大作业

情感分析 (Sentiment Analysis)

王瑞
上海交大计算机系
wangrui12@sjtu.edu.cn
TA：胡天祥
hutianxiang@sjtu.edu.cn

目录

目录

任务介绍

- **任务介绍：**

在自然语言处理中，情感分析和观点挖掘是文本数据挖掘领域的一个重要方向。情感分析一般指判断一段文本所表达的情绪状态，属于文本分类问题。主要任务是对文本中的主观信息（如观点、情感、评价、态度、情绪等）进行提取、分析、处理、归纳和推理。本次任务中，我们需要通过模型对一段文本进行情绪的正负判断，可以作为一个简单的二分类任务。

- **数据集介绍：**

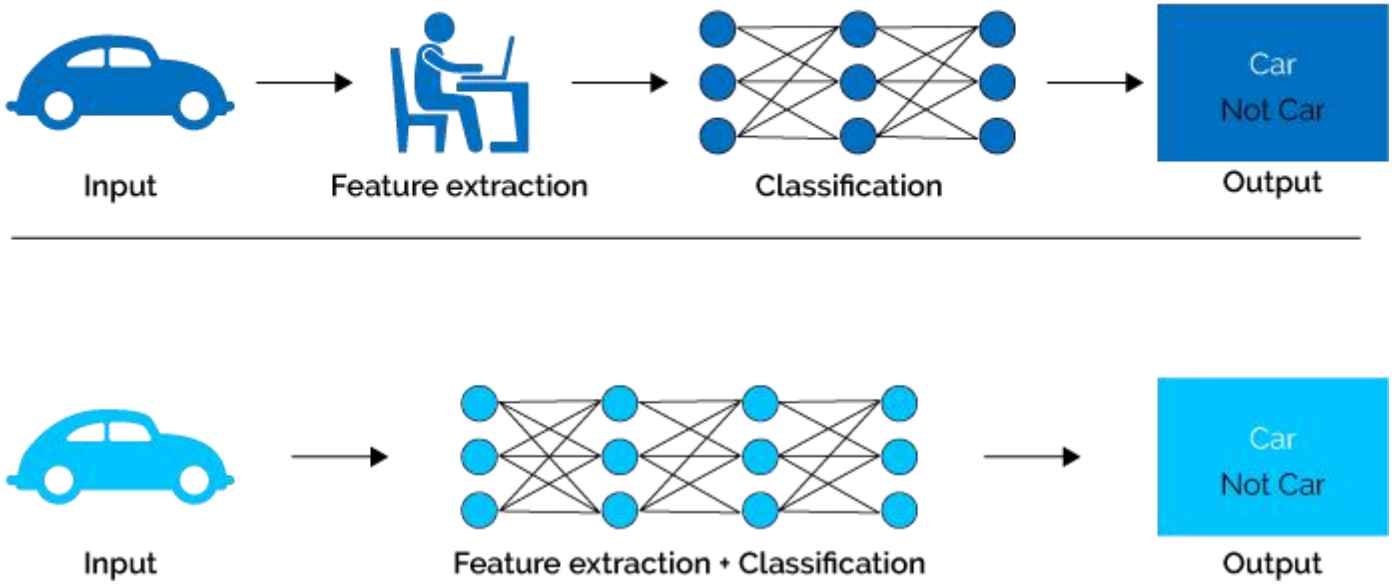
本次使用的数据集从网络上爬取的推文，其中训练集160万条。数据集拥有六个特征，分别是标签、ID、日期、话题、用户以及推文内容。

传统的情感分类

- **简单的模型**
  - 简单的情感分类可以利用词袋模型，通过统计文本中出现的正负极性词的比例来实现。
  - This restaurant is fantastic. So gorgeous decoration and meticulous service! I felt I’ m a true nobility and really like it.
- **更精细的情感分类模型**
  - 建模文本中的其他的句式、结构等特征，如：否定词、连接词、反问、转折、让步、假设、虚拟语气等。

