arXiv App 第二次实验报告

彭培烜 2018202201

一、 进度概要

1. 完善了自然语言问答系统；
2. 实现了标题归纳；
3. 对数据进行了分类处理，生成了部分可视化图表，为后续可视化提供了基础。

二、 自然语言问答系统

1. 概要

此次问答系统所用的框架为 farm-haystack (https://github.com/deepset-ai/haystack/)。该框架的各个主要部件已经在前一次报告中说明。其中最为关键的部分是 Reader。在这一阶段，我选择了更加合适的模型，解决了之前程序中的部分 bug，该程序已经能够成功运行。

2. Reader 训练

在实现问答系统时，需要对 Reader 进行神经网络训练，使其能够通过阅读文本来寻找答案。自然语言，即使用我们平时交流的语法方式进行提问，然后该系统通过 Reader 对储存的文件进行搜索，通过内容和问题的关联给出答案，此次我采用了 Hugging face 模型中心的 deepset/roberta-base-squad2 进行训练。

```
reader = TransformersReader(model_name_or_path='deepset/roberta-base-squad2',
                            tokenizer='deepset/roberta-base-squad2',
                            context_window_size=500, use_gpu=-1)
```

3. 运行情况演示

以自然语言的方式提出问题

```
How do we explore the Cs_2 Feshbach molecules?
```

搜索给出答案

```
{   'answer': 'rich internal structure',
    'context': '  We explore the rich internal structure of Cs_2 Feshbach '
               'molecules. Pure\n'
               'ultracold molecular samples are prepared in a CO_2-laser '
               'trap, and a multitude\n'
               'of weakly bound states is populated by elaborate '
               'magnetic-field ramping\n'
               'techniques. Our methods use different Feshbach resonances '
               'as input ports and\n'
               'various internal level crossings for controlled state '
               'transfer. We populate\n'
               'higher partial-wave states of up to eight units of '
               'rotational angular momentum\n'
               '(l-wave states). We investigate the molecular structure by '
               'measurements of the\n'
               'magn'},
```

其中 context 是摘要中的原文，answer 是根据这些内容得出的最终回答。

三、 标题归纳

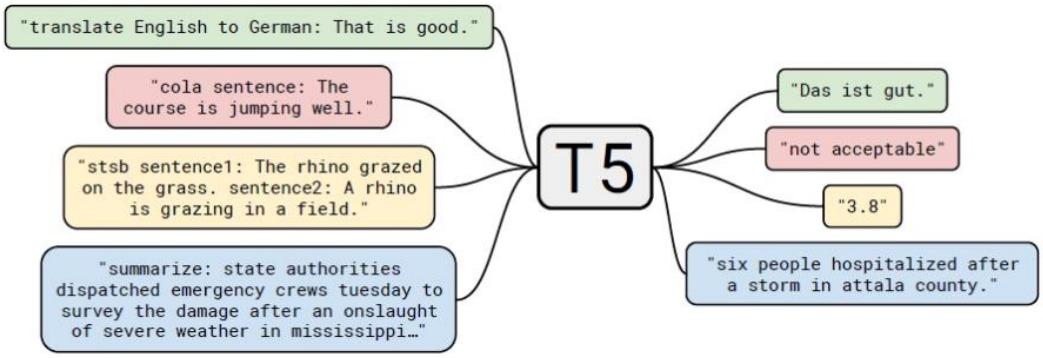
1. 概述

我是在 kaggle 页面上看到了这个 Task，这个任务的目的是为了找出是否可以训练出 NLP 模型来对给定的摘要生成文章标题。这是一个文本到文本的任务，因此通过使用 T5 模型解决问题。

2. T5 模型

T5 主要包含模型结构、注意力掩码机制、Objectives、C4 和训练策略五部分。

其模型结构为 Encoder-Decoder 结构，注意力掩码机制为 Fully-visible，Objectives 包含常用预训练方法、Mask 策略、Mask 比率和 Span 长度四个部分。



3. 具体实现

考虑到计算机配置，本次实验中使用了 2017-2020 年 4 年的论文数据。

```
metadata = get_metadata()
titles = []
abstracts = []
years = []
for paper in metadata:
    paper_dict = json.loads(paper)
    ref = paper_dict.get('journal-ref')
    try:
        year = int(ref[-4:])
        if 2016 < year < 2021:
            years.append(year)
            titles.append(paper_dict.get('title'))
            abstracts.append(paper_dict.get('abstract'))
    except:
        pass
```

