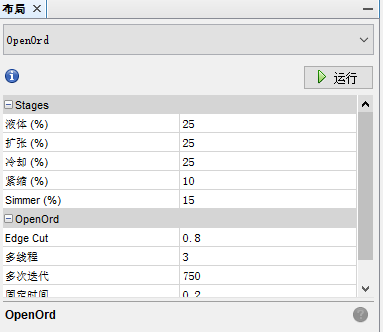
* + - 1. 导入创建好的gml文件
      2. 设置模块化参数



* + - 1. 设置设置外观节点颜色

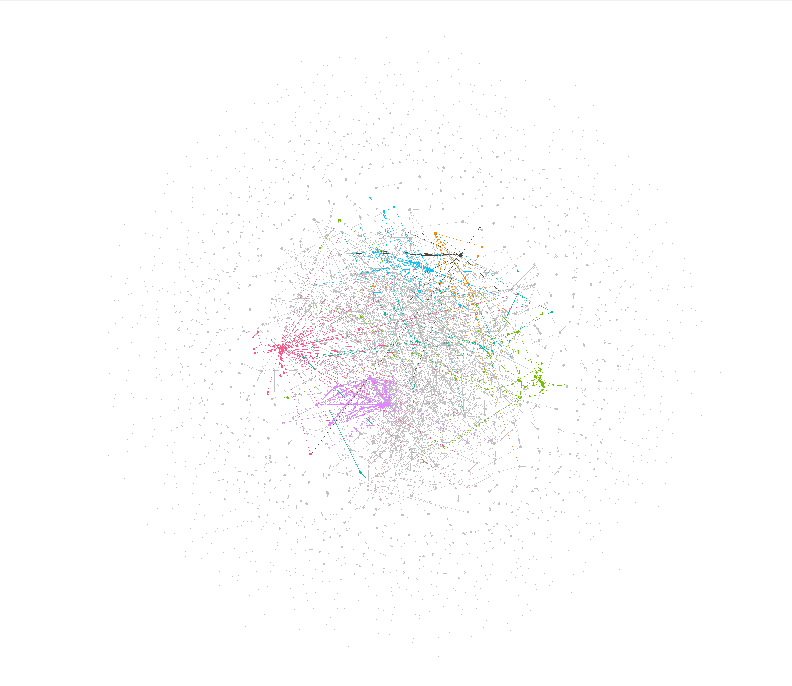
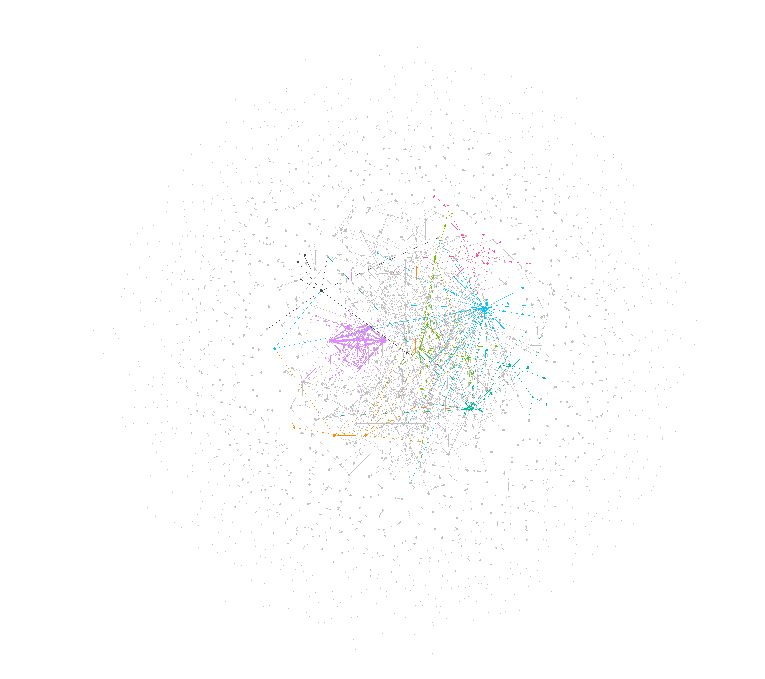


