Natural Language Processing I 2022F

WU Xiaokun 吴晓堃 xkun.wu [at] gmail

课程考核说明

考核标准

平时课堂参与：**20%**，课程理解：**20%**，课程项目：**50%**，荣誉（加分）：**10%**。

实践报告要求

- 完整个人信息：学号、姓名、专业。
- 课程理解：简答题，按照课程讲解、自己理解回答。
- 课程项目：写明选题，及原因。
- 正文：详细阐述解决思路。
- 代码：必须使用**Jupyter**笔记，并保留原始输出，其中包括评测指标，否则课程原则上最高**80**分。
- 总结：对照评分规则给出自我评价分数，并说明依据。

截止日期、提交要求

结课后2周，即6月10日（第16教学周）。需提交：

1. 实践报告：纸版、电子版（PDF）
2. 代码：Jupyter笔记

电子版命名格式：

```
nlp_{专业班级}_{组长姓名}.(pdf|ipynb)
```

- 专业班级：计科、软件、计科B。
- 实践报告保存成PDF，代码保存成ipynb。

课程理解问题

1. 简述N元语法的假设、两种公式与含义、计算方法、稀疏性问题及解决方案。
2. 简述序列标注任务、两大基本问题，并结合架构图简述三类主要算法（HMM、CRF、RNN）的原理。
3. 简述启发向量语义的两种观点及词嵌入表示的特点、优缺点；简述如何构造词嵌入的两种常用模型（tf-idf、word2vec）。
4. 简述循环神经网络的原理、架构图，并阐述LSTM、GRU架构的改进思路、优点与不足。

课程项目

本课程（自然语言处理）不限方法，但要求必须解决自然语言处理问题，否则课程原则上最高**70**分。

注意：选题不能与《深度学习》课程相同，否则分数只能记到一门课程。

项目等级分类

- I类：教材例子改版，原则上最高**80**分，**90**分封顶。
- II类：自选进行中的Kaggle或其他竞赛问题。
  - 注意：需要通过题目审核，否则视为I类。
- III类（荣誉加分）：实际工业问题获奖，Kaggle或其他竞赛获奖。

基本评分规则

评分细则参见附录。

- 实验报告完成度、规范程度：15%。
- 代码完成度、正确度：数据预处理，模型设计、训练、预测，评测模型：25%。
- 模型架构创新，超过基准实现、实际预测应用：10%。

注意：严禁代码造假！否则课程原则上最高**60**分。

以下为可选项目参考。

IMDB影评正负面分类问题

本题视为I类项目。

可以参考15.2. 情感分析：使用递归神经网络、15.3. 情感分析：使用卷积神经网络。基准实现的测试精度：89%。

准备新测试数据

参考附录，或在IMDB网站复制10条评论：

- 其中应该包括一半正面评价、一半负面评价。
- 代码部分：能够打印出这10条评论内容。
- 计算正确率：每识别正确一个计1%，总计10%。

自然语言推断

本题视为I类项目。

可以参考15.5. 自然语言推断：使用注意力、15.7. 自然语言推断：微调BERT。

机器翻译：德汉互译

本题视为II类项目。

可以参考9.7. 序列到序列学习（seq2seq）、10.4. Bahdanau 注意力、10.7. Transformer。

准备新测试数据

参考附录：共两部分。《查拉图斯特如是说》要求计算BLEU分数评测。《德汉互译 试题》只要求翻译。

- 代码部分：能够打印出这10条语句的内容。
- 计算正确率：计算BLEU分数。超过0.6算正确。

机器作文

本题视为II类项目。

自行设计文本生成算法，并能够以"我永远不能忘记……"开头完成一篇作文，字数控制在800字左右。可以参考CFG句法结构、语言模型可视化、朴素贝叶斯、隐式马尔科夫模型、风格迁移学习、GAN等。

- 要求：内容积极、健康，严禁极端言论！

FAQ

Q: 是否可以按小组提交?