分类任务框架

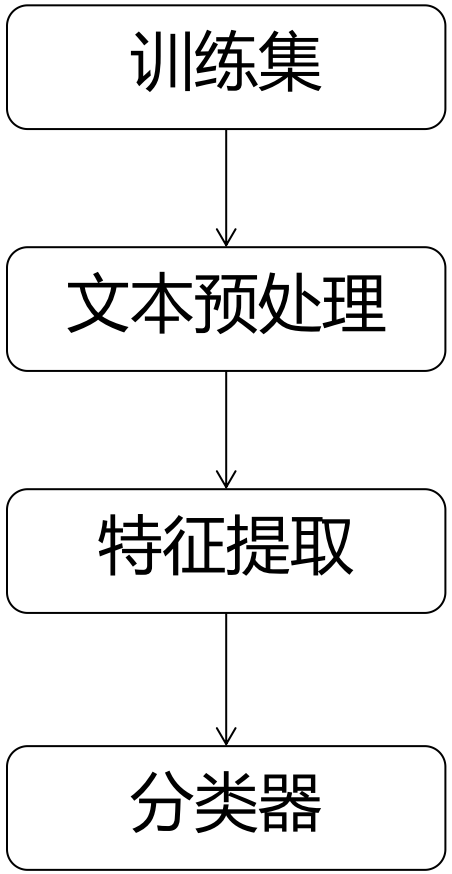
- **Feature-based method**
- **Deep neural network**



目录

Feature-based method

- **Tokenization**
- **Feature Extraction**
  - Bag of words
  - TF-IDF
  - Word2vec
- **Classification**
  - Naïve Bayes
  - MaxEnt
  - SVM

