电商平台评论数据情感分析

报告人：汤琼

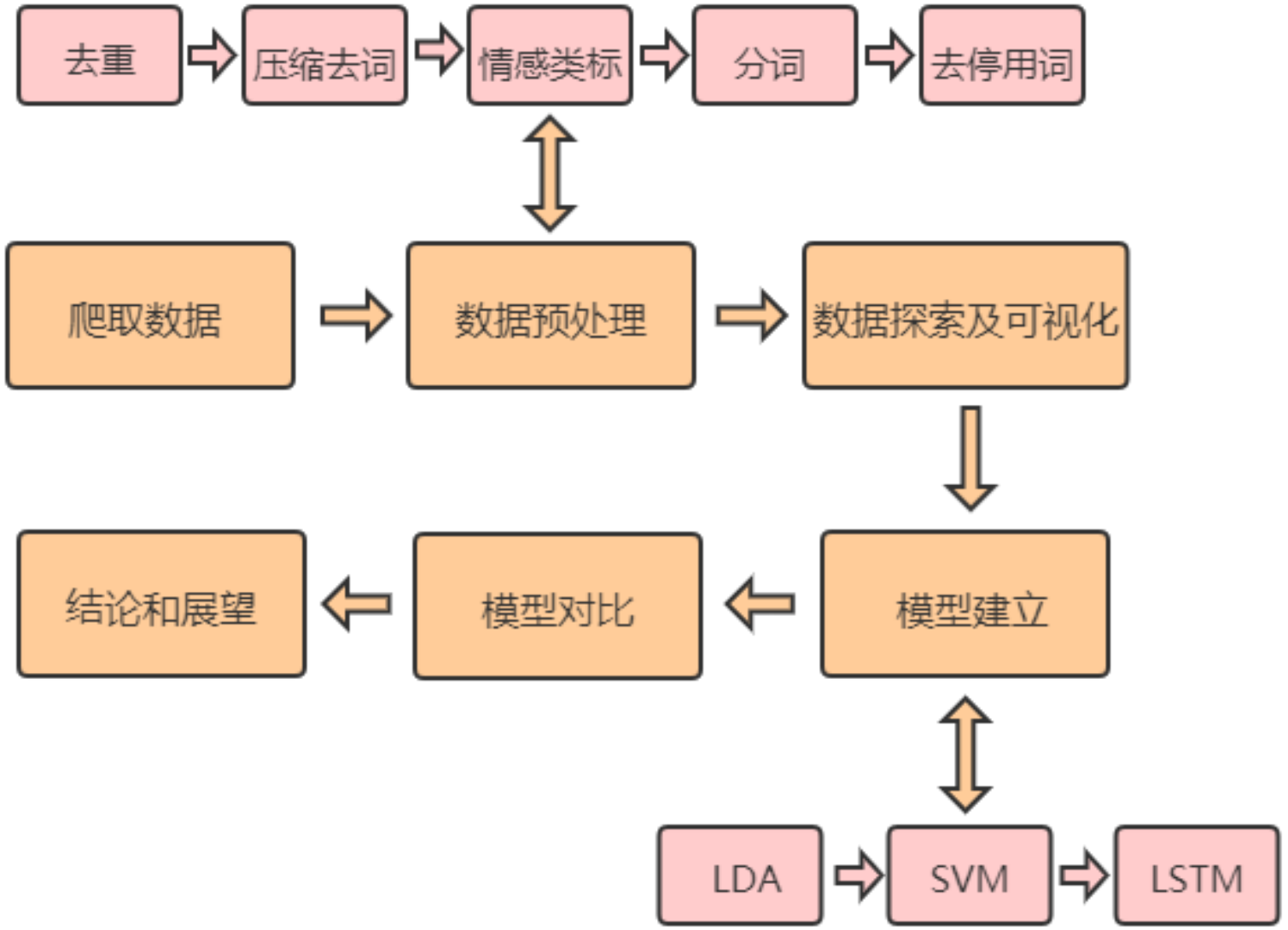
目录

背景和概要

问题背景

随着电子商务的快速发展，网上购物对人们消费模式产生巨大影响。选购商品时，用户的评论信息具有很高的参考价值，应用**自然语言处理**和**文本挖掘**的方法对产品评论进行自动挖掘，在很大程度上可以改善和提升用户体验，**评论情感分析**(sentiment analysis,SA)也因此成为越来越多的研究者和工业界的兴趣焦点。

概要


去重
压缩去词
情感类标
分词
去停用词
爬取数据
数据预处理
数据探索及可视化
结论和展望
模型对比
模型建立
LDA
SVM
LSTM

研究目标和创新点

研究目标

- 分析 iPhone 用户情感倾向
- 从评论文本中挖掘出 iPhone 的优点和不足
- 建立三种模型进行文本情感分析并进行对比

创新点

- 非结构化数据
- 无监督的LDA主题模型，有监督的SVM机器学习模型和深度学习的LSTM模型
- 随机分割和 K-fold Cross Validation K折交叉验证

数据获取与预处理

数据获取

通过调用京东API接口采集到约一万条原始评论数据，数据集大小及评论数据样例内容如下：

```
print(len(data))#数据集大小
```

```
9980
```