A: 可以，但每组最多不超过5人。封面上需标明负责人1人、其他组员（不标注组长视为全是组员）。此外：

- 为了肯定组长的工作，每位组员的分数扣除1%，并将这些分数加在组长的成绩中。
- 为了鼓励独立解决问题，每位同学的最终成绩扣除[高重复度报告的总人数]%。

举例：

- 独立完成、无重复：无分数修正。
- 三人组、无重复：组长$(2-3=)-1$，组员$(-1-3=)-4$。
- 三人组，另有1份四人组报告重复度高：组长$(2-3-7=)-8$，组员$(-1-3-7=)-11$。

Q: 如何保存代码?

A: 使用 Jupyter Notebook 导出成HTML格式，然后再转成PDF格式或直接打印。注意：Jupyter 对中文的支持不好，直接转成PDF格式可能中文是空白或乱码。也可以只用 Jupyter 完成代码部分并导出打印，而正文部分在另外一个文档中撰写。

附录：影评测试集示例

以下影评均节选自：Avengers: Endgame (2019)[1]。

若使用本示例：1-5 Star 视作负面评论；6-10 Star 视作正面评论。

[1 Star] Endgame is everything wrong with today's cinema

Constant stupid jokes, annoying cgi, fake deaths, predictable outcomes, product placement, no sense of danger, self indulgent ending, idiotic fan service, brainless action, and cliche-ridden idiocy is the best way to describe this film. It is fast food for the brain. If you want to take your kids to a fun movie, skip endgame, take them to Ford vs Ferrari. It may be just as sad at times, but it is based off a true story and has great acting, intelligent humor, a fantastic script, and heartwarming

moments. It is enjoyable for everyone, and it doesn't have to be as idiotic as this trash to get people to enjoy it. Scorsese is right.

[2 Star] A Terrible Conclusion

This movie lacked everything that made its predecessor Infinity War so great. The comedic take on Thor was the worst part of the whole 22 movie franchise. The pace of the movie was slow and boring. Overall, this movie is one big disappointment.

[3 Star] Infinity war was soooo much better

WHAT THE HELL IS WRONG WITH THE RUSSO BROTHERS,,,,, After you did a terrible job of messing up the hulk in infinity war you then made it worst in endgame with nonsense professor hulk and to add insult to injury you also mess up thor,,, you suck Russo Brothers,,,, you did such a fantastic job with the winter soldier and civil war two of my favorite movies,, what went wrong with endgame,,,,, you took two of the strongest and best avengers and turn one into a big slob and turn hulk into a big stupid teddy bear,,,,, seriously disappointed.

[4 Star] Extremely overhyped movie

The movie is average, but severely overhyped. There is nothing special and this really needs to stop being said as Marvel's best movie. Everything that it tried to add in the name of science fiction was extremely confusing and despite trying to make it seem logical, it simply comes off as mumbo-jumbo and in the end you are left confused wondering what just happened. It does not have any substance. Unlike Infinity War, which managed to hold itself together as a film, balancing all aspects such as action, emotion while also keeping the audience engaged and being understandable, Endgame is simply confusing. It is slightly redeemed by the final battle scene, and the visual effects and action is commendable, it does not completely redeem the film and it still remains an average experience.

[5 Star] Doesn't make sense

Writting about time traveling is always difficult. But there's been so many inconsistencies in this movie that I lost Interest after 1 hour watching. The writting was so poorly done that I felt no emotions and got bored at some point. Infinity war were so well done that I'm really surprised that the last Avengers movie turned down like this.

[6 Star] So much potential lost ....

After Infinity War I had huge expectations. Unfortunately I think it doesn't follow the comic book at all because they wanted some kind of closure of the story. I saw a lot of 10 star reviews and I cannot believe the majority of people like this movie more than Infinity War. I'm not saying it wasn't entertainment to watch but the feeling is that all heroes were forced to fit in a movie paying a big price to the story. There are moments in the movie where the rules from all other Marvel movies are broken and characters which were build over hours of screen time were reduced in order to make space for others. The story is average at best with big flaws.

