目录

[主题模型（Topic Model） 2](#_Toc498632302)

[简介 2](#_Toc498632303)

[主题1：分手 5](#_Toc498632304)

[主题2：人生 7](#_Toc498632305)

[主题3：两性 9](#_Toc498632306)

[主题4：校园生活 11](#_Toc498632307)

[主题5：宿舍 13](#_Toc498632308)

[基于时序变化的情感分析 16](#_Toc498632309)

[简介 16](#_Toc498632310)

[以月份为单位 16](#_Toc498632311)

[以星期为单位 18](#_Toc498632312)

[以天为单位 20](#_Toc498632313)

[以小时为单位 22](#_Toc498632314)

# 主题模型（Topic Model）

## 简介

**主题模型**（Topic Model）在[机器学习](https://zh.wikipedia.org/wiki/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0" \o "机器学习)和[自然语言处理](https://zh.wikipedia.org/wiki/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86)等领域是用来在一系列文档中发现抽象主题的一种统计模型。直观来讲，如果一篇文章有一个中心思想，那么一些特定词语会更频繁的出现。比方说，如果一篇文章是在讲狗的，那“狗”和“骨头”等词出现的频率会高些。如果一篇文章是在讲猫的，那“猫”和“鱼”等词出现的频率会高些。而有些词例如“这个”、“和”大概在两篇文章中出现的频率会大致相等。但真实的情况是，一篇文章通常包含多种主题，而且每个主题所占比例各不相同。因此，如果一篇文章10%和猫有关，90%和狗有关，那么和狗相关的关键字出现的次数大概会是和猫相关的关键字出现次数的9倍。一个主题模型试图用数学框架来体现文档的这种特点。主题模型自动分析每个文档，统计文档内的词语，根据统计的信息来断定当前文档含有哪些主题，以及每个主题所占的比例各为多少。ｗ

ｉ

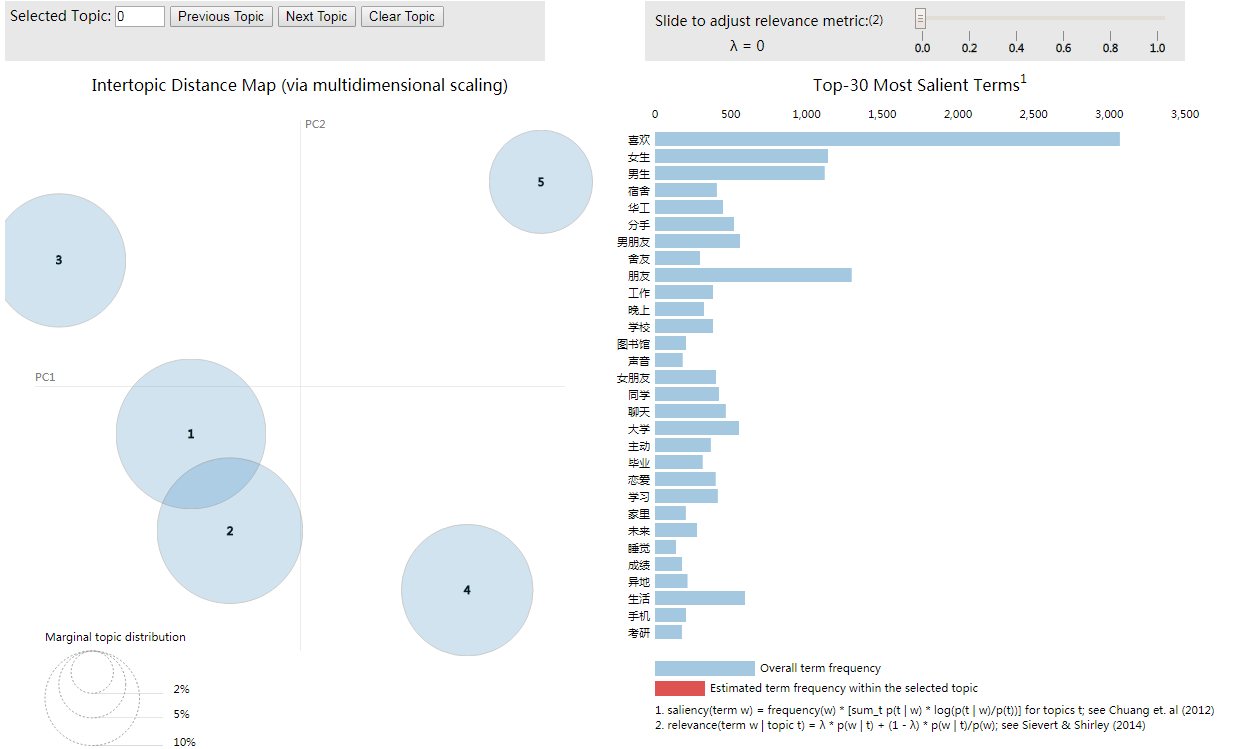
**隐含狄利克雷分布**简称LDA(Latent Dirichlet allocation)，是一种[主题模型](https://zh.wikipedia.org/wiki/%E4%B8%BB%E9%A2%98%E6%A8%A1%E5%9E%8B" \o "主题模型)，它可以将文档集中每篇文档的主题按照[概率分布](https://zh.wikipedia.org/wiki/%E6%A6%82%E7%8E%87%E5%88%86%E5%B8%83)的形式给出。同时它是一种[无监督学习](https://zh.wikipedia.org/wiki/%E9%9D%9E%E7%9B%A3%E7%9D%A3%E5%BC%8F%E5%AD%B8%E7%BF%92)算法，在训练时不需要手工标注的训练集，需要的仅仅是文档集以及指定主题的数量k即可。此外LDA的另一个优点则是，对于每一个主题均可找出一些词语来描述它。

LDA首先由Blei, David M.、[吴恩达](https://zh.wikipedia.org/wiki/%E5%90%B4%E6%81%A9%E8%BE%BE" \o ")和Jordan, Michael I于2003年提出[[1]](https://zh.wikipedia.org/wiki/%E9%9A%90%E5%90%AB%E7%8B%84%E5%88%A9%E5%85%8B%E9%9B%B7%E5%88%86%E5%B8%83" \l "cite_note-blei2003-1)，目前在[文本挖掘](https://zh.wikipedia.org/wiki/%E6%96%87%E6%9C%AC%E6%8C%96%E6%8E%98" \o "文本挖掘)领域包括文本主题识别、文本分类以及文本相似度计算方面都有应用。

