Task 1 赛题理解

2群-181-datago-StarLEE

1. 赛题数据

数据为新闻文本，但是经过了字符级别的匿名处理，将字符串变成了数字，这将是进行NLP的一个难点。

数据共分为14类（以数字0~13表示）：

```
{'科技': 0, '股票': 1, '体育': 2, '娱乐': 3, '时政': 4, '社会': 5, '教育': 6, '财经': 7, '家居': 8, '游戏': 9, '房产': 10, '时尚': 11, '彩票': 12, '星座': 13}
```

训练集中数据的样式：

```
In [3]: 1 train_df.head()
```

Out[3]:

	label	text
0	2	2967 6758 339 2021 1854 3731 4109 3792 4149 15...
1	11	4464 486 6352 5619 2465 4802 1452 3137 5778 54...
2	3	7346 4068 5074 3747 5681 6093 1777 2226 7354 6...
3	2	7159 948 4866 2109 5520 2490 211 3956 5520 549...
4	3	3646 3055 3055 2490 4659 6065 3370 5814 2465 5...

label为文本所属类别，text为文本内容

2. 评测标准

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

预测正，实际正：true positive（TP），预测正，实际负：false positive（FP）

预测正，实际正：true positive（TP），预测正，实际负：false positive（FP）

预测	实际	
+	+	true positive (TP)
+	-	false positive (FP)
-	+	false negative (FN)
-	-	true negative (TN)

$$N_{pre} = TP + TN$$

$$N_{total} = TP + TN + FP + FN$$

定义正样本个数为*P*，负样本个数为*N*，则

$$P = TP + FN, N = TN + FP$$

定义识别率（acc），召回率（recall），精确率（precision）以及F1-score：

$$acc = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{N_{total}}$$

$$recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

$$precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2TP}{2TP + FN + FP} = 2 * \frac{precision * recall}{precision + recall}$$

可见，recall体现了分类模型对正样本的识别能力，precision体现了模型对负样本的区分能力，而F1为recall和precision的调和平均数，综合体现了模型的稳健程度

3. 解题思路

思路1：TF-IDF + ML分类器

TF-IDF（term frequency–inverse document frequency）是一种统计方法，评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

思路2：FastText

FastText 是由 Facebook's AI Research (FAIR) lab 创造的 word embedding 和 text classification 库。该模型允许监督学习和非监督学习算法来获取代表文本的向量，且 fastText 运用NN来进行word embedding。

思路3：WordVec + DL分类器

Word2vec 是一组用来进行 word embedding 的模型，这些模型为浅层双层的 NN，用来训练以重新构建语言文本的词汇。Word2vec 以大量文本语料库为输入，以向量空间为输出。语料库中的每一个独特的单词都对应到空间中的一个向量。具有相似文本的向量在空间中将离得很近。

思路4：Bert词向量

Bert（Bidirectional Encoder Representations from Transformers）在 pre-train 上运用了 Masked Language Model 和 Next Sentence Prediction 两种方法分别捕捉词语和句子级别的 representation。MLM 是指在输入的词序列中，随机遮挡15%的词，并遮挡部分词语进行双向预测。