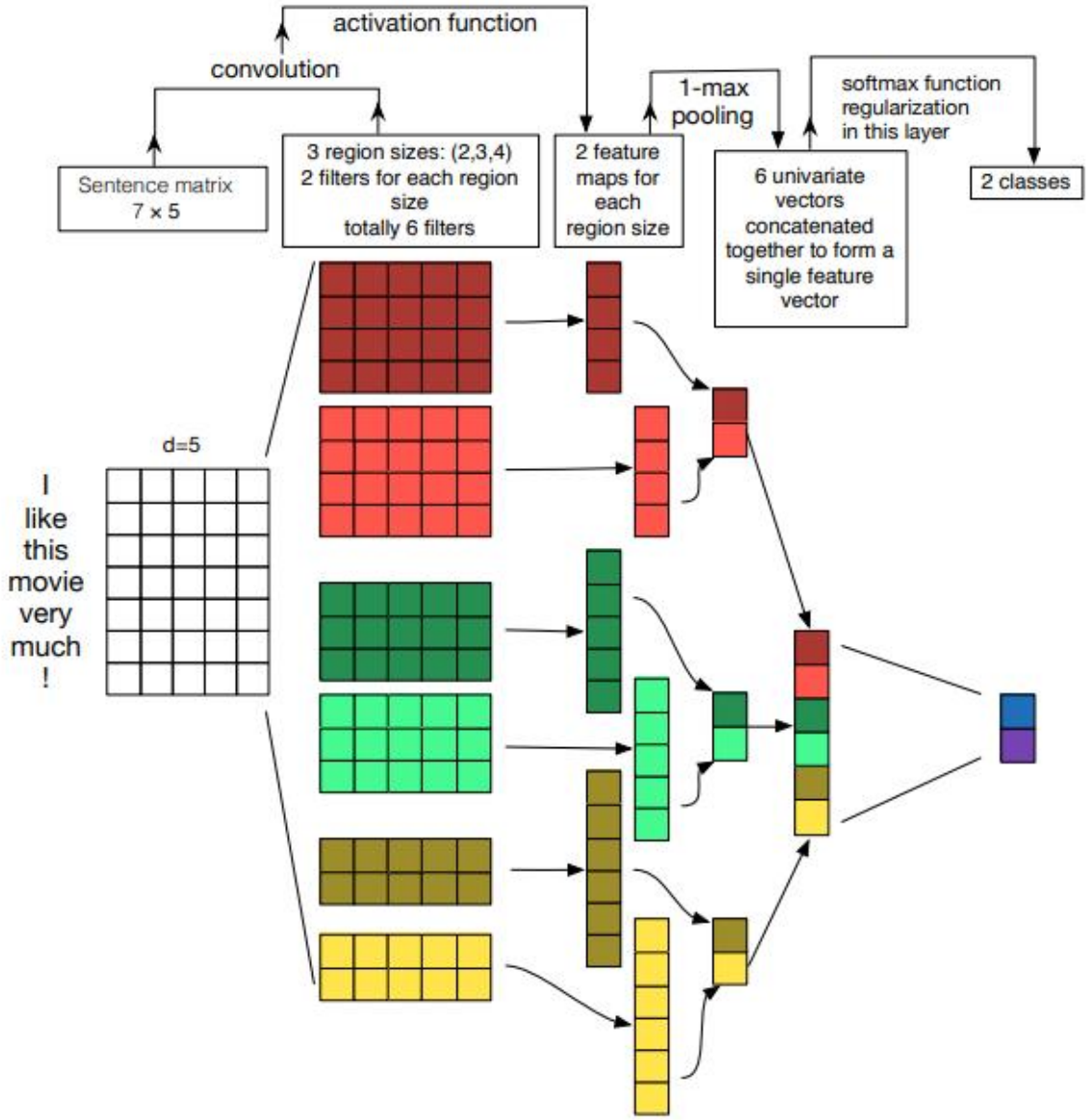


目录

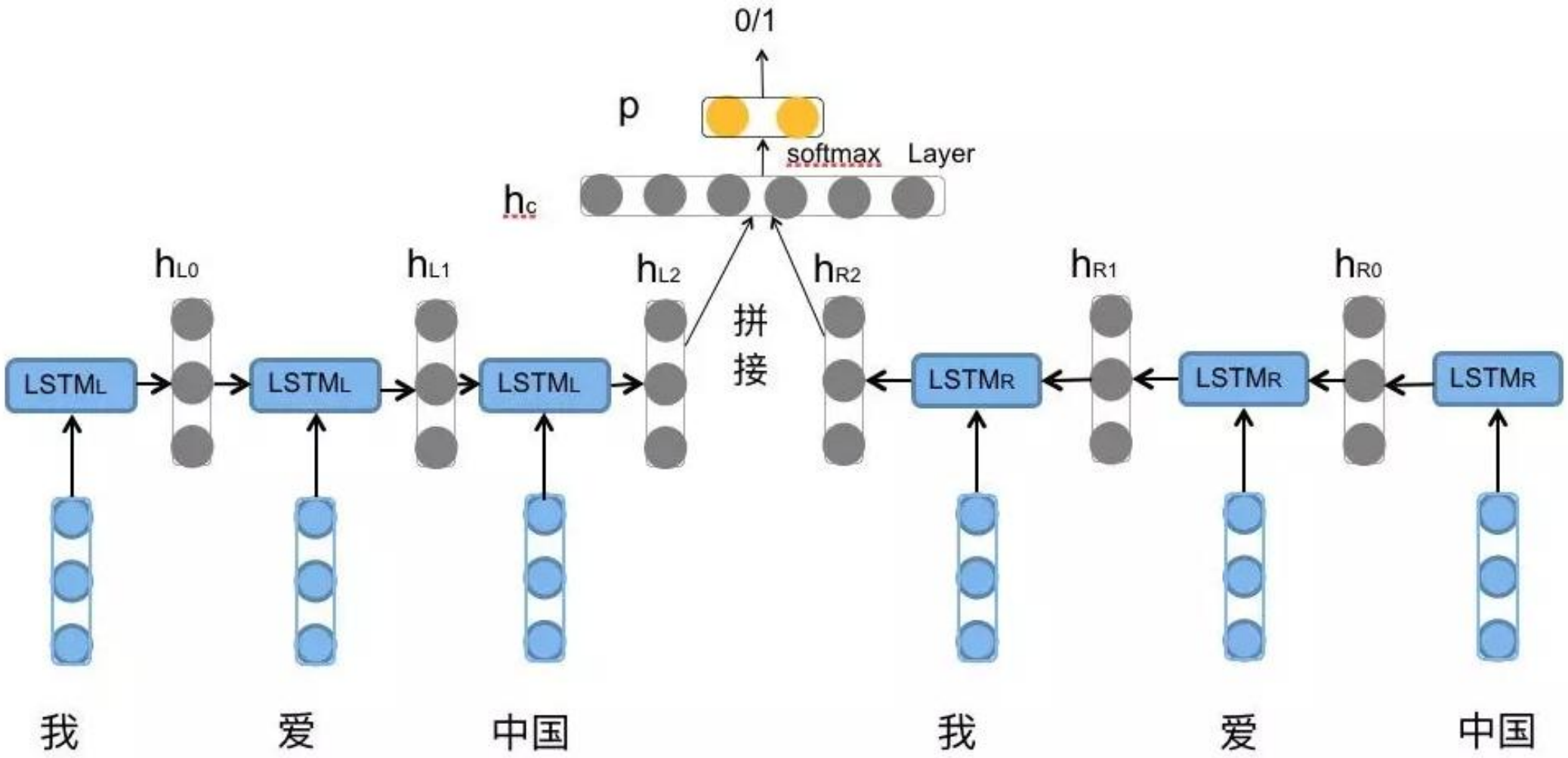
Deep neural network

- **TextCNN, TextRCNN**
- **RNN, LSTM，BiLSTM, BiLSTM_Attention**
- **Transformer，BERT+fine-tuning**

