复旦大学计算机科学技术学院

2018-2019 学年第一学期期末论文课程评分表

课程名称： **自然语言处理** **课程代码：** **COMP130141.01**

开课院系： 计算机科学技术学院

学生姓名： 于泳欣 **学号：** 16307130307 **专业：** 计算机科学与技术

论文名称：“云中谁寄锦书来”——基于复旦大学表白墙等的文本分析和情书生成

（以上由学生填写）

成绩：

论文评语（教师填写）：   任课教师签名： 日　　期：

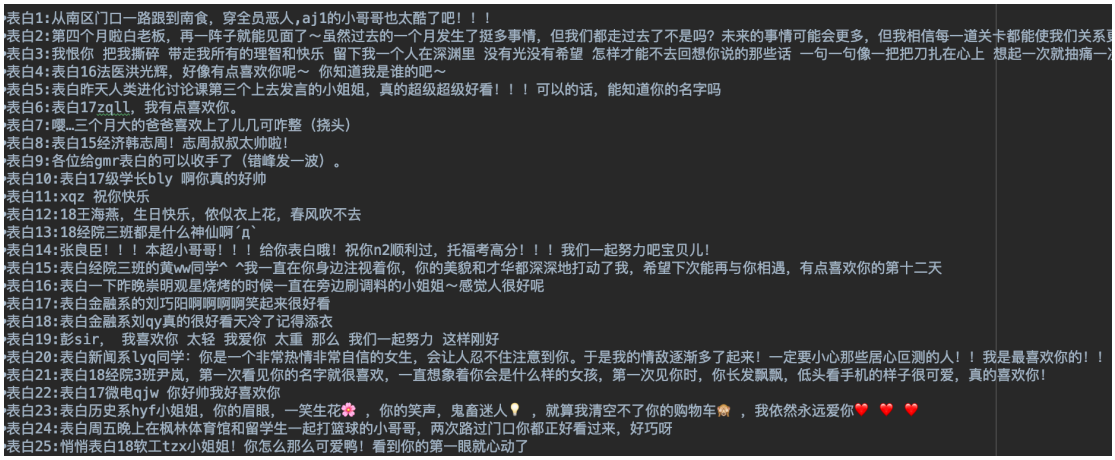
1. 背景介绍：

从古至今，人们都习惯用诗文寄托相思之意。思念一个人时有“悠哉悠哉，辗转反侧”，亦有“思君如满月，夜夜减清辉”。踏入网络时代，人们有了更多现代化的通讯设备，情书越来越少地出现在人们的日常生活中，取而代之的是成百上千条短信、微信。我注意到校园里有表白墙这一个平台，供同学们投稿抒发自己的情感，也作为一个窗口来了解当代大学生的感情生活。所以这次期末项目选择基于表白墙的文本分析和自动生成文本。

2. 数据集来源及分析

2.1 数据集来源

因为数据集是复旦微生活公众号的推文，所以选择在对于公众号文章搜索表现较好的搜狗微信搜索上爬取推文，共爬取一百篇文章，时间范围为本学期，即 2018 年 9 月之后。



观察爬取到的推文，发现一些特点：

- 关于表白对象。大部分发表表白的同学比较腼腆，会选择用名字的拼音缩写（图片中的表白 6，表白 9 等），或者没有明确的对象，只能对对方的形象特征进行描写，这种属于天知地知你知我知，不能定位到具体的年级（表白 1 和表白 5 等）。
- 语料比较口语化。因为现在网络通信的发达，发表者将表白墙只作为一次普通的短信发表，并没有字斟句酌，呈现出的语料也显得口语化，不规范，语气词较多。（图中最规范的表白 2 和表白 3，和正规的出版物相比也只能算语句通顺）
- 使用表情。有的是图形的表情，有的是用语言表达的表情（比如表白 7：挠头）。在处理的时候应该考虑这一点进行清洗。