根据4800多条悄悄话数据，通过LDA主题建模，大致分为五个主题，LDA主题模型通过计算会给出五组描述不同模型的关键词，这五组关键字就刻画了五个主题的特征，粗略命名如下：

1. 分手
2. 人生
3. 两性
4. 校园生活
5. 宿舍

借助可视化工具pyLDAvis，可以更好地描述主题模型。



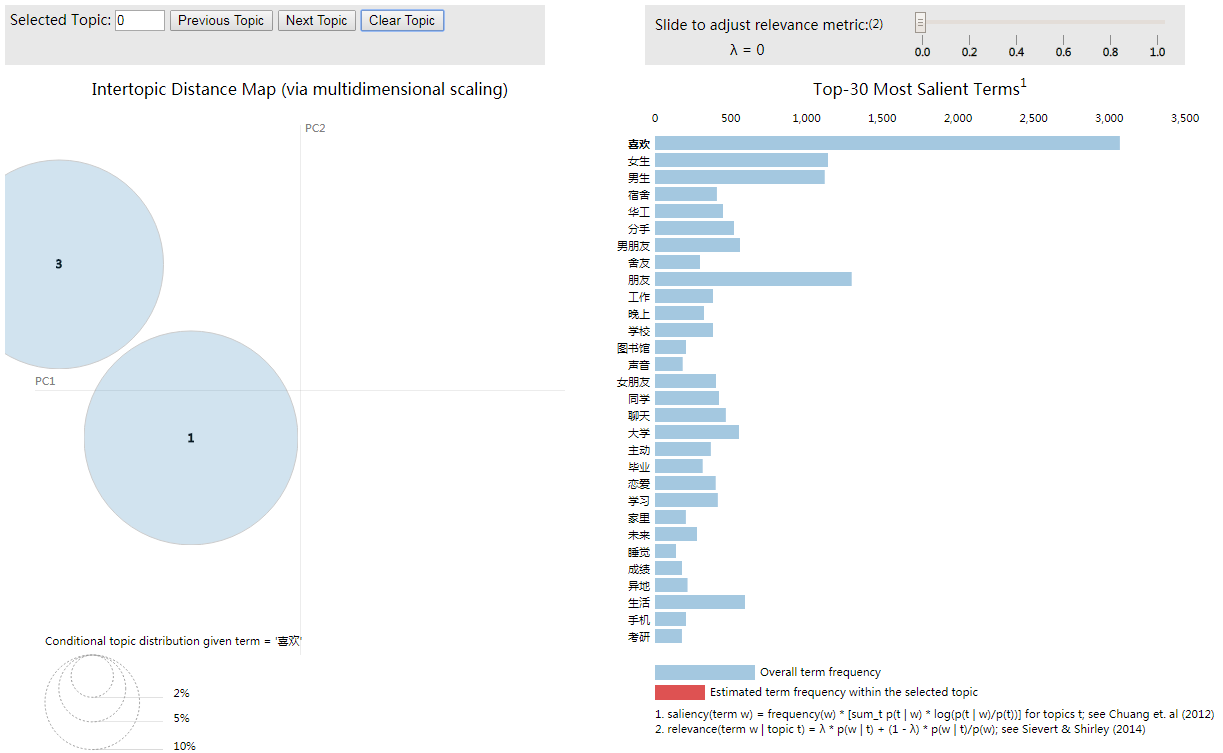
在图中左边有一个二维坐标轴，其中有5个圆形，圆心标号1到5，正对应着主题号码1到5。

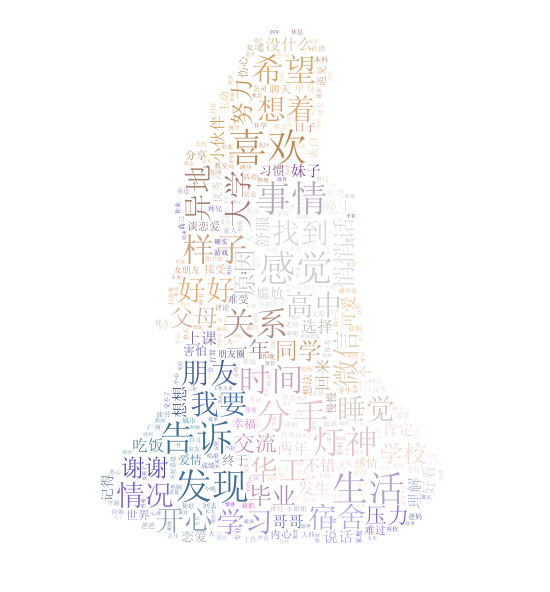
圆形的面积越大，说明包含的悄悄话数量越多，从图中可以看到圆形面积从1到5递减，这说明悄悄话主人们经常发送关于“分手”、“人生”等悄悄话，“宿舍”主题数量是最少的，查看悄悄话数据，这一点是基本符合的。

两个圆形距离越近，说明对应两个主题联系越密切，如图中的主题1（分手）与主题2（人生），接下来的关键字刻画也可以佐证这一点，而主题5（宿舍）则与其他4个主题距离较远，这也是在情理之中的。

在图中右边，是从所有悄悄话中归纳出来的前30个主要关键词（Top-30 Most Salient Terms）。对应柱状越长，则说明词频越大（具体公式见论文），可以看到“喜欢”这个词是悄悄话经久不衰的关键词，这和我们的直观感受相符。