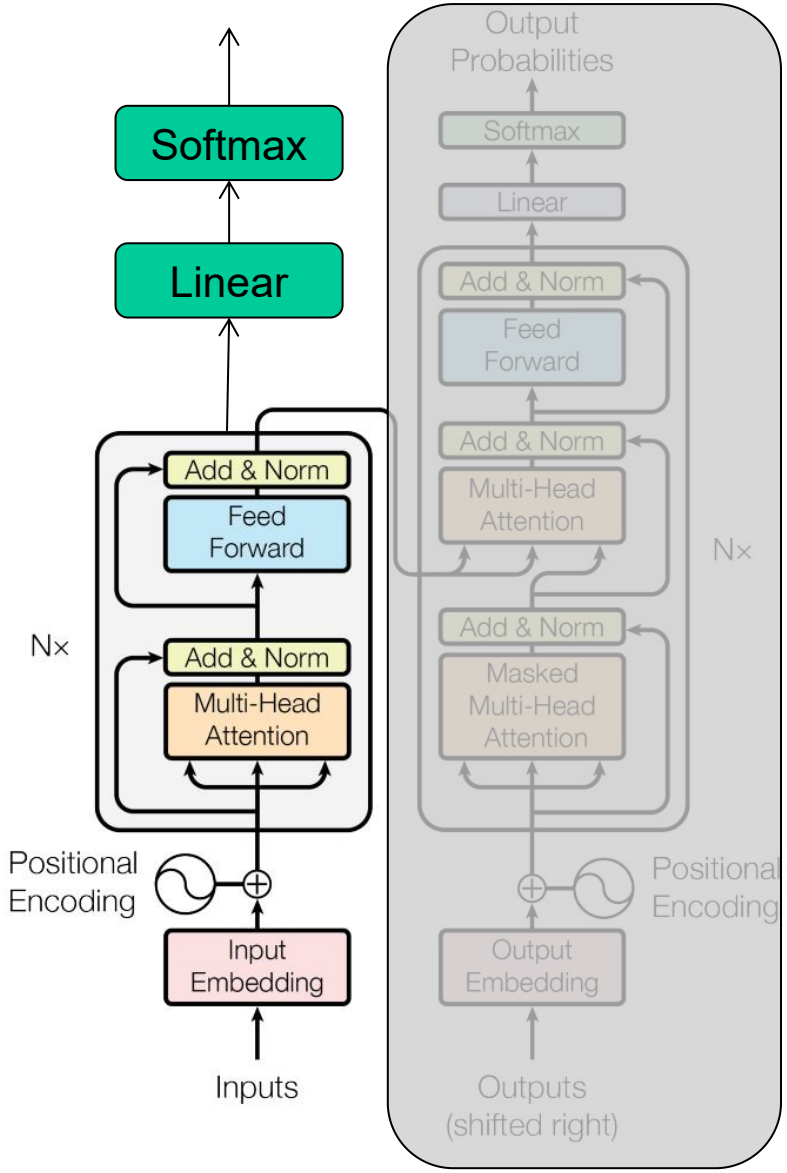
TextCNN


activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

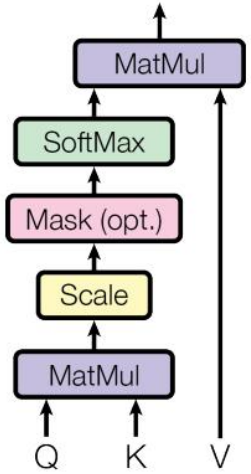
BiLSTM


0/1
p
softmax Layer
hc
hL0
hL1
hL2
拼接
hR2
hR1
hR0
LSTML
LSTML
LSTML
LSTMR
LSTMR
LSTMR
我
爱
中国
我
爱
中国

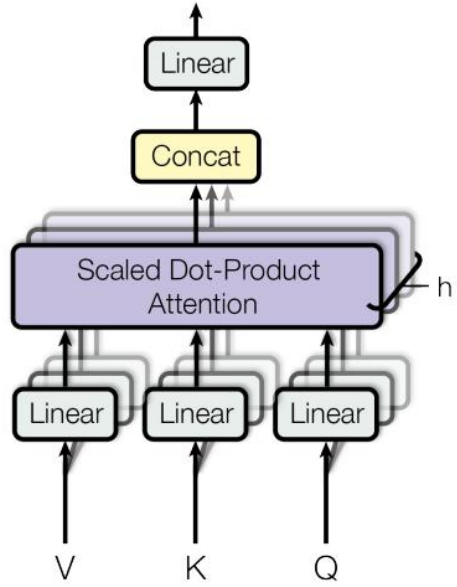
Transformer



Scaled Dot-Product Attention



Multi-Head Attention



目录

实验流程——数据预处理

- **提取source sentences & labels;**

4	3	Mon May 1	kindle2	tpryan	@stellargirl I looooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right.
4	4	Mon May 1	kindle2	vcu451	Reading my kindle2... Love it... Lee childs is good read.
4	5	Mon May 1	kindle2	chadfu	Ok, first assesment of the #kindle2 ...it fucking rocks!!!
4	6	Mon May 1	kindle2	SIX15	@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big one is huge! No need for remorse! :)
4	7	Mon May 1	kindle2	yamarama	@mikefish Fair enough. But i have the Kindle2 and I think it's perfect :)
4	8	Mon May 1	kindle2	GeorgeVHu	@richardebaker no. it is too big. I'm quite happy with the Kindle2.
0	9	Mon May 1	aig	Seth937	Fuck this economy. I hate aig and their non loan given asses.
4	10	Mon May 1	jquery	dcostalis	Jquery is my new best friend.
4	11	Mon May 1	twitter	PJ_King	Loves twitter
4	12	Mon May 1	obama	mandanico	how can you not love Obama? he makes jokes about himself.
0	14	Mon May 1	obama	kyleselle	@Karoli I firmly believe that Obama/Pelosi have ZERO desire to be civil. It's a charade and a slogan, but they want to destroy conservatism
4	15	Mon May 1	obama	theviewfa	House Correspondents dinner was last night whoopi, barbara & sherri went, Obama got a standing ovation
4	16	Mon May 1	nike	MumsFP	Watchin Espn..Jus seen this new Nike Commerical with a Puppet Lebron..sh*t was hilarious...LMAO!!!
0	17	Mon May 1	nike	vincentx24	dear nike, stop with the flywire. that shit is a waste of science. and ugly. love, @vincentx24x