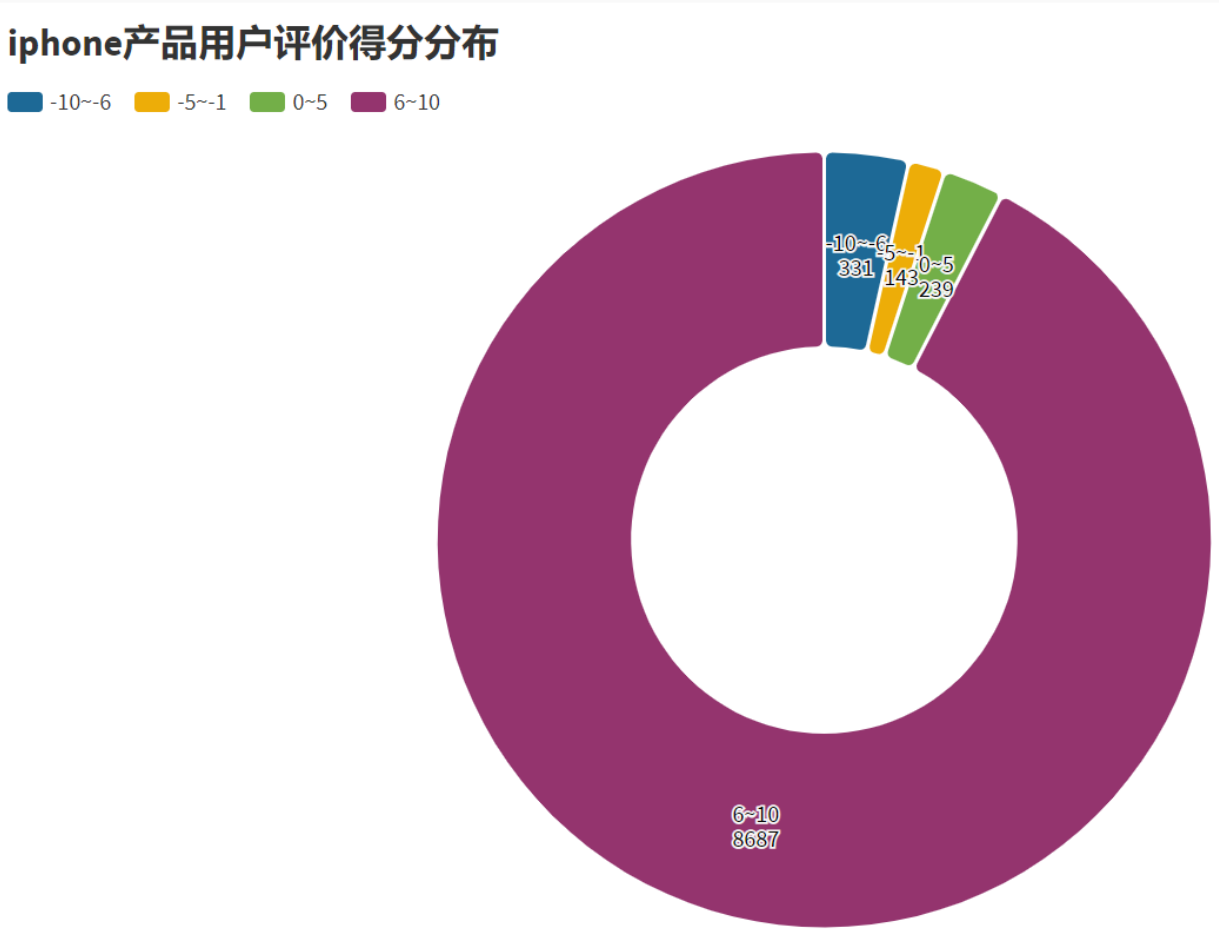
	会员	级别	评价星级	评价内容	时间	点赞数	评论数	商品属性	页面网址	页面标题	采集时间
0	1***p	NaN	star5	京东的快递很快！收到后急忙上手，外观特别漂亮、大气！屏幕音效很好！拍照技术更是无与伦比，效果...	43812.67778	10	17	暗夜绿色 256GB ...	https://item.jd.com/100004770237.html#none	【AppleiPhone 11 Pro Max】Apple iPhone 11 Pro Ma...	0.002025463
1	瓶子丶	PLUS会员	star5	外形外观：这次的背面细磨砂质感非常特别，不容易留下指纹也不怕手汗，颜色主要的一大亮点就是暗夜...	43817.07986	9	0	暗夜绿色 256GB ...	https://item.jd.com/100004770237.html#none	【AppleiPhone 11 Pro Max】Apple iPhone 11 Pro Ma...	0.00202662

数据预处理

	words	label	words_punc	words_cuts	words_cuts_stop	length	space_words
0	买的第二台了，第一台是暗夜绿色。外观方面，三个摄像头外观上吸睛度非常高。iPhone11 m...	0	买的第二台了，第一台是暗夜绿色。外观方面，三个摄像头外观上吸睛度非常高。iPhone11 m...	[买, 的, 第二台, 了, ，, 第一台, 是, 暗夜, 绿色, 。, 外观, 方面, ，, ...	[买, 第二台, 第一台, 暗夜, 绿色, 外观, 三个, 摄像头, 外观, 吸睛, 度, ...	62	买 第二台 第一台 暗夜 绿色 外观 三个 摄像头 外观 吸睛 度 高 iPhone11 m...
1	13号20点抢的，后来想改地址还不能改，不得不让快递小哥帮我改签，晚收到了一天。手机很喜欢，...	0	13号20点抢的，后来想改地址还不能改，不得不让快递小哥帮我改签，晚收到了一天。手机很喜欢，...	[13, 号, 20, 点, 抢, 的, ，, 后来, 想改, 地址, 还, 不能, 改, ...	[13, 号, 20, 点, 抢, 想改, 地址, 改, 快递, 小哥, 帮, 改签, 晚,...	32	13 号 20 点 抢 想改 地址 改 快递 小哥 帮 改签 晚 收到 手机 喜欢 绿色 磨...