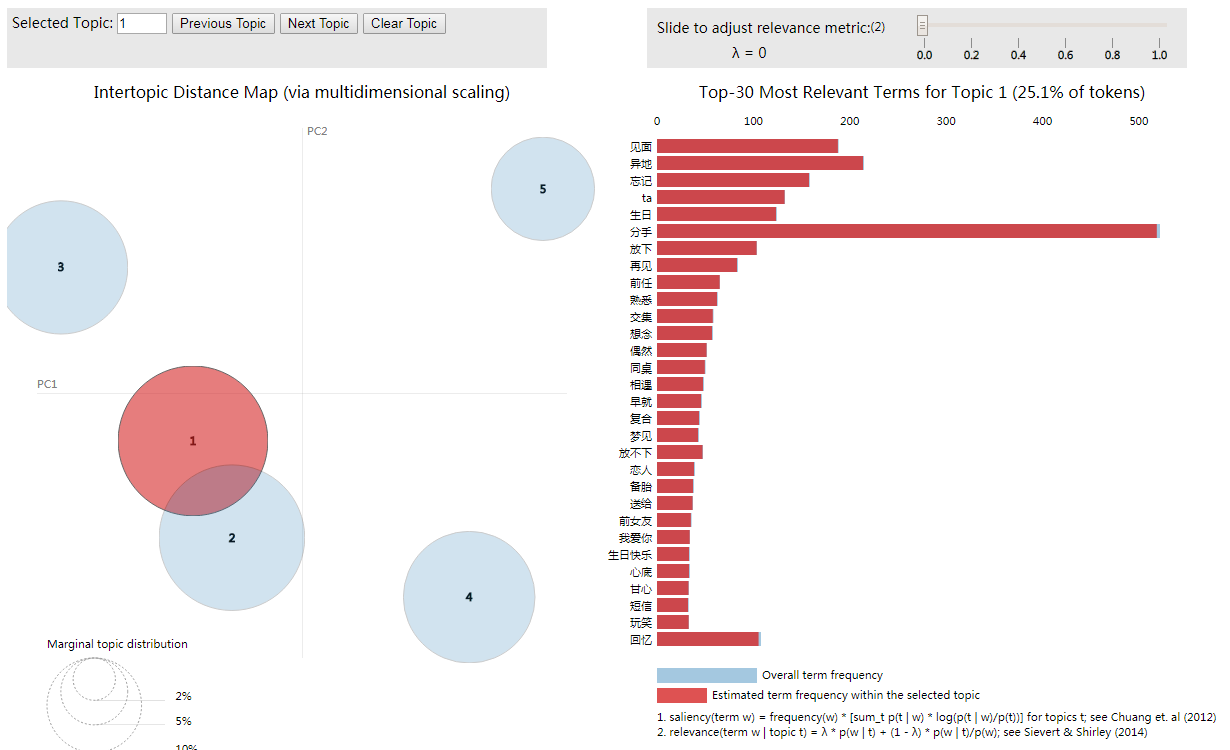
通过选定“喜欢”关键词，可以直接看到“喜欢”与主题1（分手）、主题3（两性）相关，另外三个主题则不会出现“喜欢”关键词。





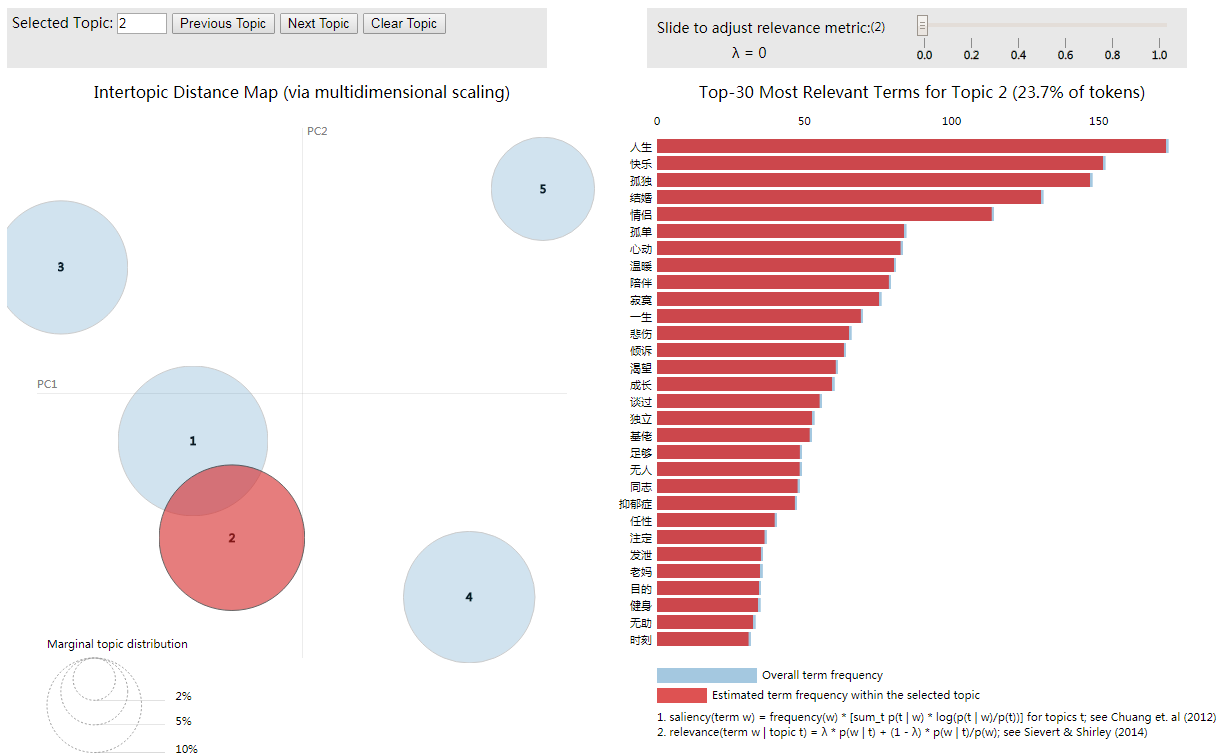
## 主题1：分手

虽然大家都希望拥有美满的感情，但是事与愿违，情侣之间经常会出现感情破裂直至分手，很多时候悄A、悄B、悄C都是关于分手的故事。看一看刻画分手主题特征的关键词有哪一些：