2.2 数据清洗

首先将爬取到的语料全部组合成一个文件，使用结巴分词进行初步的分词处理和停用词处理。这里的停用词库使用的是哈工大版本。

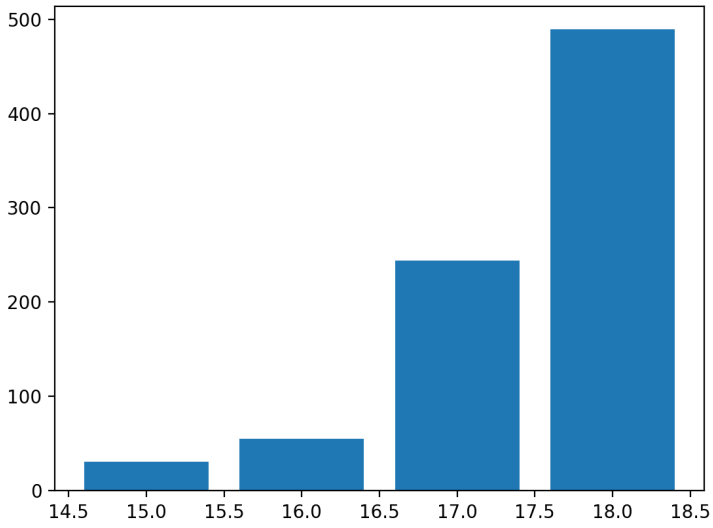
```
表白 1 : 表白 数院 某 大神 ， 怎么 有 这样 温和 宽容 有趣 深刻 谦虚有礼 细致 谨慎 自制 的 人 ！ 你 这个 芳心 收割机 ， 到底
表白 2 : 表白 化学系 wjy 助教 小姐姐 ！ 小姐姐 真是 我 入学 以来 碰见 的 最 漂亮 最 和蔼 的 助教 啦 吹 爆 小姐姐 ！ 祝 小
表白 3 : 期末 季完 了 ， 求 对象 ！
表白 4 : to 18 自科 4gy ～ 可能 是 心里 比较 乱 吧 ， 之前 一直 没有 想过 要 怎么 和 你 相处 ， 所以 在 墙 墙上 贴 了
表白 5 : 送给 我 拼图 的 那位 谢谢 啦 ！ ！ 祝 你 早日 脱单 ， 绩点 全 A ！ ！
表白 6 : 表白 昨天 表白 十四 ！ 国 足 加油 [illegible] 复旦 旁友 看 亚洲杯 进 qq 群 吧 645516108
表白 7 : 表白 心中 小漫 ， 所有 失去 的 东西 都 会 以 另 一种 方式 回来
表白 8 : 在 【 文图 一楼 书架 】 拿 了 《 视差 》 《 有趣 的 时代 》 《 美国政府 对 中国 国家 形象 认知 》 的 天使 ， 拜托
表白 9 : 17 国务 dsw 哥哥 你 今天 是 彩虹 彩色 条状 绚丽多彩
表白 10 : 回 1.9 表白 2 在线 不回 你 就是 不 礼貌 吗 ？ 也许 有事 呢 ？ 也许 ， 你 的 消息 对 他 来说 就是 一种 打扰
表白 11 : 回 昨天 表白 19 ， zan 同加 一
表白 12 : 表白 苏文 臭文爱 你
表白 13 : 表白 我 闺蜜 仙女 小姐姐 q1394855273 期末 加油 ， 愿 美好 的 你 邂逅 所有 美丽
表白 14 : 表白 于 大宝 国足 牛 逼 [illegible]
表白 15 : sjc 哥哥 ， 考完试 你 有 时间 么
表白 16 : 女生 都 不 知道 在线 不回 男生 消息 是 不 礼貌 的 嘛 （ 手动 狗头 昨日 第二条
表白 17 : 我 不管 ， 双 马尾 的 xc 同学 最 可爱 ！
表白 18 : 回复 昨天 表白 19 : 超级 支持 你 ！ ！ ！ ！ ！ ！
表白 19 : 在电 宝救 我 狗命 表白 ybc
```

进行词频统计后发现还是有一部分无意义的词，手动更新停用词库再分析结果。

2.3 文本分析

2.3.1 关于年级

考虑在校的 2015-2018 级，一般会有 16 和 2016 两种形式，需要将二者加和，而且注意在表白的序号处也有相应的数字，因为共 100 天的数据，所以可以简单的减去 10 作为最终的数据。



