自然语言处理与文本挖掘
课程大作业

2020.4.22

自然语言处理和文本挖掘涵盖了非常广泛的研究内容，有着大量的研究和应用。课程大作业希望大家基于自己的兴趣或研究中遇到的实际问题，参考课程中介绍的内容，自选题目，以独立探索研究的方式，完成一项有价值的工作。

题目：

由大家自行选择，下面提供一些角度以供参考：

1. 应用自然语言处理和文本挖掘方法解决实际问题：
   a. 挖掘和处理互联网上的信息，例如从社交网络上获取电影口碑；
   b. 结合自己的学科背景，例如挖掘基因序列；
   c. 常见应用，例如聊天机器人；
   …
2. 改进已有的研究工作：
   以一个现有工作为基础，找到你认为可以改进的地方，实现并验证你的想法。
3. 针对一个公开的数据集/评测，实现一个模型，完成要求的任务。

要求：

1. 由 1~3 人组成小组，以小组为单位完成；4 月 30 日前向助教上报组队及具体题目（j-song17@mails.tsinghua.edu.cn）；
2. 16 周课上，以小组为单位进行汇报展示，展示应以 PPT 为主，也可自由发挥；每组展示时间 10 分钟，提问 2 分钟；
3. 第 16 周周末（6 月 7 日）24:00 前提交报告和程序代码，其中报告应包括问题描述，方法和实验（包括对实验结果的分析）等方面。
4. 报告中对于使用论文中的方法和引用论文中的图片，需标明引用信息。
5. 可以使用任何开源工具和开源代码，但需标明引用信息。

注：大作业的评分以工作的完整程度和深入程度为主，实验结果仅作为参考。

参考：

1. Stanford NLP 课程的往届 project：
   https://nlp.stanford.edu/courses/cs224n/
   http://cs224d.stanford.edu/reports_2016.html
   http://web.stanford.edu/class/cs224n/reports.html
   http://web.stanford.edu/class/cs224n/archive/WWW_1617/reports.html
2. 一些可以获取论文的地方：ACL，EMNLP，arxiv
3. 寻找公开任务和数据集
   https://www.kaggle.com/datasets
   https://rajpurkar.github.io/SQuAD-explorer/
   https://github.com/niderhoff/nlp-datasets