## 主题2：人生

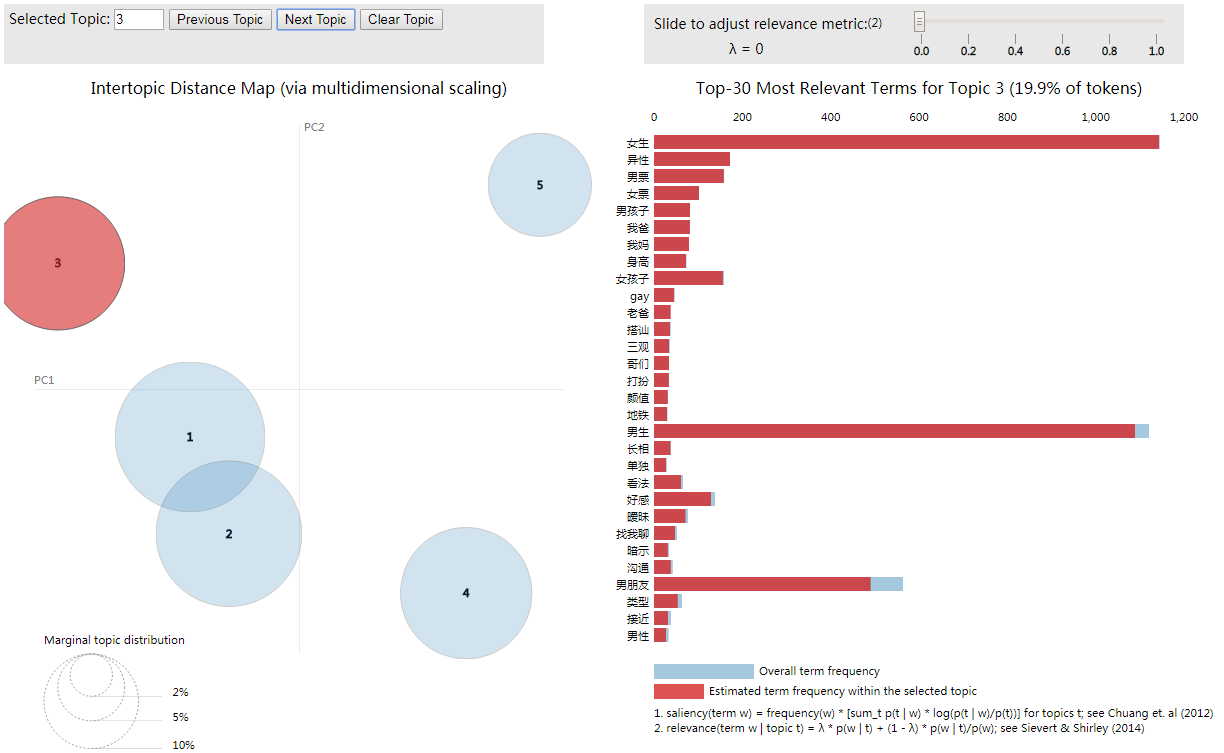


主题2的关键词就很耐人寻味了，快乐、孤独、结婚、情侣、孤单、心动、温暖、陪伴、寂寞、一生、悲伤、倾诉、渴望、成长、谈过（恋爱）等等。

这里面既有积极的，如快乐、温暖、陪伴，又有消极的，如孤独、寂寞、悲伤、抑郁症。在悄悄话中可以一览人生百味。



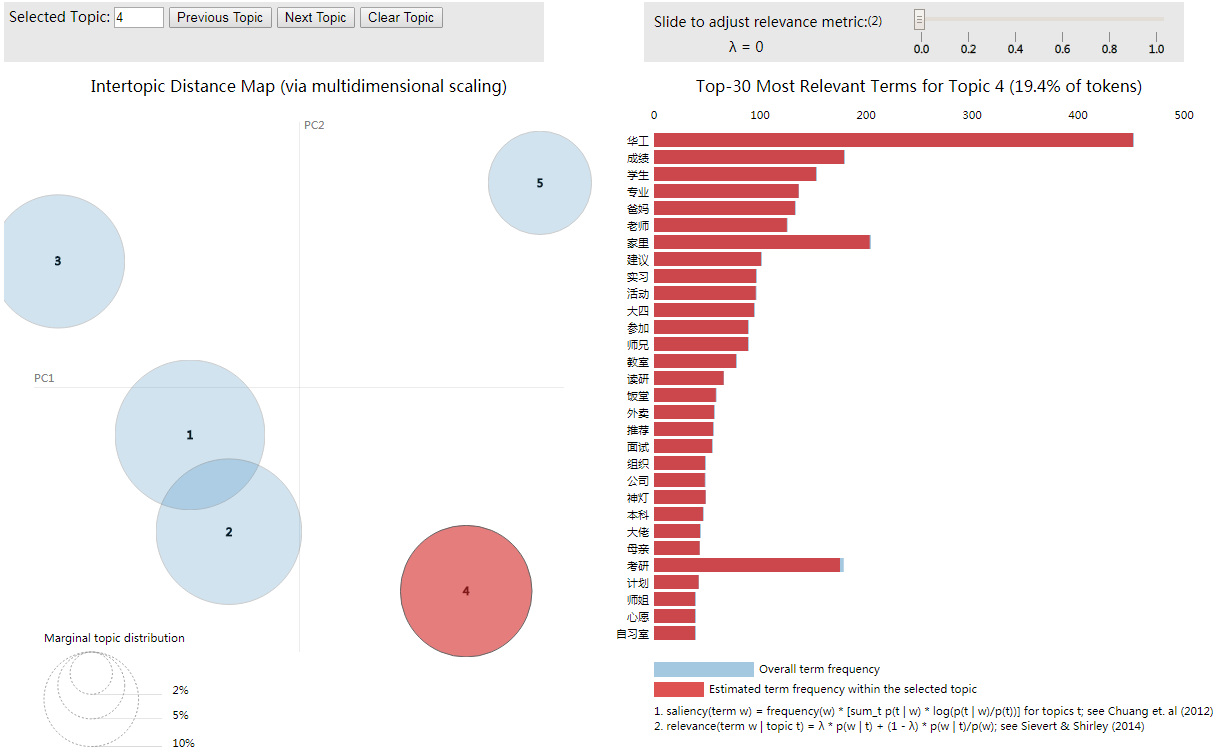
## 主题3：两性



主题3中，女生、男生占了绝大部分的比重，所以将其称之为两性主题，关键词还有三观、搭讪、打扮、颜值、看法、暗示、接近等等，甚至还有一个地点关键词：地铁。这种主题的悄悄话主人们看来是想找人，或者发表一些对于喜欢的异性/同性的观点。



