NLP Project

Part 1 20%

从JSON文件中提取文本列表, 尽量提取全文而不仅仅是标题/摘要.

Part 2 30% 分割

遍历提取的文本列表并分割成单词, 这部分任务被划分成了3个小任务和一个bonus

2.1 使用split()通过指定分隔符对字符串进行切片 **10%**

使用语法`str.split(str="", num=string.count(str)).`

- str -- 分隔符，默认为所有的空字符，包括空格、换行(\n)、制表符(\t)等。
- num -- 分割次数。默认为 -1, 即分隔所有。

2.2 使用NLTK或者SciSpacCy python的自然语义处理库 **10%**

这两个库区别不大, 用哪个都行.这个库的使用需要学习一下, 安装起来也可能会遇到一些小问题

功能还是很多的, 使用也很方便, 比如文本切分为语句， sent_tokenize()

```
from nltk.tokenize import sent_tokenize
text=" Welcome readers. I hope you find it interesting. Please do reply."
print(sent_tokenize(text))
```

2.3 使用BPE算法 **10%**

BPE是一种压缩算法，是一种自下而上的算法。将单词作为单词片段处理（word pieces），以便于处理未出现单词。

> 基本过程如下：
>
> （1）首先将统计text中单词，做成词汇表（单词-频率），然后按照unigram进行分解。
>
> 5 l o w
> 2 l o w e r
> 6 n e w e s t
> 3 w i d e s t
>
> 词汇表：l, o, w, e, r, n, w, s, t, i, d,
>
> （2）寻找频率最大的片段（字符），进行组合，将组合片段加入词汇表。
>
> 5 l o w
> 2 l o w e r
> 6 n e w es t
> 3 w i d es t
>
> 词汇表：l, o, w, e, r, n, w, s, t, i, d, es
>
> （3）继续重复上述操作，直到达到设定的阈值（词汇数+操作数）->操作数是唯一的超参数
>
> 5 l o w
> 2 l o w e r
> 6 n e w est
> 3 w i d est

> 词汇表：l, o, w, e, r, n, w, s, t, i, d, es，est
>
> 5 lo w
> 2 lo w e r
> 6 n e w est
> 3 w i d est
>
> 词汇表：l, o, w, e, r, n, w, s, t, i, d, es,est,lo

这是一个已经训练好的模型, 需要 install Huggingface's transformers

2.4 bonus +5

上述模型并不适用于生物医学领域, 因为它与日常生活中的用法非常不同, 所以或许我们可以为生物医学领域建立并且训练一个新的BPE模型.

Part 3 30% 表示单词

这部分需要我们对于提取到的单词构建表示, 维度可以限制在256个, 这三个部分基本上分开的, 可以一人做一个

3.1 N-gram语言模型 **15%**

N-Gram语言模型简单的说就是假设当前词只和前N个词相关。那我们在训练神经网络时，就会用(前N个词,当前词)构建训练样本对，用大型的语料库训练完之后，会得到每个词的word embedding。我们可以通过前N个词来猜测出第N+1个, 或许可以看看这篇

3.2 Skip-gram **15%**

通过中间词来预测上下文

3.3 Bonus **5%**

实时上下文感觉不是很容易

Part 4 20% 可视化

4.1 用tsne来可视化, 网上有代码, 最终需要给出一个`A diagram by t-SNE based on representations of up to 1000 words.` 只要给出随机1000个单词或者整个词汇表的可视化就好了 **5%**

4.2 生物医学单词的可视化, 这部分需要聚类, 因为同一类别的疾病应当给与相同的颜色 **5%**

4.3 Covid-19 给出`A sorted list of biomedical entities and description on how the entities are selected and sorted.` **5%**

4.4 找出Covid-19最相近的biomedical entity **5%** 4.3 4.4几乎捆绑

Part 5 10%

基于CORD-19数据集构建生物医学知识图, 并从中挖掘有用信息.

project里给出了一种解题思路, 先把前面的做好了再做这个吧, 跟前面的project关联度挺高的.

DDL

Part 1 1/28

Part 2 2/2(考虑到过年)

Part 3 2/5

Part 4 2/10

Report建议在每个part完成后就可以写了

关于NLP的综述可以让一个人前期主要写, Part 1 2 3应该也不用5个人做, 其他人把看的文献笔记发给写综述的同学就好了