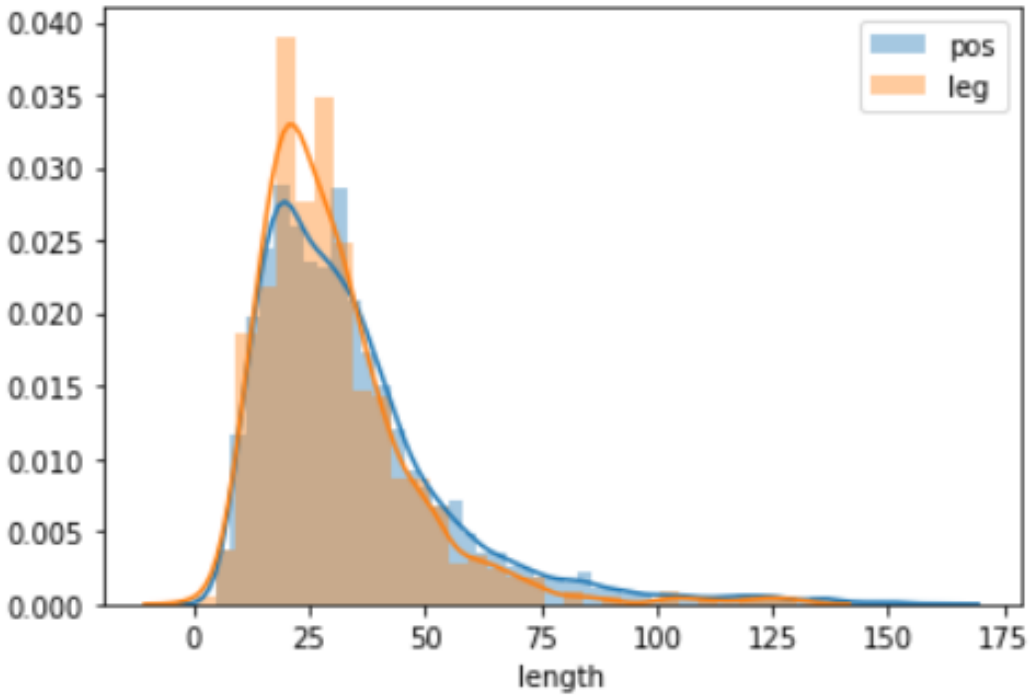
数据探索及可视化

数据探索及可视化

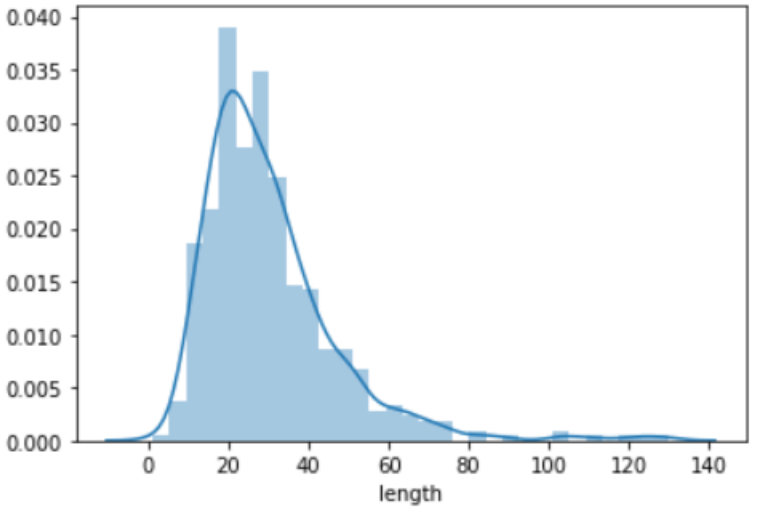


我们根据类标标注的评论得分绘制饼状图，由图可知，用户评论得分集中在6~10分，得分为-5~-1的评论数最少。

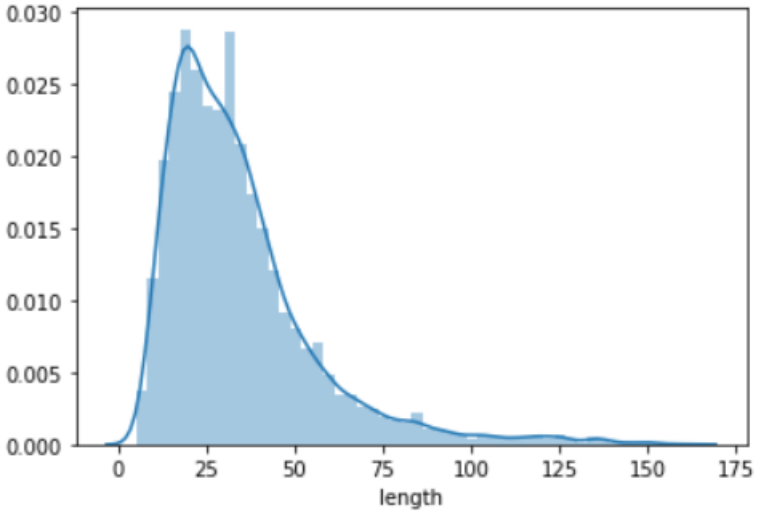
数据探索及可视化



正向，负向语言词数分布情况

