

Datawhale零基础入门NLP赛事

新闻文本分类

Baseline和比赛介绍

分享人：阿水





目录

contents

Part 1 比赛介绍

Part 2 baseline

Part 3 比赛知识点

天池新人赛由天池与Datawhale联合发起，并提供学习内容和组织学习：

- ❑ Datawhale是一个专注于数据科学与AI领域的开源组织；
- ❑ NLP直播PPT 可关注Datawhale，回复关键词 **NLP直播** 下载；
- ❑ 同时可以加入Datawhale数据竞赛交流群，一起组队参赛，交流学习；





个人介绍

阿水, <https://www.zhihu.com/people/finlayliu>

- ✓ 天池数据大神;
- ✓ 知乎小V, 全网文章阅读量20W+;
- ✓ 数据科学爱好者, 擅长计算机视觉;
- ✓ Datawhale成员, 内容学习 & 分享者;

数据竞赛开源项目发起人:

<https://github.com/datawhalechina/competition-baseline>



Part 1 比赛介绍

本次新人赛是Datawhale与天池联合发起的零基础入门系列赛事第三场：零基础入门NL[赛事之街景字符识别。

零基础入门系列赛事：

- ❑ 比赛为个人赛，但可以组队学习；
- ❑ Datawhale提供组队学习社群，组队打卡学习；
- ❑ 全程360度分享，包括baseline和比赛知识点；

零基础入门数据挖掘之二手车交易价格预测大赛

第一场：<https://tianchi.aliyun.com/competition/entrance/231784/introduction>

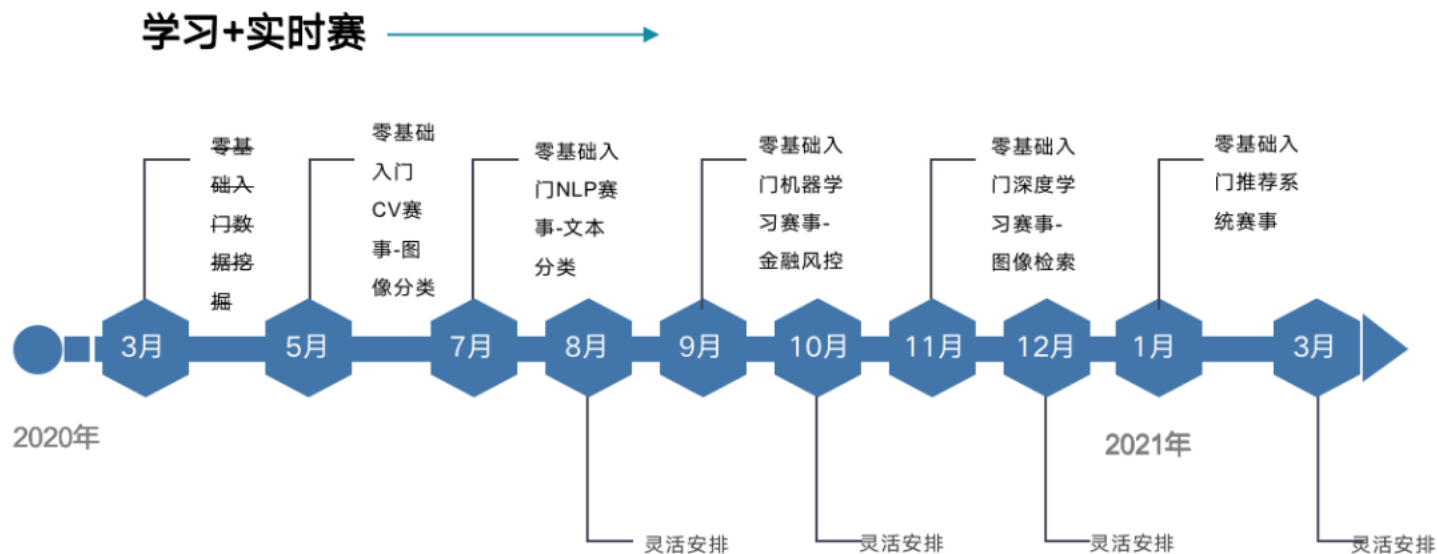
零基础入门CV赛事之街景字符识别

第二场：<https://tianchi.aliyun.com/competition/entrance/531795/introduction>

零基础入门NLP赛事之街景字符识别

第三场：<https://tianchi.aliyun.com/competition/entrance/531810/introduction>

本次竞赛是Datawhale联合天池的系列学习（数据挖掘、CV、NLP、机器学习、深度学习、推荐系统）第三场。



赛题主题：以自然语言处理为背景，要求选手根据新闻文本字符对新闻的类别进行分类，这是一个经典文本分类问题。

赛题目标：通过这道赛题可以引导大家走入自然语言处理的世界，带大家接触NLP的预处理、模型构建和模型训练等知识点。

赛题数据：赛题需要选手根据匿名新闻字符识别新闻的类别，总共包括14类新闻。

label	text
6	57 44 66 56 2 3 3 37 5 41 9 57 44 47 45 33 13 63 58 31 17 47 0 1 1 69 26 60 62 15 21 12 49 18 38 20 50 23 57 44 45 33 25 28 47 22 52 35 30 14 24 69 54 7 48 19 11 51 16 43 26 34 53 27 64 8 4 42 36 46 65 69 29 39 15 37 57 44 45 33 69 54 7 25 40 35 30 66 56 47 55 69 61 10 60 42 36 46 65 37 5 41 32 67 6 59 47 0 1 1 68