结果显然随着年级的上升，被表白的同学数量也越来越少。刚入学的 2018 级新生对于爱情总是更为积极一些，也可能是学长学姐们已经更为成熟，选择了私下里内敛地表达。

2.3.2 关于专业

对于专业的表达方式更加多种多样，尤其是复旦采取大类招生，细分专业的培养方案，如计算机微电子等学院还有缩写 CS, ME 等，让专业名更加难以统计。总结下来有以下几种：

- 直呼其名：如计算机学院 2016 级通常被称为 16 计算机（然而好像没看到有人跟 16 计算机的同学表白）
- 学院缩写：依然以计算机学院为例，我们都是计院，计科的一份子，这种通常可以用 X 院来匹配
- 英文缩写：如上文所述的 CS, ME，以及电工的 EE，软工的 SS 等
- 大类招生：多见于大一同学，技术科学试验班简称技科，自然科学试验班简称自科，可以用 X 科来匹配

结果：

0	自科	142
1	经院	89
2	管院	79
3	技科	50
4	新闻	43
5	临床	41
6	社科	40
7	物理	39
8	药学	37
9	公卫	33

10	生命	30
11	微电子	29
12	法学	28
13	历史系	27
14	预防	25
15	中文	15
16	软工	15
17	外院	12
18	国务	11
19	微电	9

20	化学系	9
21	历史	7
22	电工	7
23	化学	6
24	护理	6
25	高分子	5
26	材料	5
27	金融	5
28	计算机	5

自科一骑绝尘，以绝对优势成为最受欢迎的专业，经管双雄紧随其后。比较意外的是以为会很受欢迎的中文和外院名次都不高，可能因为人数不占优势。正如上面提到的，自科、技科、社科都是大一同学特有的专业，名词都很靠前，这也暗合了我们分析的新生发表的表白数目较多。除去榜中的这些专业，其实还有很多专业，但因为抓取到的次数太少，不具代表性，就去除掉了。

2.3.3 关于地点

表白一个矮矮的扎马尾的女生，好几次在食堂看到你，好像是新闻学院的日本留学生？

今天文图二楼自习室156号桌穿红色校名棒球服的小哥哥也太可爱了叭！眼睛里有小星星★！！全程偷瞄

一见钟情的戏码总是让人难忘，表白墙上也有很多同学试图通过这种方式抓住一闪而过的缘分。根据经验总结出现频率较高的地点分别为图书馆、食堂、宿舍，讨论课上通过自己的发言吸引粉丝的也大有人在。所以同样建立了一张表来统计词频。

图书馆	文图	22
总计：47	图书馆	19
	理图	6
宿舍	南区	22
总计：39	北区	11
	东区	8
教学楼	三教	17
总计：37	二教	6
	六教	5
	五教	5
	四教	4
食堂	食堂	6
总计：15	南食	5
	春晖	4
校区	枫林	12
总计：26	张江	9
	江湾	5
运动	篮球场	14

图书馆依然是最适合邂逅的地方，文图又以其座位多，较为方便拔得头筹。文图的座位设计和其他几个图书馆相比，间隔也比较小，不知道和这一点有没有关系。教学楼中三教没有意外成为第一名。三教中经常会安排通识教育课、政治课等多专业共同研讨的课程，通宵自习室的存在也赋予它其他教学楼没有的美丽。另外意外的发现篮球场这一词语出现频率颇高，或许大家应该考虑多去篮球场走一走。

2.3.4 心动词语

众所周知，情书中通常会有一些词语来形容心仪的对象。比较好奇大家会倾向于在表白墙中使用怎样的词语，所以采用了词云生成，背景设定为一个可爱的女孩子，也去除了一些表意性不强的比如“哥哥”、“小姐姐”等词。可以看出“可爱”是排名最高的形容词，和我的想法还是很契合的，因为喜欢一个人不管她漂不漂亮，聪不聪明，在你的心里总归是可爱的。