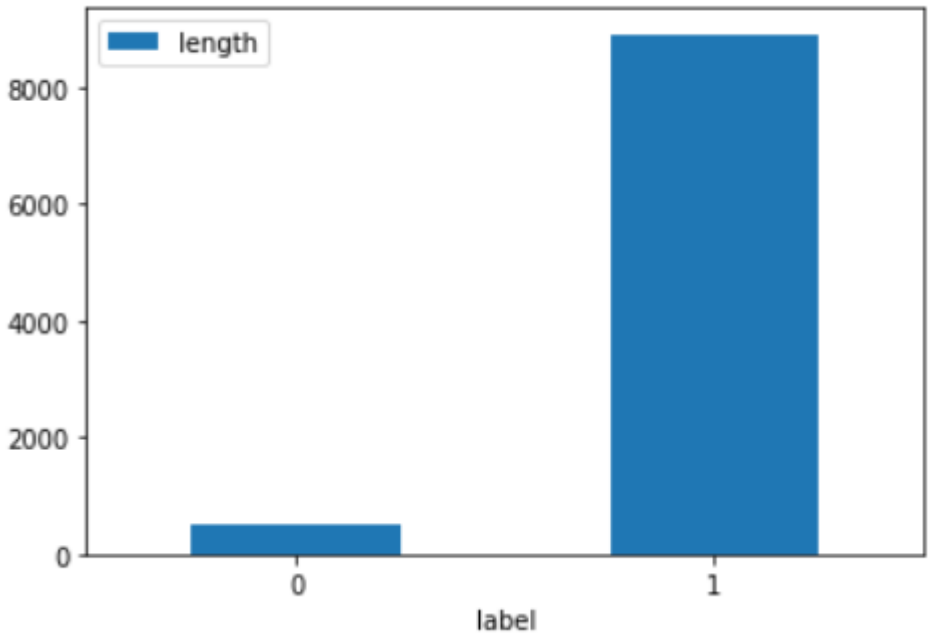


正向语言词数分布情况



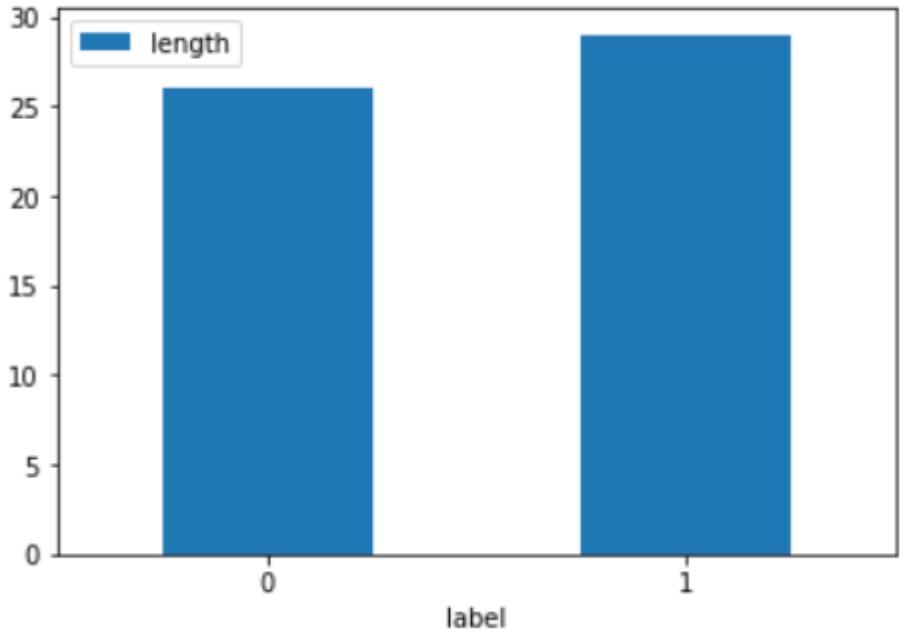
负向语言词数分布情况

数据探索及可视化



正负向语料条数分布情况

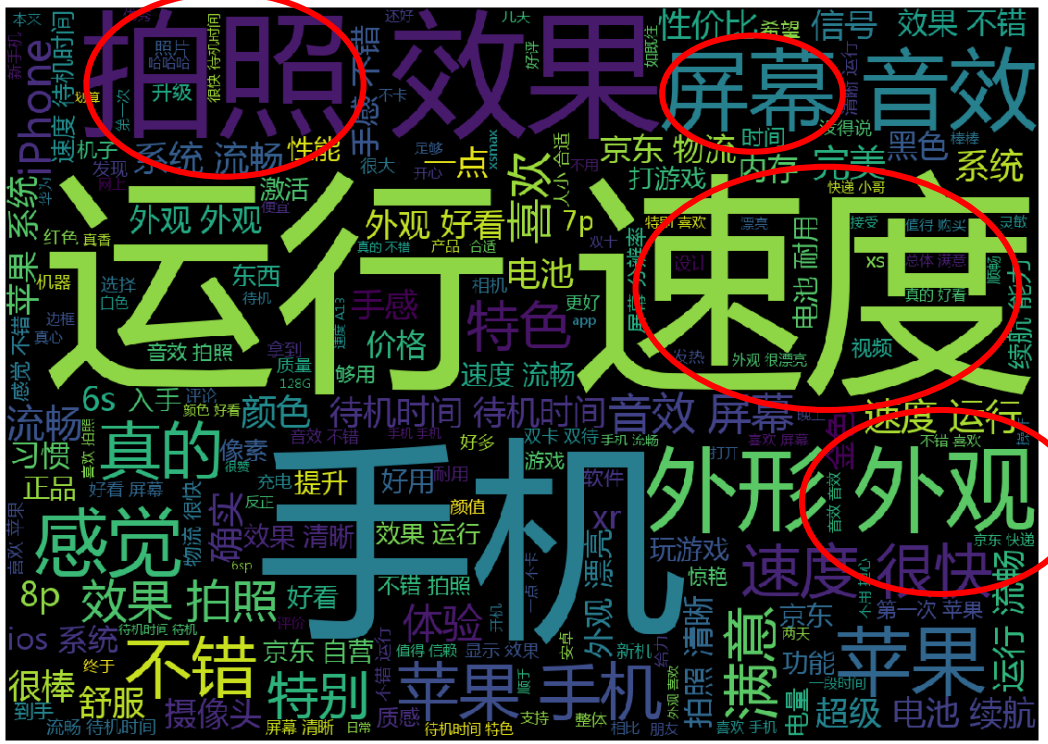
由柱状图可知，正向评论数远多于负向评论数，正向评论数约9000条而负向评论数约500条。