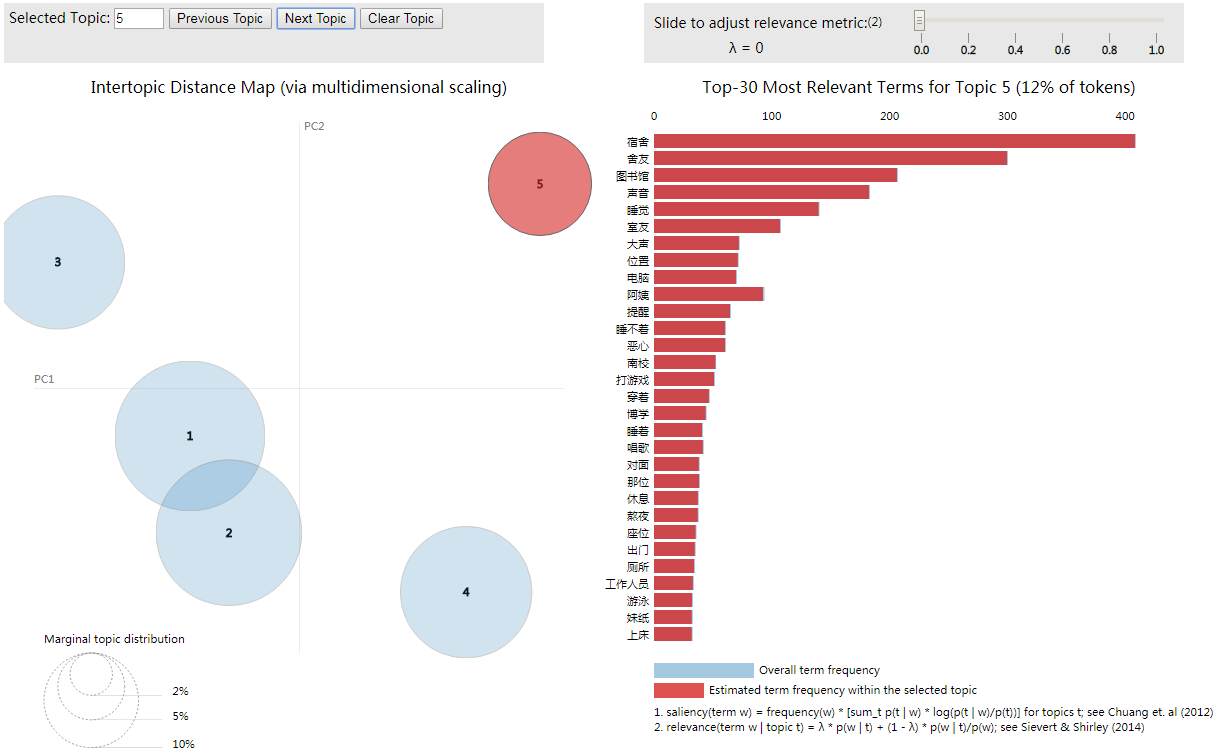
## 主题4：校园生活



在主题4中，关键词“华工”占了主导地位，从其他关键词可以看到有考研党、外卖党、面试党、实习党、学习党等等。可以说这个主题反映了主人们的校园生活，而且还有师兄、师姐，一般来说，高年级一般会经常给低年级一些学习建议、面试建议、成长建议等等。



## 主题5：宿舍



主题5相对于前面的主题就显得“狭隘”很多，这里主要关于宿舍的问题，宿舍、舍友是显著度最高的两个关键词。一提起“宿舍”，人们会想到什么？舍友大声说话、打游戏的鼠标键盘声、熬夜修仙党等等。可以看到，许多关键词是有联系的。比如：“对面”“那位”同学“唱歌”，影响到我“休息”了。的确，在同一个屋檐下生活，难免会有许多摩擦，多一点宽容和理解，宿舍就会和谐。



# 基于时序变化的情感分析

## 简介

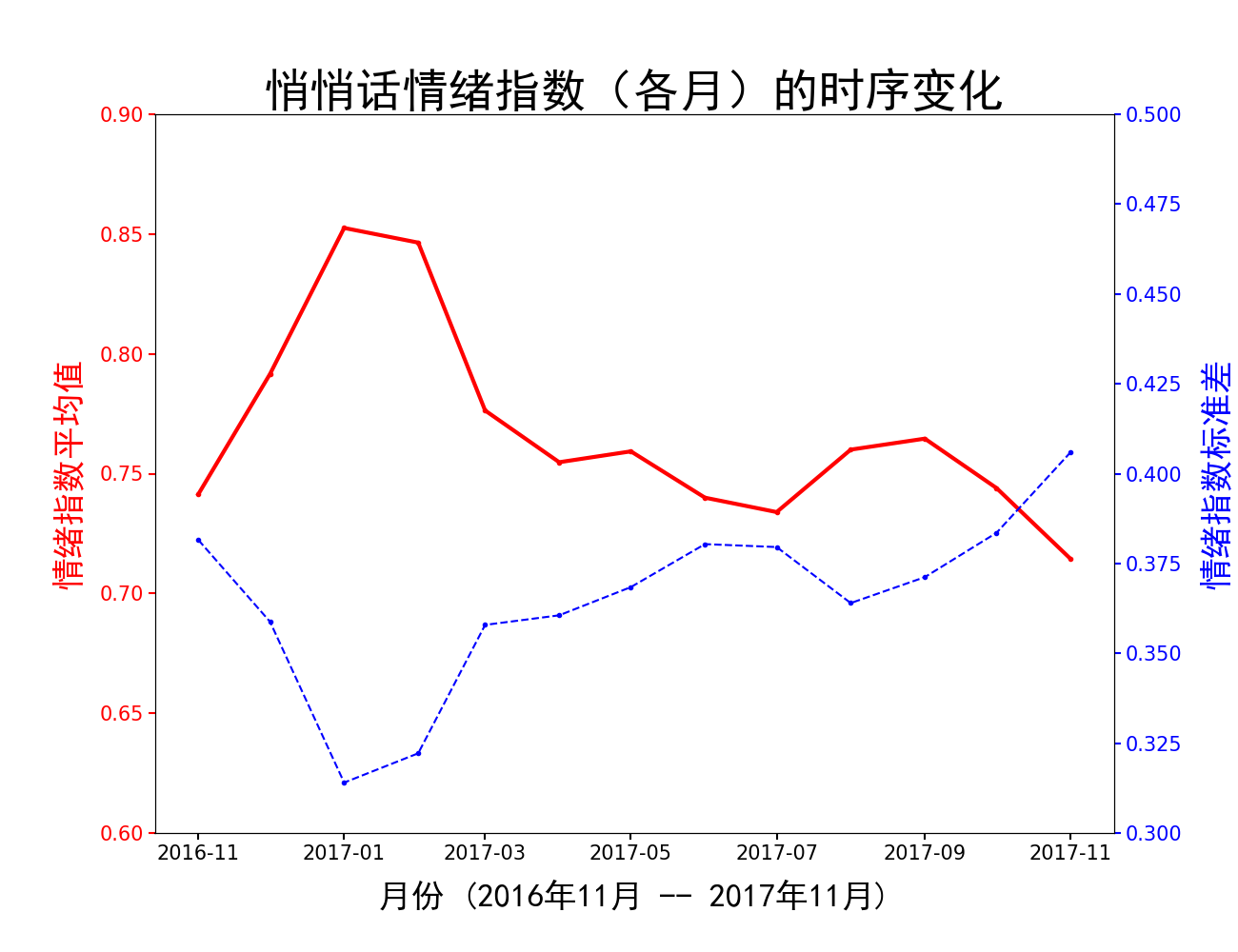
**文本情感分析**（也称为**意见挖掘**）是指用[自然语言处理](https://zh.wikipedia.org/wiki/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86)、[文本挖掘](https://zh.wikipedia.org/wiki/%E6%96%87%E6%9C%AC%E6%8C%96%E6%8E%98)以及[计算机语言学](https://zh.wikipedia.org/wiki/%E8%AE%A1%E7%AE%97%E6%9C%BA%E8%AF%AD%E8%A8%80%E5%AD%A6)等方法来识别和提取原素材中的主观信息。