可爱这个词也带有一些偏爱的意味，更具包容性。总的来说大家使用过的词语还是很积极向上的，怀有“希望”，想在“一起”，把握“现在”和“今天”，也期待“以后”的“生活”。

3. 基于深度学习的情书自动生成

3.1 为什么选择文本生成

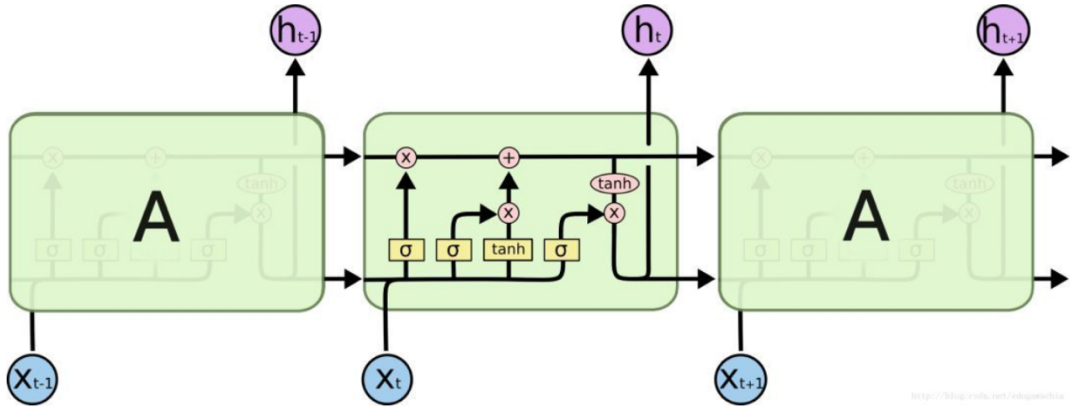
表白墙本身的性质不同于微博等，它大部分数据的情感倾向都是积极的，比较单一，针对表白墙来做情感分析的意义不是很大。而摘要生成这一方面，需要语料本身有一定的长度，比如人民网上的一篇新闻或者美团上的一个点评，这样方便于关键词提取排序等。但是表白墙每一条的内容已经足够短小，再去提取关键词就很刻意。另外的问答系统也考虑过，感觉不是很适合，所以最后还是选择了文本生成，希望能做出来一些有意思的东西。

3.2 核心思想

RNN 模型中，当前节点的输出值由当前的输入和上一个节点的输出两部分共同决定。使其很适合处理上下文相关的问题，目前的文本自动生成任务也大多采用了 RNN 及其变种模型。我在这学期学习了 RNN 的 LSTM 和 GRU 模型，在这次的 PJ 中也分别采用了这两种模型，既作为一次练习，也是在实践中提高的过程。

3.3 LSTM

文本的上下文存在着序列关系，所以可以使用基于概率的模型，根据输入的数据预测下一个可能出现的文本。但是 RNN 模型的一个缺点是处于不可避免的指数爆炸，对长时记忆的能力比较弱，如果我们输入一本完整的文档，最后输出的可能只是文档最后的部分。而 LSTM 引入了门的设计，对长时信息和短时信息分别对待，使长信息不至于被短信息所埋没。



如图 LSTM 中引入了被称为细胞状态的记忆向量 c(t)，使用了输入门 i(t)、遗忘门 f(t)、输出门 o(t)等控制遗忘门 h(t)和 c(t)的更新。公式如下

$$\boldsymbol{i}^{(t)} = \sigma(\boldsymbol{W}_i \boldsymbol{h}^{(t-1)} + \boldsymbol{V}_i \boldsymbol{x}^{(t)})$$

$$\boldsymbol{f}^{(t)} = \sigma(\boldsymbol{W}_f \boldsymbol{h}^{(t-1)} + \boldsymbol{V}_f \boldsymbol{x}^{(t)})$$

$$\boldsymbol{o}^{(t)} = \sigma(\boldsymbol{W}_o \boldsymbol{h}^{(t-1)} + \boldsymbol{V}_o \boldsymbol{x}^{(t)})$$