[7 Star] Everyone just calm down...

... and pull the hype train over. Amazing fun, stunning spectacle great performances especially RDJ. It's just not the whole sum of its parts. Lots of great parts but not cohesively held. Too much Comedy with one character essentially now just full comedy relief. The most ridiculous coincidences and a couple of plot holes. You will laugh, you will get gut punched and for some you will cry, but then you will wonder what really happened in that 3 hours. Did it need 3 hours? No. Is it better than infinity war? No. Will you enjoy it? Definitely.

[8 Star] Endgame

There isn't too much in the way of suspense or surprises when it comes to the story, but there are some shocking moments and funny lines in this epic finale. Again, like many of the best Marvel films, the holes and flaws are covered up with humor and fan service, making everything okay. That being said, I did prefer Infinity War to this film, which really misses the leads of the other Marvel franchises that were "snapped" out. Overall, however, there are only a few ways you can wrap up the main story of the MCU, and this was a solid direction.

[9 Star] Perfect..no. Epic..absolutely

I think how viewers receive this movie is dependent upon when you were born. I was 12 years old and glued to our big screen color TV on 9/8/66 for the premier of Star Trek. I still have my Leonard Nimoy autographed fan club membership cards from 1966-69. I was in the audience in 1977 opening weekend of Star Wars after a 3+ hour wait in line. What the Marvel team created here was nothing short of spectacular. To bring these characters to life, weave their stories together bringing some to a close and launching others while still managing to cross the T and for the I, well that's epic. I've experienced a few.

[10 Star] The ending made all 22 movies worth it

If you're going to watch this movie, avoid any spoilers, even spoiler free reviews. Which is why I'm not going to say anything about the movie. Not even my opinion. All I'm going to say is:

The crowd applauded 3 times during the movie, and stood up to clap their hands after. This I have never witnessed in a Dutch cinema. Dutch crowds aren't usually passionate about this. I checked the row where I was sitting, and people were crying. After the movie, I was seeing people with smudged mascara. That's all I have to say about the movie.

附录：德汉互译示例

节选自《查拉图斯特如是说》。„Also sprach Zarathustra. Ein Buch für Alle und Keinen"。

Du großes Gestirn! Was wäre dein Glück, wenn du nicht Die hättest, welchen du leuchtest!
你，伟大的星球！假若你没有被你照耀的人们，你的幸福何在呢！

Siehe! Ich bin meiner Weisheit überdrüssig, wie die Biene, die des Honigs zu viel gesammelt hat, ich bedarf der Hände, die sich ausstrecken.
看啊！我像积蜜太多的蜂儿一样，对于我的智慧已经厌倦了；我需要有人伸手来领受这智慧。

Und also sprach der Greis zu Zarathustra: "Verwandelt ist Zarathustra, zum Kind ward Zarathustra, ein Erwachter ist Zarathustra: was willst du nun bei den Schlafenden?
老者对查拉图斯特拉如是说道："查拉图斯特拉变了，查拉图斯特拉变成了孩子，查拉图斯特拉是个觉醒者：现在你要到沉睡者那里去干什么呢？"

Als Zarathustra aber allein war, sprach er also zu seinem Herzen: "Sollte es denn möglich sein! Dieser alte Heilige hat in seinem Walde noch Nichts davon gehört, dass Gott tot ist!"
可是当查拉图斯特拉独自一人时，他对他的心如是说道："难道有这种可能！这位老圣人在森林中竟毫无所闻，不知道上帝已经死了！"

Ich lehre euch den Übermenschen. Der Mensch ist Etwas, das überwunden werden soll. Was habt ihr getan, ihn zu überwinden?
我教你们何谓超人。人是应被超越的某种东西。你们为了超越自己，做过些什么呢？

Wahrlich, ein schmutziger Strom ist der Mensch. Man muss schon ein Meer sein, um einen schmutzigen Strom aufnehmen zu können, ohne unrein zu werden.
确实，人是一条不洁的河。要能容纳不洁的河流而不致污浊，人必须是大海。