- **清洗clean(比如网址、颜文字等);**
- **分词tokenization;**
- **验证集划分 (例如，|Train|:|Valid| = 9:1);**

实验流程——模型训练

- 选择合适的目标函数(loss function)；
- 写好train step, valid step；
- 设置合适的超参数(embed_dim, num_layers, ...)；
- 结合自身算力资源设置训练参数(epoch, batch size, lr, ...)；
- Train.....

实验流程——性能测试

- **评估指标**
  - 准确率：x/498，y%；
  - 精度(Precision)、召回率(Recall)、F1 score;

- **可以将预测错误的句子列出，尽量作出分析。**

目录

评分细则——实验内容+presentation

实验内容	
模块	分值
完成度	10
性能	5
探究性	10
报告	10
代码	5
合计	40

Presentation	
模块	分值
PPT表现清晰程度	5
语言表达陈述能力	5
实验内容丰富程度	5
时间控制	3
提问	2
合计	20

评分细则——实验内容

- **完成度：**有模型、能训练、能测试、保证合理准确率；
- **性能：**依据准确率排名，按组数比例2：5：3，分别得5分，4分，3分；
- **探究性：**进行探究性实验(超参数、不同模型对比、是否用word2vec初始化、新的loss function、分类任务转回归任务、实验结果分析、结合其他上课讲过的方法等)；
- **报告：**内容充实、写作规范；
- **代码：**包含注释、Readme文件，可复现；

评分细则——Presentation

- **汇报内容：**
  - 团队成员及分工
  - 研究任务的内容
  - 研究方法和研究思路
  - 实验结果与探索分析
  - 其他

- **汇报形式：**每组内推选1人进行课堂报告，报告完后，评委会进行提问

- **汇报时间：**15周开始，具体每组汇报时长(大概10-15min)等过后通知

评分细则——组内分数计算

- **最终报告要写明小组分工，即每人所做工作内容**
  - 如果组内工作量不同，还需写出每人工作量所占百分比，则每个人的最终分数=小组得分*人数*(工作量百分比/小组总百分比)，不超过60分。
  - 例如：三人小组工作量相等，可以都写33%，小组得分54分，则每人得分仍为36*3*(33%/99%)=54分。

目录

组队介绍

- 自由组队，原则上每组不超过三个人，如果本组有文科同学，则最多可以四个人一组。
- 如果找不到队友，则助教会对没有组队的同学随机分配组队。确实想单人一组的同学，请联系助教说明。
- 小组人数与最终小组评分无相关性。

报告模式

- 1. 简介 Introduction
- 2. 模型/方法 Methodology
- 3. 实验 Experiments(Setup+results)
- 4. 分析 Analysis（探究性）
- 5. 结论 Conclusion
- 参考文献 References
- Appendix (组内分工及其他想放进来的)

中英文均可且不影响分数，最终上传报告为pdf格式。

最终提交材料

- 1. 汇报PPT
- 2. 实验报告
- 3. 实验代码(不包括模型)

将以上材料打包压缩后，上传到canvas平台，每组仅需一人提交即可。

References

- Kim, Yoon. "Convolutional Neural Networks for Sentence Classification." proceedings of the Eighth International Joint Conference on Natural Language Processing. 2014.
- Zhang, Ye, and Byron C. Wallace. "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification." Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2017.
- S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no.8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." NIPS 2014 Workshop on Deep Learning, December 2014. 2014.
- Huang, Zhiheng, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging." arXiv preprint arXiv:1508.01991 (2015).
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In Advances in neural information processing systems, pp. 5998-6008. 2017.
- Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019.
- Sun, Chi, et al. "How to fine-tune bert for text classification?." China National Conference on Chinese Computational Linguistics. Springer, Cham, 2019.