T5 模型配置如下

```
model_args = {
    "reprocess_input_data": True,
    "overwrite_output_dir": True,
    "max_seq_length": 512,
    "train_batch_size": 16,
    "num_train_epochs": 4,
}

# model = Seq2SeqModel(encoder_decoder_type="bart",
                       # encoder_decoder_name="facebook/bart-base",
                       # args=model_args)
model = T5Model("t5-small", args=model_args, use_cuda=True)
model.train_model(train_df)
results = model.eval_model(eval_df)
```

此处还可以选择 facebook/bart-base 模型进行训练，但消耗过大，无法运行。

最后的期望结果如下

```
Actual Title: Computational intelligence for qualitative coaching diagnostics:
  Automated assessment of tennis swings to improve performance and safety
Predicted Title: ['Personalized qualitative feedback for tennis swing technique using 3D vid
eo video']
Actual Abstract: ['summarize:   Coaching technology, wearables and exergames can provide qua
ntitative\nfeedback based on measured activity, but there is little evidence of\nqualitative
feedback to aid technique improvement. To achieve personalised\nqualitative feedback, we dem
onstrated a proof-of-concept prototype combining\nkinesiology and computational intelligence
that could help improving tennis\nswing technique utilising three-dimensional tennis motion
data acquired from\nmulti-camera video. Expert data labelling relied on virtual 3D stick fig
ure\nreplay. Diverse assessment criteria for novice to intermediate skill levels and\nconfig
urable coaching scenarios matched with a variety of tennis swings (22\nbackhands and 21 fore
hands), included good technique and common errors. A set\nof selected coaching rules was tra
nsferred to adaptive assessment modules able\nto learn from data, evolve their internal stru
ctures and produce autonomous\npersonalised feedback including verbal cues over virtual came
ra 3D replay and\nan end-of-session progress report. The prototype demonstrated autonomous\n
assessment on future data based on learning from prior examples, aligned with\nskill level,
flexible coaching scenarios and coaching rules. The generated\nintuitive diagnostic feedback
consisted of elements of safety and performance\nfor tennis swing technique, where each swin
g sample was compared with the\nexpert. For safety aspects of the relative swing width, the
prototype showed\nimproved assessment ...\n']
```

目前由于环境以及机器配置原因，还无法运行。

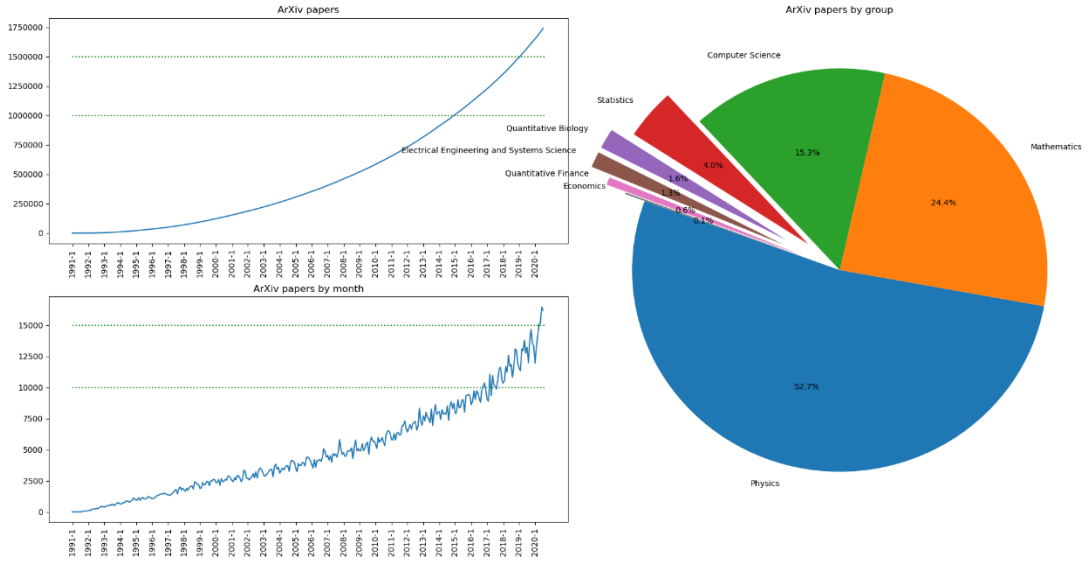
四、 数据预处理

1. 概述

为了实现后续的知识图谱以及可视化，需要对 arXiv 的论文数据进行分类处理。这一阶段主要完成了学科论文的统计以及找出了各个学课最具影响力的文章。同时还找出了最热门的话题，并为其生成了一个词云。