Was gross ist am Menschen, das ist, dass er eine Brücke und kein Zweck ist: was geliebt werden kann am Menschen, das ist, dass er ein Übergang und ein Untergang ist.
人类之伟大处，正在它是一座桥而不是一个目的。人类之可爱处，正在它是一个过程与一个没落。

Seht, ich bin ein Verkündiger des Blitzes und ein schwerer Tropfen aus der Wolke: Dieser Blitz aber heißt Übermensch.
看，我是闪电的宣告者，从云中落下的一滴沉重的雨点：而这个闪电就叫做超人。

Man muss noch Chaos in sich haben, um einen tanzenden Stern gebären zu können.
存在于混沌之中，才能生出舞动的星。

Zu meinem Ziele will ich, ich gehe meinen Gang; über die Zögernden und Saumseligen werde ich hinwegspringen. Also sei mein Gang ihr Untergang!
我要朝着我的目标行进，我要超越那些迟疑者和拖延者。我的行进便是他们的没落！

节选自《上海外国语大学 2017 年硕士研究生入学考试德汉互译试题》

1. Märchen sind nicht auf dem Rückmarsch, ganz im Gegenteil.
2. Der Preisträger des Europäischen Märchenpreises 2012 setzt sich seit über 40 Jahren mit der Bedeutung und Weiterverbreitung von Märchen auseinander.
3. Das Prinzip von Gut und Böse ist auch in modernen Sagen wie Harry Potter zu finden.
4. Seine Liebe für deutsche Märchen und die deutsche Sprache entdeckte der Germanist und Folklorist auf Umwegen: nach dem Abitur ging der gebürtige Deutsche zum Studium in die USA, eigentlich wollte er Mathematiker werden.
5. Heute sieht das natürlich anders aus: Durch das digitale Zeitalter hat sich die Verbreitung enorm verändert.
6. 别的表情等待反应，例如悲哀等待怜悯，威严等待慑服，滑稽等待嬉笑。唯美貌无为，无目的，使人没有特定的反应义务的挂念，就不由自主地被吸引，其实是被感动。
7. 其实美貌这个表情的意思，就是爱。这个意思既蕴藉又坦率地随时呈现出来。拥有美貌的人并没有这个意思，而美貌是这个意思。
8. 用美貌这个先验的基本表情，再变化为别的表情，特别容易奏效(所以演员总是以美貌者为上选。日常生活中，也是美貌者尽占优势)，那变化出来的别的表情，既是含义清晰，又反而强化美貌。
9. 美貌的人睡着了，后天的表情全停止，而美貌是不睡的，美貌不需要休息；倒是由于撤除附加的表情，纯然只剩美貌这一种表情，就尤其感动人，故曰：睡美人。
10. 老人睡着，见得更老，因为别的附加的表情率尔褪净，只剩下衰败的美貌这一种惨相，光荣销歇，美貌的废墟不及石头的废墟，罗马夕照供人凭吊，美貌的残局不忍卒睹。

附录：详细评分规则

- 实验报告完成度、规范程度：15%。
- 代码完成度、正确度：数据预处理，模型设计、训练、预测，评测模型：25%。
- 模型架构创新，超过基准实现、实际预测应用：10%。

实验报告完成度：15%

- 结构清晰、完备（封面、正文、代码、总结）：5%。
- 解决思路严谨、合理：5%。
- 代码逻辑正确、保留原始输出：5%。

代码完成度：25%

参考3.6. softmax回归的从零开始实现、4.10. 实战Kaggle比赛：预测房价。

- 数据获取与预处理：5%。
- 模型定义与训练：5%。
- 模型选择（使用验证方法调节超参数）：5%。
- 使用模型预测（`predict`、`argmax`）：5%。
- 正确评测模型（`evaluate`）：5%。

模型架构创新等：10%

- 模型架构创新：5%。
- 效能超过基准实现、实际预测应用：5%。

---

1. https://www.imdb.com/title/tt4154796 ↩