正负向预料长度中位数分布

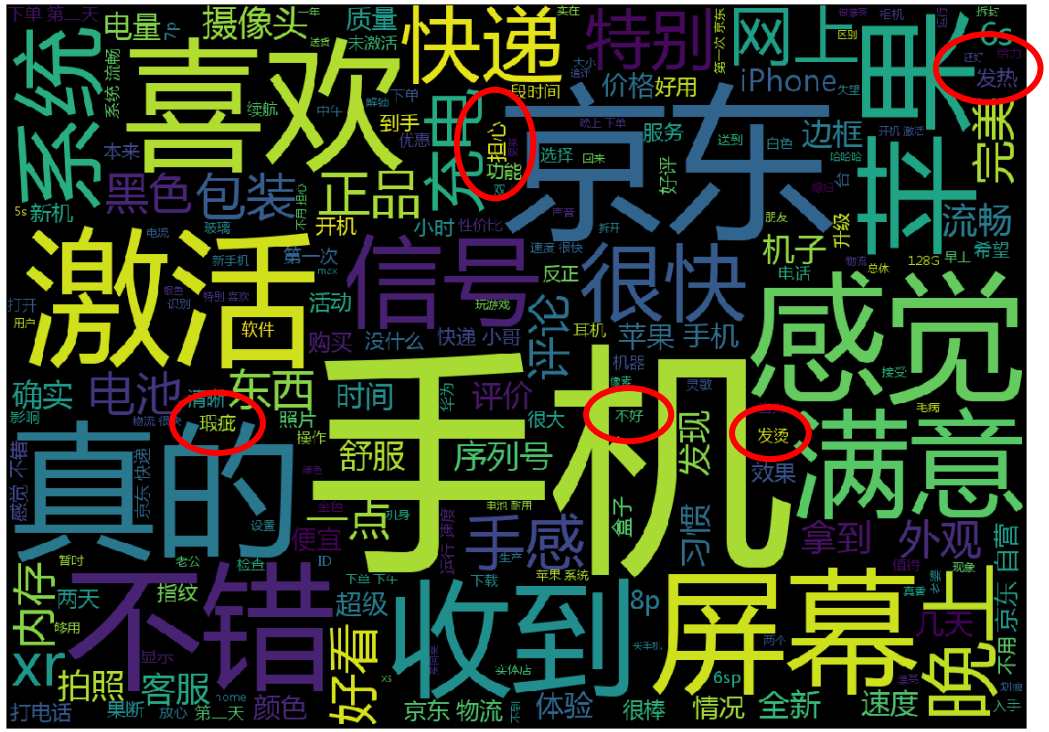
由柱状图可知情感倾向正负向语料长度中位数分布情况，看起来评价正面的话比较多。

数据探索及可视化



正向语料词云

速度 拍照 屏幕 外观



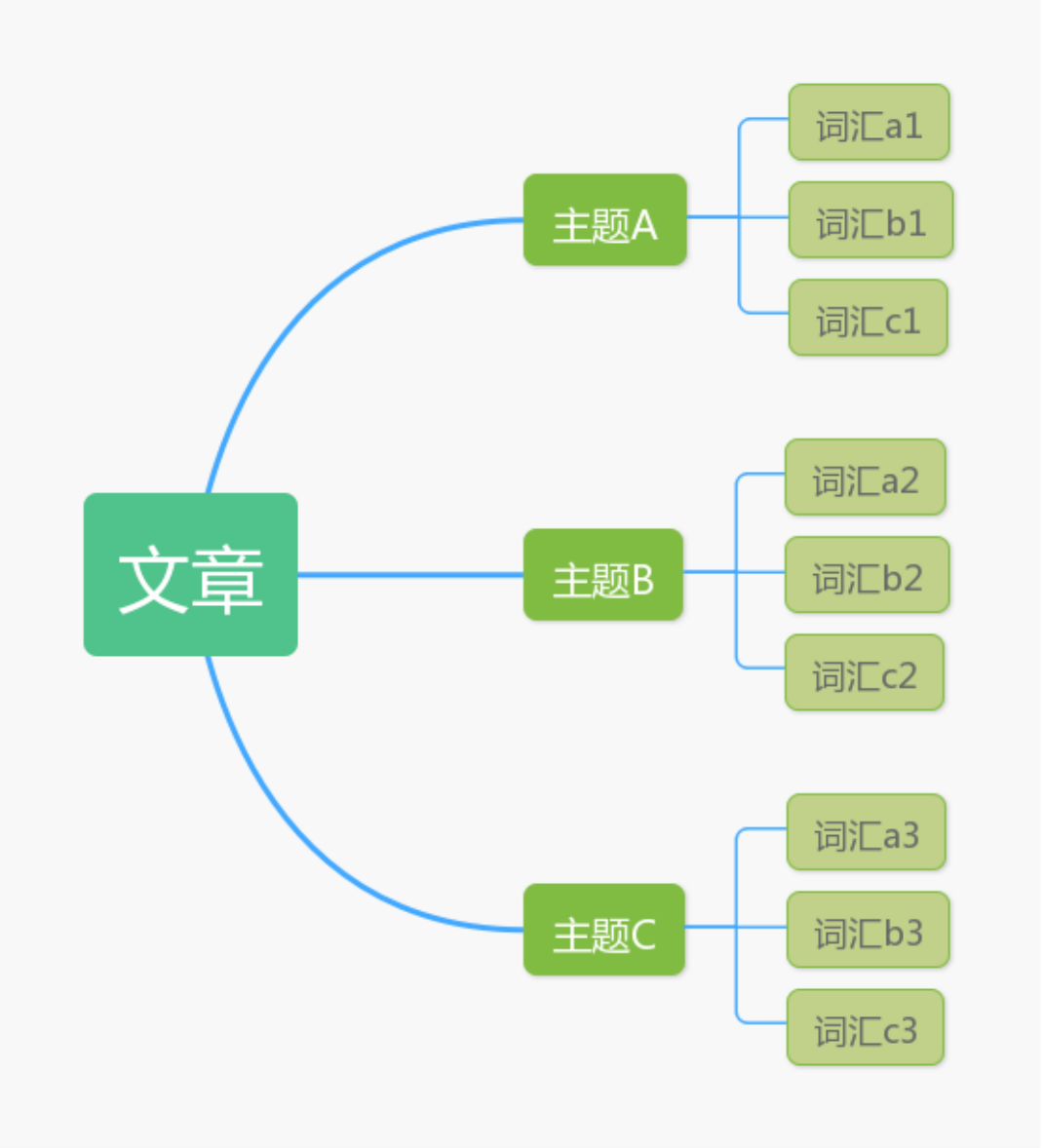
负向语料词云

瑕疵 担心 发烫 发热 不好

模型建立及对比

基于 LDA 主题模型的情感分析

- LDA定义两个分布：**主题与词汇分布，文章与主题分布**。
- 经过 LDA 主题分析之后，评论文本被聚为**3个主题**，每个主题下面生成若干最有可能出现的**词语**以及**相应频率**。



基于 LDA 主题模型的情感分析

下表展示了正面评价文本中的潜在主题

主题1	主题2	主题3
效果	手机	手机
运行	618	苹果
拍照	苹果	买
速度	兔	不错
屏幕	年	喜欢
时间	妈	速度
待机	键	京东
外观	128g	流畅
机时	32G	屏幕