2. 论文分类统计

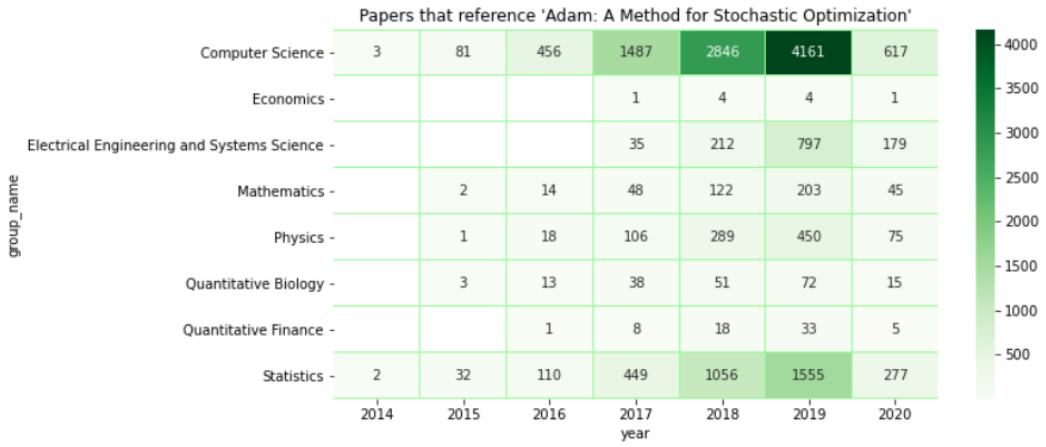
具体情况如下图，可以看到论文主要涉及物理，数学和计算机科学。



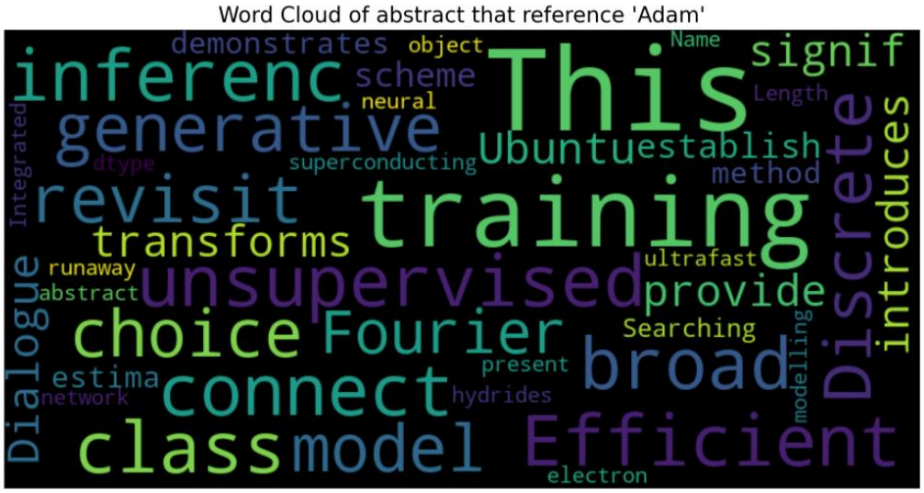
3. 最具影响力的文章

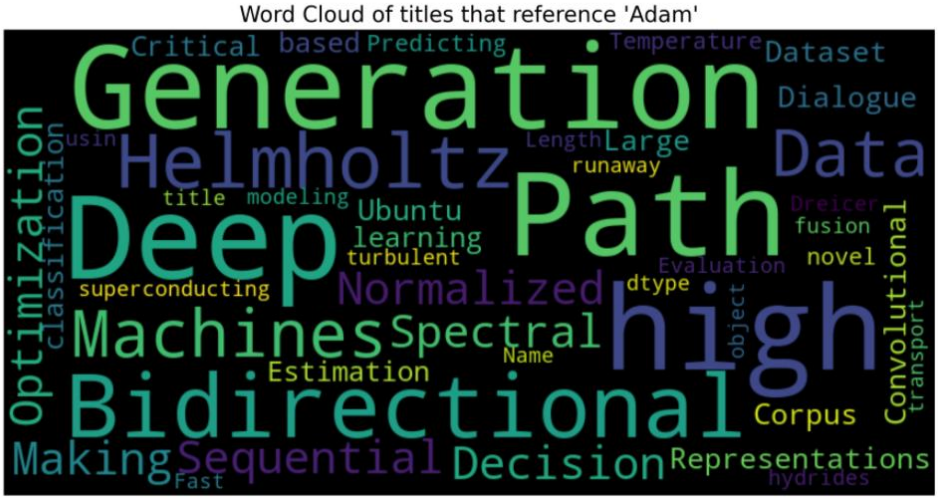
我们经过引用分析，得出了近年来最具影响力的文章：

Adam: A Method for Stochastic Optimization



并根据其引用关系生成了词云图





五、 下一阶段目标

下一阶段主要是完成标题归纳，以及知识图谱和数据可视化。