为什么使用匿名数据集？

- ✓ 匿名数据集没有版权问题；
- ✓ 匿名数据集需要大家从头构建词向量；

Part 2 baseline

本次赛题：赛题本质是文本分类问题，需要对图片的字符进行识别。但赛题给定的数据为匿名数据（以字为级别进行了匿名处理），因此没办法进行分词。

赛题思路：

- ✓ TF-IDF + RidgeClassifier ;
- ✓ FastText;
- ✓ Word2Vec + TextCNN;
- ✓ Bert;

思路1：将字符进行TF-IDF统计，然后送入线性分类器进行训练；

- ✓ CountVectorizer + RidgeClassifier;
- ✓ TfidfVectorizer + RidgeClassifier;

TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Term frequency

Number of times term t appears in a doc, d

Inverse document frequency

$$\log \frac{1 + n}{1 + \text{df}(d, t)}$$

of documents n

Document frequency of the term t

思路1: 将字符进行TF-IDF统计, 然后送入线性分类器进行训练;

- ✓ CountVectorizer + RidgeClassifier;
- ✓ TfidfVectorizer + RidgeClassifier;

```
from sklearn.feature_extraction.text import TfidfVectorizer
import pandas as pd
train_df = pd.read_csv('../input/train_set.csv', sep='\t')
test_df = pd.read_csv('../input/test_a.csv', sep='\t')

tfidf = TfidfVectorizer(max_features=2000).fit(train_df['text'].iloc[:].values)
train_tfidf = tfidf.transform(train_df['text'].iloc[:].values)
test_tfidf = tfidf.transform(test_df['text'].iloc[:].values)

clf = RidgeClassifier()
clf.fit(train_tfidf, train_df['label'].iloc[:].values)

df = pd.DataFrame()
df['label'] = clf.predict(test_tfidf)
df.to_csv('submit.csv', index=None)
```