通常来说，情感分析的目的是为了找出说话者/作者在某些话题上或者针对一个文本两极的观点的态度。这个态度或许是他或她的个人判断或是评估，也许是他当时的情感状态（就是说，作者在做出这个言论时的情绪状态），或是作者有意向的情感交流（就是作者想要读者所体验的情绪）。

文本情感分析的一个基本步骤就是将文本中的某段已知文字的两极性进行分类，这个分类可能是在句子级、功能级。分类的作用就是判断出此文字中表述的观点是积极的、消极的、还是中性的情绪。更高级的“超出两极性”的情感分析还会寻找更复杂的情绪状态，比如“生气”、“悲伤”、“快乐” 等等。

## 以月份为单位

取出各月内所有的悄悄话，进行情感强度分析，然后以月份分类取平均值。



由图可知：

红色为平均值，数值越大代表情绪越积极，数值越小代表情绪越消极。

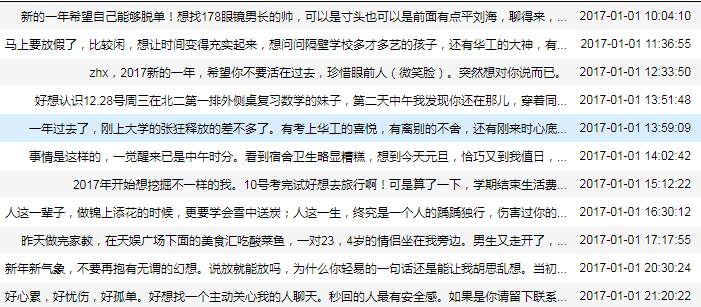
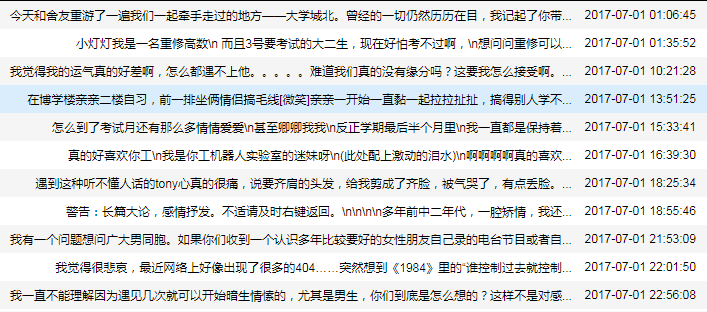
蓝色为标准差，数值越大代表情绪波动越大（不稳定），数值越小代表情绪波动越小（越稳定）。

分析：

通过红色曲线的趋势可以看出，2017-01的悄悄话的情绪指数平均最高，且波动最小，2017-02仅次之，其他月份情绪指数较低，特别是是年中、年末的时候，波动剧烈。

可以发现，年初时，悄悄话主人的情绪高涨，有很多对新的一年的打算，有充实的计划，有美好的希望等。到了年中、年末时，发现有许多毕业分手党、期末考砸党、考研焦躁党等等，所以情绪波动十分剧烈。