- 主题1的高频特征词反映：iPhone 拍照效果好，待机时间长，外观好看。
- 主题2的高频特征词，关注点是：兔年，妈妈，128G，32G，反映 iPhone 可以作为新年礼物，iPhone 的内存空间也经常被关注。
- 主题3的高频特征词，关注点是：流畅，速度，屏幕，反映 iPhone 使用流畅，速度快等。

基于 LDA 主题模型的情感分析

下表展示了负面评价文本中的潜在主题

主题1	主题2	主题3
手机	手机	手机
买	买	买
京东	京东	京东
激活	激活	喜欢
喜欢	说	感觉
感觉	苹果	激活
真的	不错	说
说	感觉	苹果
快递	屏幕	真的

- 高频特征词除了买，喜欢，不错，感觉，真的等情感词，提到了激活，快递等，主要反映：激活流程复杂，快递速度慢等问题。

基于有监督学习的情感分析

抽取特征

TF-IDF是一种用于信息检索与数据挖掘的常用加权技术，常用于挖掘文章中的**关键词。**

词频　　逆文档频率

$$\text{词频(TF)} = \text{某个词在文章中的出现次数}$$

$$\Downarrow$$

$$\text{词频(TF)} = \frac{\text{某个词在文章中的出现次数}}{\text{文章的总词数}}$$

$$\Downarrow$$

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1}\right)$$

$$\Downarrow$$

$$\text{TF} - \text{IDF} = \text{词频(TF)} \times \text{逆文档频率（IDF）}$$

基于有监督学习的情感分析

- k折-交叉验证分类结果

精确率(Precision)
召回率(Recall)

```
[0.93426916 0.78141392 0.72652455 0.72849755 0.78667665]
[0.5541747  0.55249026 0.56537435 0.60273112 0.62888809]
```