而细胞状态和隐藏状态也通过公式计算得出。

$$\tilde{\boldsymbol{c}}^{(t)} = \tanh(\boldsymbol{W}_c \boldsymbol{h}^{(t-1)} + \boldsymbol{V}_c \boldsymbol{x}^{(t)})$$
$$\boldsymbol{c}^{(t)} = \boldsymbol{f}^{(t)} \odot \boldsymbol{c}^{(t-1)} + \boldsymbol{i}^{(t)} \odot \tilde{\boldsymbol{c}}^{(t)}$$
$$\boldsymbol{h}^{(t)} = \boldsymbol{o}^{(t)} \odot \boldsymbol{c}^{(t)}$$

显然 LSTM 能够记忆更长时间之前的信息，从而对文本的预测有更好的表现。

LSTM 模型中的输入使用字符级，即每个字符对应一个 one-hot 的词向量。但是汉字不同于英文字母只有 26 个，使用字符级的向量依然是比较大的开销，one-hot 本身也存在高维稀疏的问题。搭建神经网络使用的是 LSTM 层+全连接层的方式

```
# model
model = keras.models.Sequential()
model.add(layers.LSTM(256, input_shape=(length, len(chars))))
model.add(layers.Dense(len(chars), activation='softmax'))

optimizer = keras.optimizers.RMSprop(lr=1e-3)
model.compile(loss='categorical_crossentropy', optimizer=optimizer)
model.fit(x, y, epochs=100, batch_size=1024, verbose=2)
```

在之前的作业中，文本生成的概率分布是固定的，我们也会选择概率最大的词，所以一个确定的语句之后所接的内容通是确定的，经常会有重复的段落出现，比较无趣。在这个模型中我引入了 temperature 机制采样。通过设定一个初始的 temperature 值，改变文本的概率分布，并增强文本选择的随机性。

$$p(x_{new}) = \frac{e^{\log(p(x_i)) / temperature}}{\sum_i e^{\log(p(x_i)) / temperature}}$$

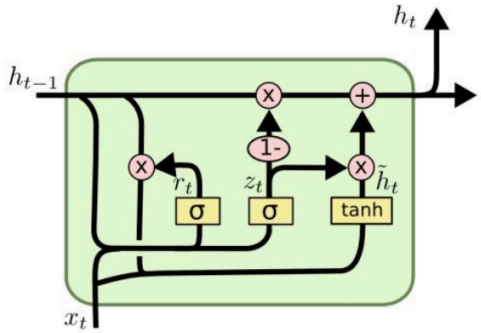
这里需要指出的是 temperature 越大，概率分布越均匀，随机性也越大，实际上是会对我们生成的文本的可读性造成一定的影响。

总结一下我们的文本预测过程：

- 通过已有的文本计算出字符的概率分布
- Temperature 机制得到新的字符概率分布
- 在新的概率分布下采样得到下一个字符
- 生成的新字符添加到最后，去掉原文本的第一个字符。比较类似于滑动窗口，窗口的大小是确定的，移动窗口决定当前框起来的文本。

3.4 GRU

GRU 是在 2014 年提出的将 LSTM 的单元状态和隐层状态合并，并做了一些改动的 RNN 辩题模型。



与 LSTM 模型不同，GRU 只有两个门，重置门 r(t)和更新门 z(t)。重置门控制是否重置，即多大程度擦除之前的状态；更新门则表示更新 hidden layer 的程度。重置门的值越小说明忽略的前一时刻的隐藏层信息更多，更新门的值越大说明前一时刻的隐层输出对当前隐层的影响越大。计算他们的公式如下：

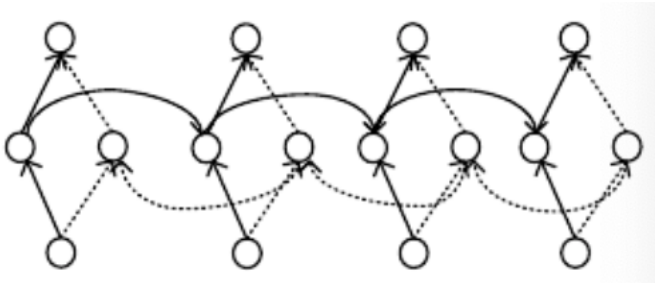
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