附2017-01与2017-07节选悄悄话

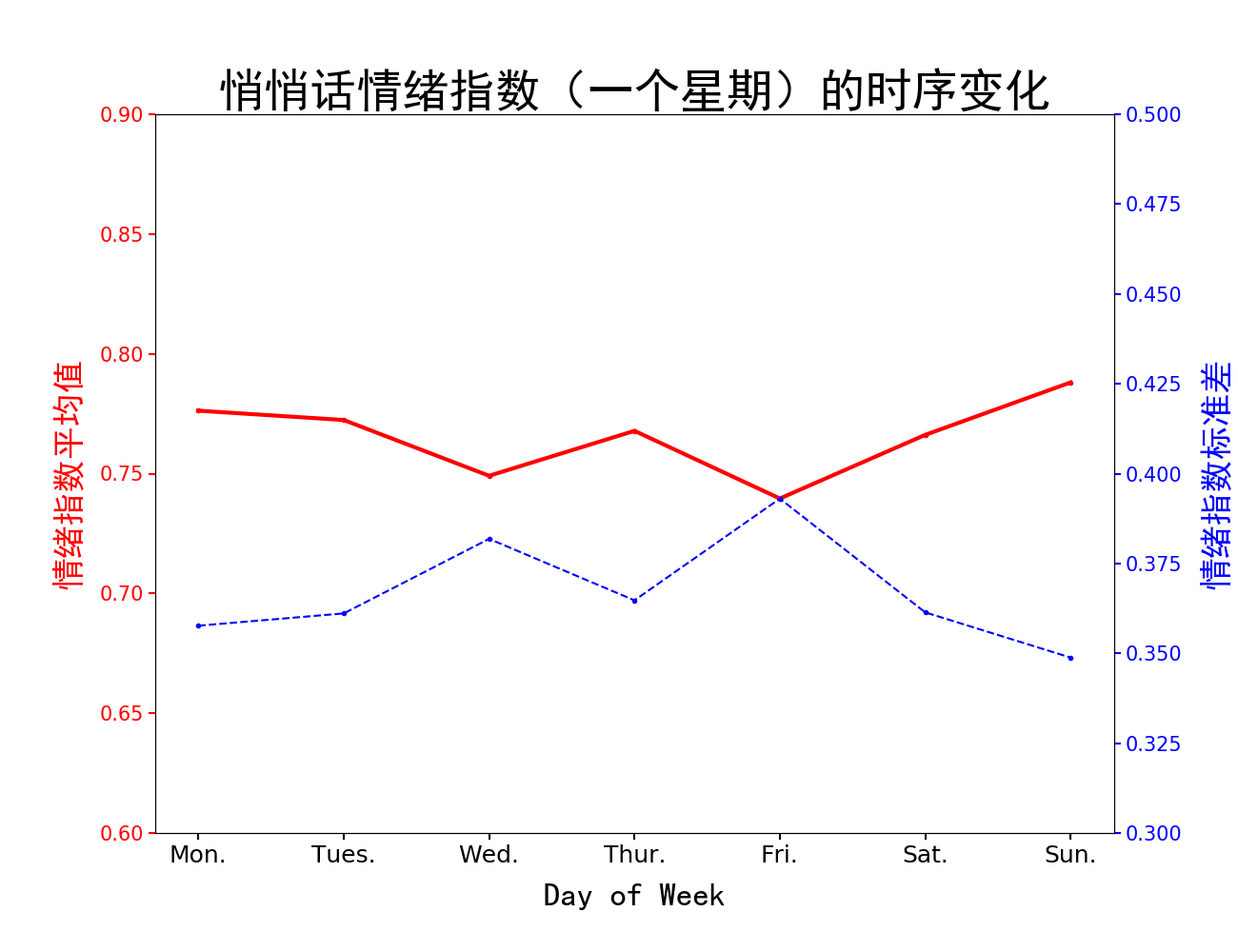
## 以星期为单位

接下来是一个星期的情绪指数变化，以及各天悄悄话数量在一个星期所占的比重：

分析：

每个悄悄话都会找到属于自己的星期几，以星期一到星期五为工作日，星期六、星期天为周末，根据每天的悄悄话数量占一个星期的比例，可以发现，以星期一为峰值，以星期五为谷底，若首尾相接，可以想象，呈现一副类似正弦函数的图像。这与我们的认知有些许偏差，可能我们会认为周末的悄悄话数量会远大于工作日悄悄话数量，但实际上并不是这样，星期一反而是悄悄话最多的时候。

不过，换一种角度来思考也可以想通这个问题，周末都在进行各种活动，星期一才有内容来发悄悄话，工作日比较乏味，一直到星期五的底谷，到星期六才开始回升。



分析：

可以很明显的发现：相对来说，工作日的情绪低落，周末的情绪愉悦。

这和我们的基础认知基本一致：工作日的学习、工作压力较大，人们的情绪普遍低落，而到了周末，工作压力骤减，人们会选择在周末出游、聚会等等，整体情绪高涨。

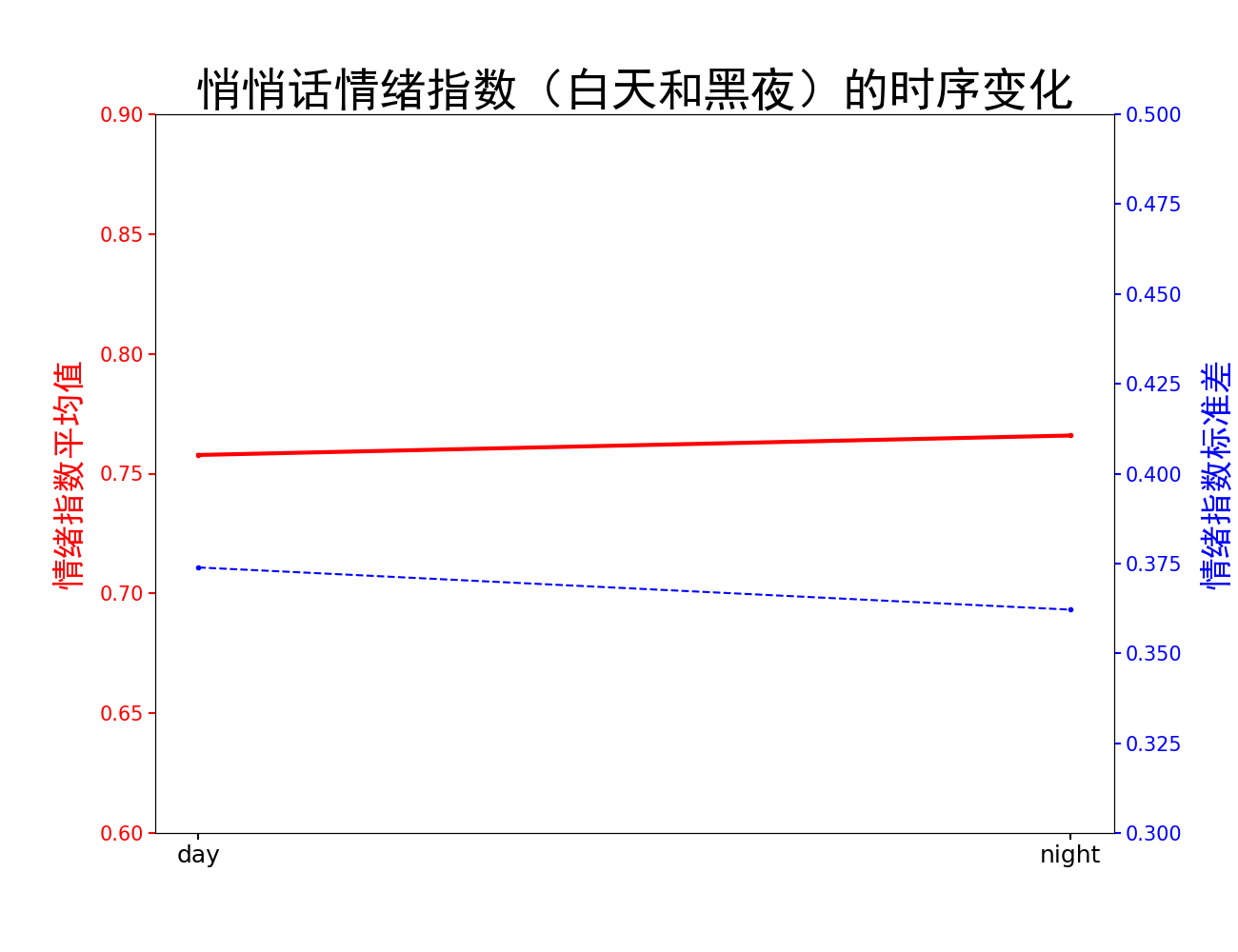
在这幅图可以发现很有趣的规律，平均值越大的时候，标准差越小，似乎是互相成反比的关系。情绪积极高涨时，情绪稳定；情绪消极低落时，情绪起伏。