- 随机分割分类结果

准确率
(Accuracy)

```
0.9410201912858661
```

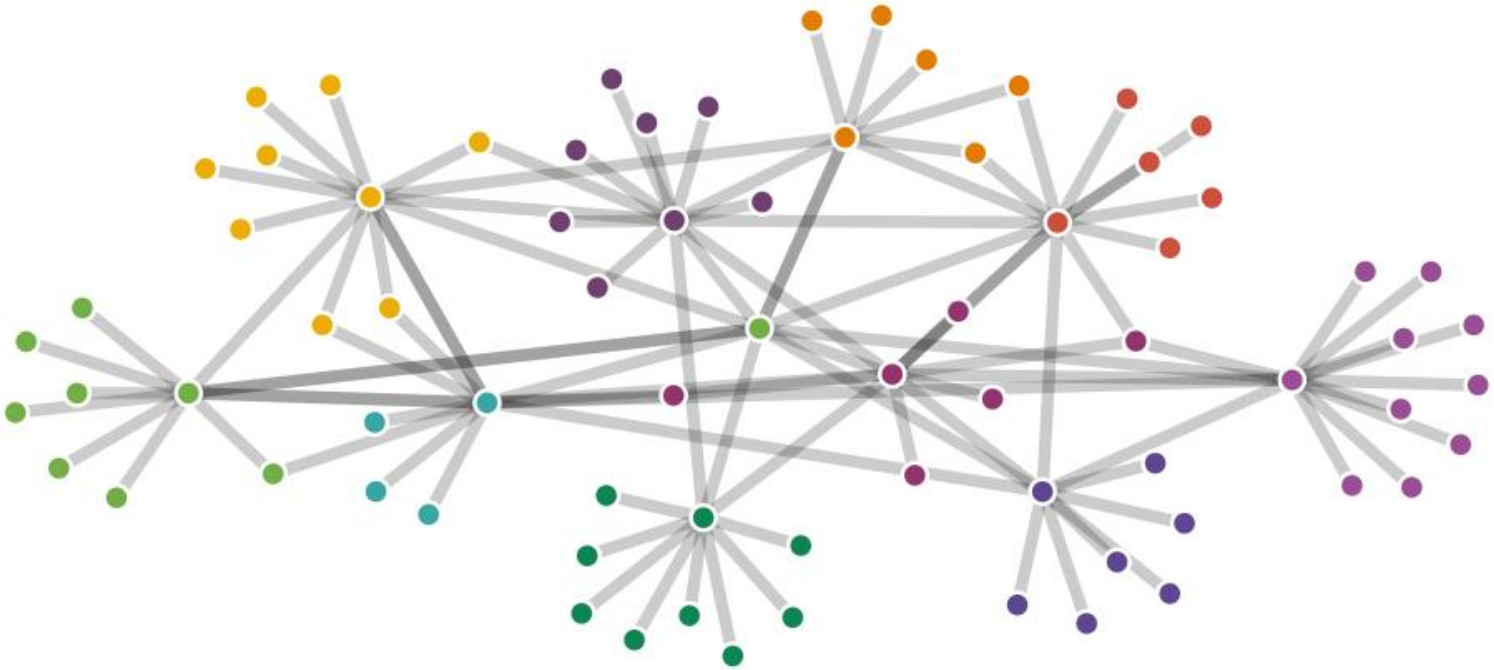
TF-IDF的优点是简单快速，而且容易理解。

缺点这种计算**无法体现位置信息**，无法体现词在**上下文**的重要性。

可以使用**word2vec算法**进行改进。

基于 LSTM 深度学习的情感分析

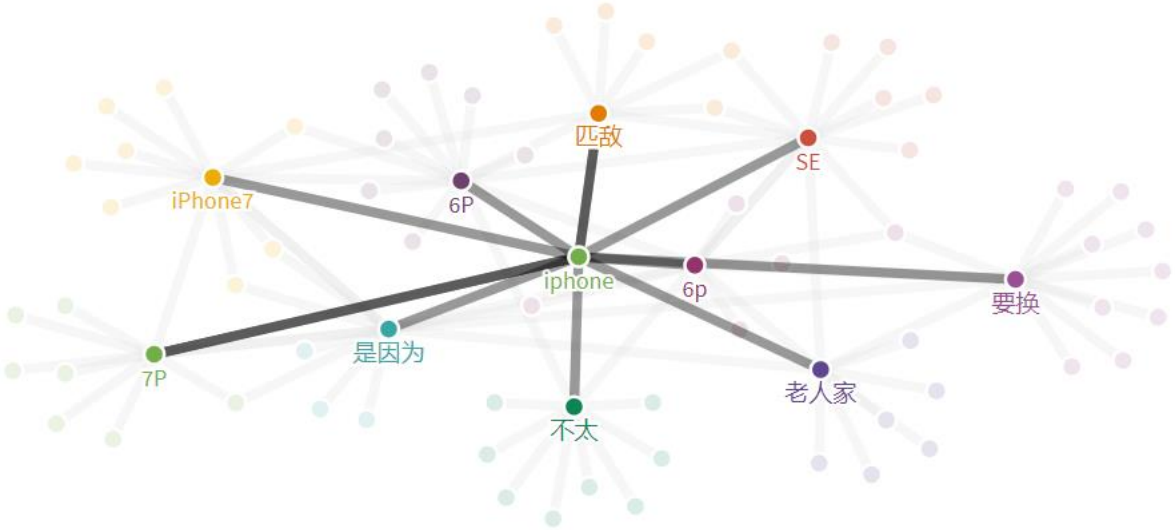
- 利用 gensim 中 Word2vec 工具，完成**词向量**转换工作，将处理好的数据导入 **LSTM** 模型中迭代100次，观察 LOSS 变化情况，并预测其情感倾向分类。



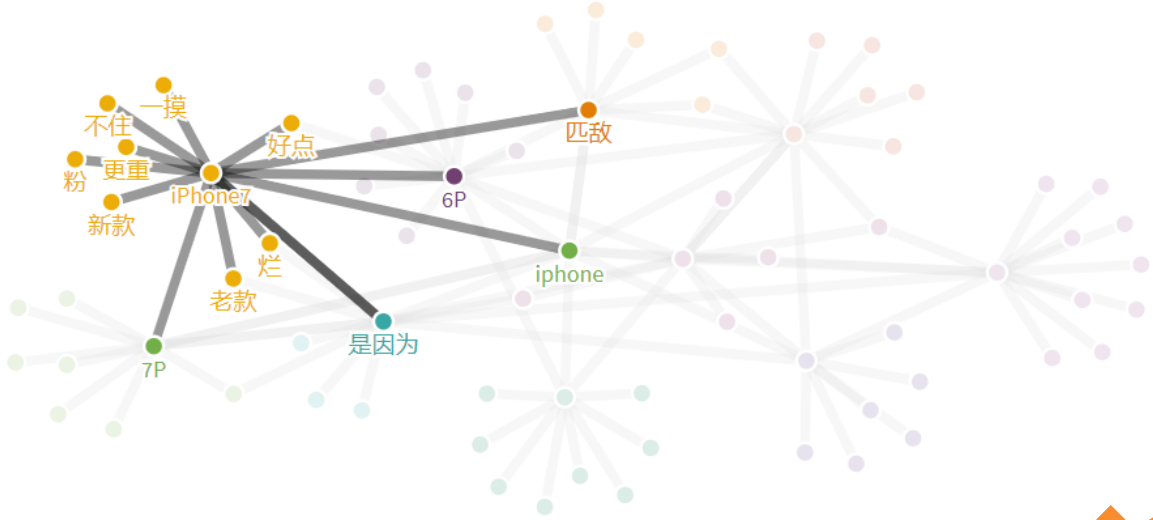
语义空间上最接近词语的 network charts

基于 LSTM 深度学习的情感分析

- 将词语完成**词向量**转化后，找到语义空间上与 **iphone** 最接近的十个词：**iPhone7，7P，6p，是因为，匹配，老人家，不太，6P，SE，要换**。并接着找分别与十个词在语义空间上最接近的十个词，画出 network charts 网络图，展现语义空间上不同词之间的关系。