与 LSTM 相比，GRU 的参数比较少，结构相对比较简单。在查找资料的时候看到了双向 RNN，可以在一定程度上解决 RNN 的序列性问题，所以在 GRU 的模型中尝试了一下 BiGRU，权做了解。BiGRU 由两个单向的、方向相反的 GRU 组成，输入由两个单向的 GRU 共同决定。



另外为了与 LSTM 模型区分，在 GRU 模型中引入了 word embedding 词嵌入模型，选择了词语级别的输入向量，即先用 jieba 分词，将分词后的数据集输入到我们的模型中，这样兼顾了字与字之间近邻的联系。和 one-hot 相比，词嵌入也能节省一部分的内存空间。

从最终的结果来看，在模型的表现上二者的差异不是很大，可能因为核心的思想都是共同的。

4. 模型评价

对于文本生成的评价历来是比较困难的一个问题，BLEU 是很多研究者选择采用的评价策略，适用于文本翻译等，但是鉴于这次的语料不很规范，所以只采用了人工评价的方法，从流畅性、关联性、表意性几方面对生成的文本进行评价。

4.1 表白墙文本生成

这个语料库生成的文本质量不是很高，意义连贯的文本比较少，推测原因除了模型本身的问题外，与表白墙语料库本身质量不高也有一些关系。表白墙中收集了不同人发表的内容，在句法结构上相似度比较小。

生成文本范例

- （看见你就很开心）

  看见你就很开心，就会继续默默许愿你好呀错过，超可爱的说哈哈哈哈，不敢生怕，做彼此的小姐姐，知道我是谁。

- （你笑起来好美）

  你笑起来好美，喜欢怎么办，在自习就在你身边，表白考试努力你愿意，包容受欢迎是天使，不要习题第一排。

可以看出流畅性不是很高，关联性和表意性尚可，一些词句靠猜测能连贯起来，但也有不知所云的部分，总体来说不理想。

4.2 书信集文本生成

因为对这个语料库不是很有信心，所以又尝试了自己比较喜欢的两本情书集，分别是王小波的《爱你就像爱生命》和朱生豪的《朱生豪情书全集》。针对一个作家的作品做文本生成，通常结果会更好一些，因为在遣词造句和连贯性上，同一个人的作品是有迹可循的，个人风格也会更加明显。

4.2.1《爱你就像爱生命》语言风格举例

当我跨过沉沦的一切，向着永恒开战的时候，你是我的军旗。

别怕美好的一切消失，咱们先来让它存在。还有一个美好的东西不会消失，就是菩提树。真

希望你是我的菩提树，我愿做你的菩提树，你知道歌里是怎么唱吗？如今我远离故乡，已经有许多年，我仍然听到呼唤，到这里寻找安谧。灵魂是活生生的，它的安慰才能使人满足。

- (我不喜欢稀里糊涂地过日子)

  我不喜欢稀里糊涂地过日子，我是这样，我喜欢你的看法，你和我讲话都是最迷人的。还有我和你…

4.2.2 《朱生豪情书全集》举例

也许真会有那么海阔天空的一天，我们大家都梦想着的一天！我们不都是自由的渴慕者吗？为了你，我也有走向光明的热望，世界不会于我太寂寞。

你说我们前生是不是冤家？我向来从不把聚散看成一回事，在你之前，除你之外，我也并非没有好朋友，不知道为什么和你一认识之后，便像被一根绳紧紧牵系住一样，怪不自由的，心也不能像从前一样轻了，但同时却又真觉得比从前幸福得多。

- （我从今天起开始盼望见你）

  我从今天起开始盼望见你，心里的意思，怎样高兴，冀念能够读到你的信，如果你不欢喜我只愿意属于你的..

5. 总结

在完成这次期末项目的过程中，我发现了自然语言处理很多有趣的应用，也有了一些新的想法，比如分析美剧中的人物关系等。因为才疏学浅，刚刚入门，目前应用的方法也都是基于前人研究的成果学习并进行改动。希望随着研究的深入，能够自己提出一些有价值的方案，进步更快！