* + - 1. 设置布局

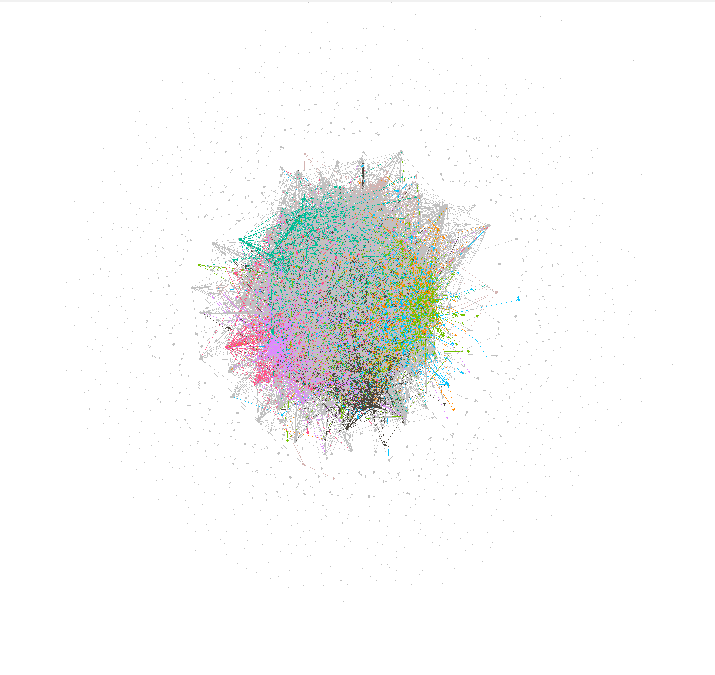
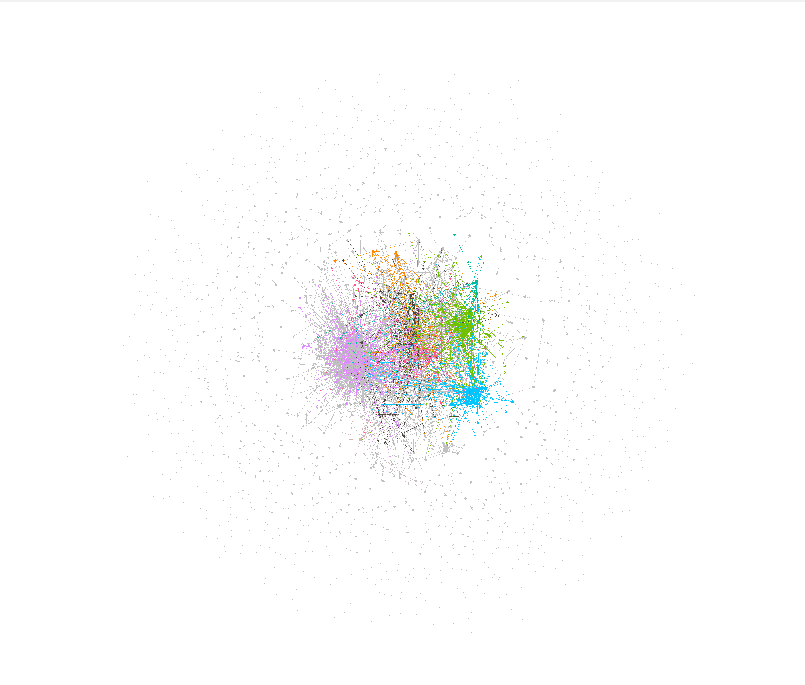


* + - 1. 对四个网路进行可视化

英语地震前 英语地震后



日语地震前 日语地震后



## 结果分析：

通过结果观察结果分析数据，训练模型得到了一个可以信赖的线性模型可以根据餐馆提供的相关数据对其的销售额进行预测，并且这个模型相对简单只要几个参数，可以很方便的存储和使用。

## 实验结论和感悟：

本次实验的数据和代码会通过.ipynb的形式上传可以随时运行查看。

具体的实现并未完全按照老师给定的代码进行实现，因此在结果上会与老师给定的代码的输出有所不同，但是在大体结果上还是一致的。

通过几次的数据科学题目也可以总结出一些共性的东西了，处理类似的数据分析题目的流程大致为：

获取数据

对数据进行预处理

对数据进行可视化操作观察数据的特点

根据数据特点和需求选择模型

筛选和处理数据，并喂给模型做训练

优化模型参数

得出结论分析结果。