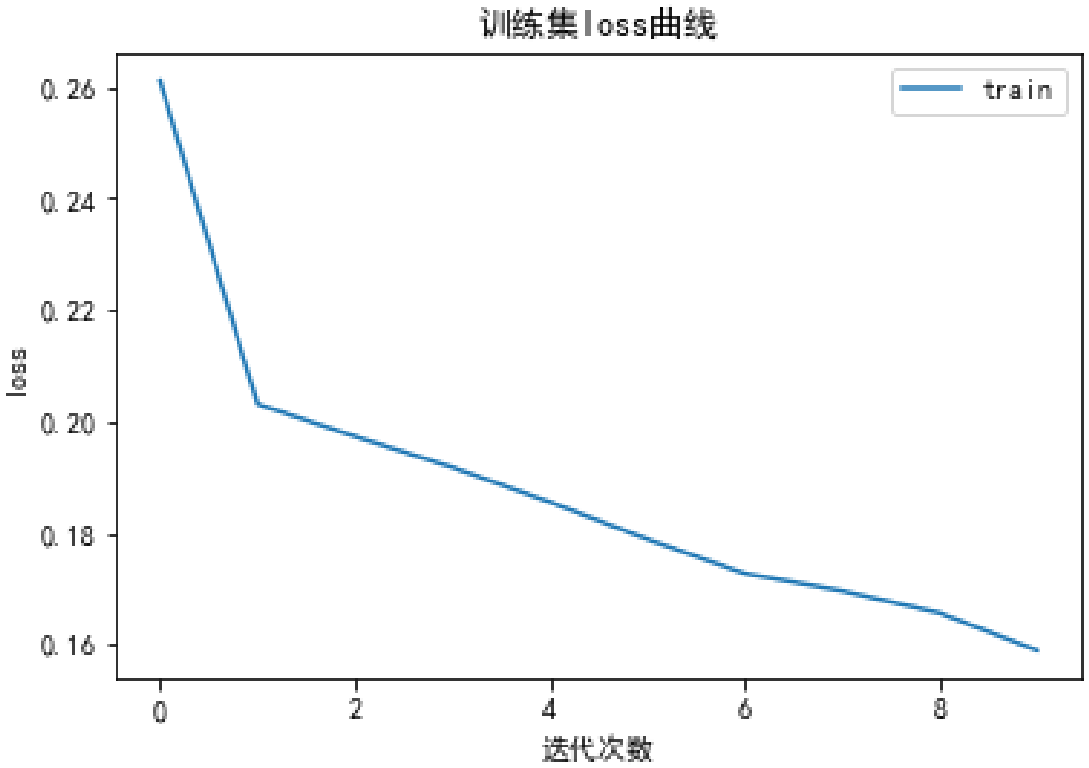


与 iphone 语义最接近词语的 network charts



与 iPhone7 语义最接近词语的 network charts

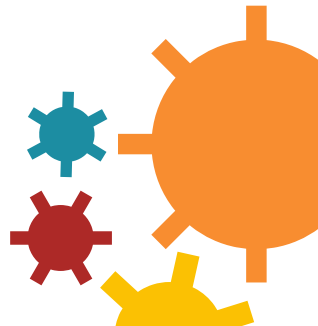
基于 LSTM 深度学习的情感分析



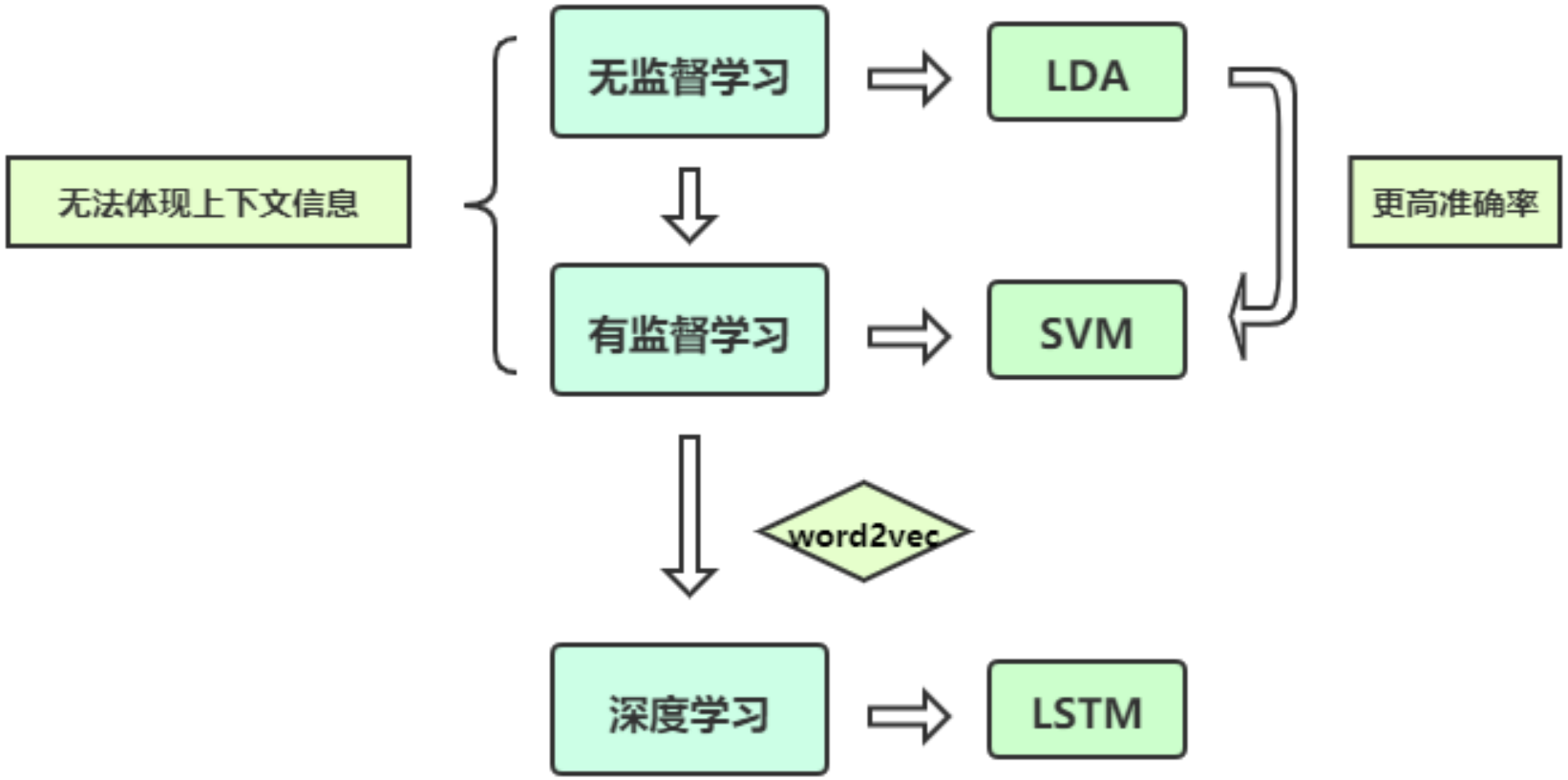
准确率结果是：

```
0.9495217853347503
```

采用基于 RNN 的优化算法——**LSTM 长短期记忆网络**。LSTM 长短期记忆网络采用一套灵活的逻辑——“只保留长序列数据中的**重要信息**，忽略不重要信息”。



模型对比



经过三种模型的对比验证，我们得到采用**LSTM 长短期记忆网络**进行文本挖掘，情感分析的**准确度最高**。

总结与展望

结论

- iPhone 产品的用户情感偏向大多为**正向**，其优势在于： 使用流畅，速度快，拍照效果好，外观好看等。劣势在于：激活流程复杂，发热等问题。

- 我们所建立的三种情感分析模型：无监督的LDA主题模型，有监督的SVM机器学习模型和深度学习的LSTM模型，通过检验发现：**LSTM 长短期记忆网络**模型准确率最高。

改进

- 1. 爬取的**数据量**不够多，增大数据量可以使模型更加完善，准确率也会提高。
- 2. 对评论进行类标标注的时候，**人工标注**的准确率会高于机器标注，但会大大提高人工成本。
- 3. 如果想进一步提高准确率，**停用词词表**需要进一步补充整理；建立 word2vec 的时候，语句长度，输出维度**参数**等可以进一步调参试试。

THANKS