## 以天为单位

把时间段再缩小一点，把一天分为两个时间段：[6-18，18-6]，即白天和黑夜，把各自时间段中的悄悄话整理出来，分析情绪指数，同样可以得出一些有趣的结论。

分析：

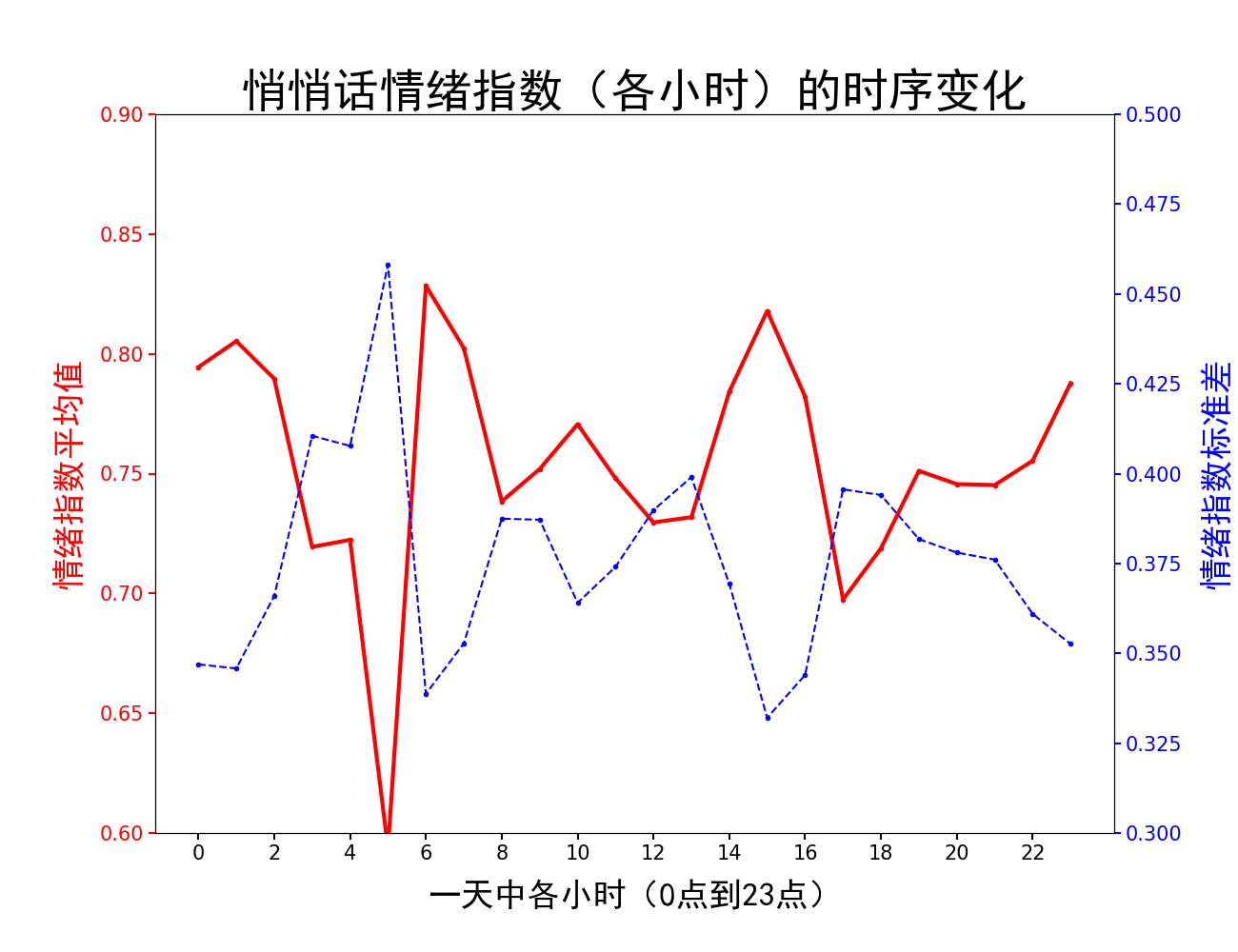
在这里，有必要找到各时间段悄悄话数量在一天中所占比重，通过整理后，可以发现在零点附近的悄悄话数量是最多的，这也是情理之中，白天主人们都比较忙，只有晚上主人们才有心思整理思绪，好好地说悄悄话。



分析：

曲线图，一天当中，白天的情绪指数更低，晚上的情绪指数高更高，这似乎与我们预期的相反，可能我们会认为：到了晚上，人会变得犹豫起来，可惜通过情感分析，这个结论似乎不完全对，对于大学生来说，可能晚上的生活更加丰富多彩。

## 以小时为单位



特别值得注意的是凌晨这个时间段，通过分析发现，所有的悄悄话主要集中于23点至次日1点之间，那如果把时间段在缩小一点，观察一天中各小时的情绪指数，则波动较大。

人越熬夜修仙，情绪就越低落，而且也更加不稳定，7点以及13点左右是一天中情绪较为积极的时刻，大概对应早上起来和午睡起来后，这时情绪很积极，而且都很稳定。而且纵观全局，可以发现，饭后的情绪较稳定，离吃饭的时间越远，情绪越不稳定。