数据挖掘 大实验

需求

欢迎来到真实世界。为了完成这个作业，你需要在一个真实存在的任务上使用数据挖掘的方法。你需要在提供的数据集上完成挖掘任务，然后提交你的挖掘结果和技术报告，来描述你是如何完成这个任务的。

- 实验结果截止日期：6月15号 23:59
- 实验报告截止日期：6月18号 23:59
- 本次大实验占总评的40%分数

数据挖掘任务

这是一个真实的软件挖掘任务。你需要为一些消息来推荐适合的emoji表情。你可以使用任何数据挖掘方法(无论是已有的方法还是提出的新方法)来使得推荐尽可能准确。

数据的背景

消息，在人类的沟通中被广泛使用。现在提供了大量的emoji表情来更好地表达人的情绪。然而，对于一个不熟悉emoji表情的人来说，选择一个适合的emoji并不是意见容易的事情(你可以试想你的父母和祖父母是如何使用emoji的)。因此，设计一个为消息自动推荐适合的emoji的系统是必要的。可以通过对消息和emoji的关联关系进行挖掘，来建设这个自动推荐emoji的模型。

数据集简介

数据集被分为两个不相交的集合，分别名为*训练集*和*测试集*。训练集中包括消息和emoji，但测试架中只有消息。在Kaggle平台的数据集压缩包中的README文件中查看更多详细信息。

提交的评判标准

1.性能评判

评判服务由Kaggle平台提供。在challenge页面查看评判与提交的详情。你需要将预测结果组织成sample_submission.csv中描述的特定的格式，然后在网页上提交。你的预测将通过平均F1-score来评判。

在报告中你需要介绍你的方法(如果你使用了已有的方法，你需要解释为什么用，以及怎么使用这个方法)并展示并分析你的结果。你需要提供一个可以运行的，充分训练的，并且可以产生在Kaggle平台上提交的数据结果的模型(可以产生Kaggle平台上最高分数对应的模型)，以及一个简短的手册来描述如何运行模型(介绍数据格式，环境以及其他一切有必要的信息)。我们会使用你的模型来复现。

2.技术报告评判

所有提交的报告会通过一下方面来评判：

- 内容：使用方法的描述
- insight：为什么这样实现
- 新颖：已有方法的创新 or 直接使用
- 文章组织：内容是否符合逻辑，语言是否精确，简洁，易于遵循
- 论文的格式：格式是否符合要求

当然，更高的平均F1-score通常意味着更高的分数。然而，实验的得分并不只依赖于评估指标的值。一个好的报告对于得分来说同样重要。总之，你需要在模型和报告书写上做同样的努力。

指导与建议

这个实验任务的难点是什么？

- 提供的数据格式是中文字符。数据集中，包括中文字符、数字、英文字母和其他多种字符。你可以尝试使用utf_8或者utf_8_sig编码来解码。解码后，你需要在建立模型之前将每条消息转换为特征向量。你需要小心地根据挖掘任务设计如何提取这些特征。
- 训练数据是的多分类的。总共有72中emoji。你不能值关注那些最频繁的类别，因为结果的评判指标是F1-score，所以每个类别都是同等重要的。

如何决定使用什么方法？

- 使用已有的方法。你可以尝试一些已有的文本挖掘的方法，比如向量空间模型、KNN、SVM、朴素贝叶斯、深度模型，等等。你也可以使用课上介绍的方法。
- 提出新方法。你可以尝试给予多个已有的方法来提出新方法，以便考虑所提供的文本数据的特征。非常鼓励这样做！

如何书写报告？

- 报告内容应包括但不限于以下内容：
    - 简要介绍你对挖掘任务的理解
    - 你选择某个具体方法的动机
    - 使用的方法的具体技术细节
    - 将你的方法与其他两种基础方法(baseline method)做对比，说明你的方法是如何更适合这个任务的
    - 对方法和结果进行讨论，并给出结论
- 你应该以技术论文的形式来组织文章，其中包括标题、摘要、关键词、主题和参考文献。请保持样式和格式与《中国软件学报》相同。

如何提交？

- 请在报告截止日期前通过web端提交报告
- 请在Kaggle平台提交你的结果。提交的文件类型应为csv，其他一律不支持
- 结果提交应遵循样例文件的格式。提交格式不正确将不被接受
- 截止后，一切提交均不被接受！
- 不接受邮件提交！

对实验有任何问题，请直接联系我或者助教。