execution queued 19:54:21 2020-07-21

baseline



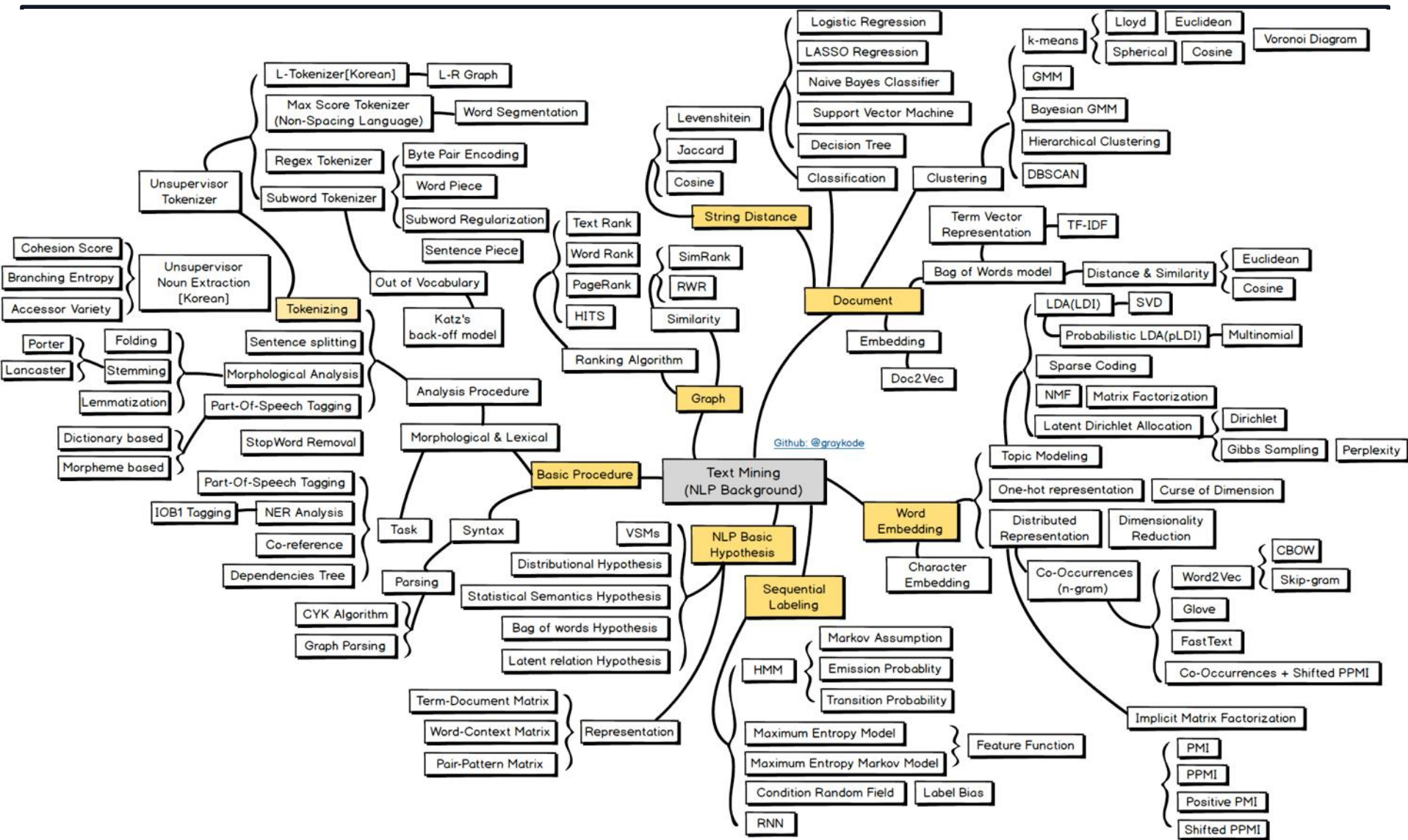
Datawhale

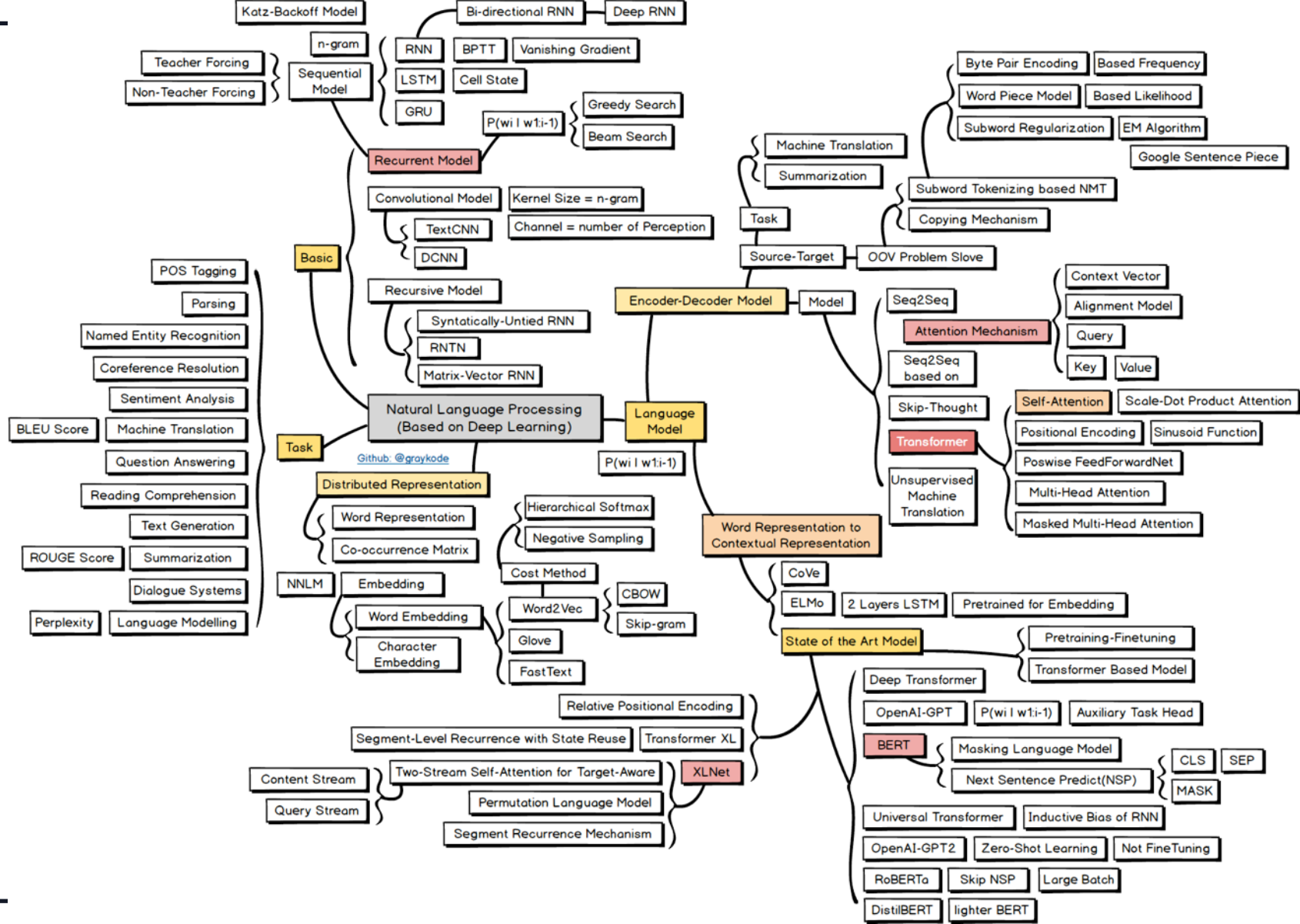
思路1实现思路:

- 理解TF-IDF



Part 3 知识点





Baseline如何继续深入，提高精度？

- ❑ 尝试其他机器学习模型；
- ❑ 对TF-IDF和ngram进行gridsearch；
- ❑ 尝试思路2、思路3、思路4；

其它尝试和思考：

- ❑ 你能分析得出匿名字符中的标点符号吗？
- ❑ 你知道NLP中哪些数据扩增方法呢？
- ❑ 线上F1打分能到0.99吗？

Methods	Prec.	Rec.	F1	Imp.
Lexical Baseline (No Data Augmentation)	.341	.342	.341	—
+ UrbanDictionary Embeddings	.343	.344	.344	0.9%
+ Twitter Embeddings*	.357	.358	.358	4.7%
+ GoogleNews Embeddings*	.364	.366	.365	6.1%
All Features Baseline (No Data Augmentation)	.365	.367	.366	—
+ Lexical (GoogleNews) and Frame-Semantic Embeddings*	.376	.377	.376	2.7%
+ Lexical (Twitter) and Frame-Semantic Embeddings*	.379	.380	.379	3.6%
+ Lexical (UD) and Frame-Semantic Embeddings*	.379	.381	.380	3.8%

- ❑ 你对比赛有什么问题?
- ❑ 你对学习有什么问题?
- ❑ 你对PPT内容有什么问题?

数据竞赛入门讲义

阿水

2020 年 7 月 14 日

目录

1 引言	6
1.1 课程目标	6
1.2 课程受众	7
1.3 课程 SMART 原则	7
1.4 课程 SQ3R 学习法	7
1.5 课程基础	8
2 数据科学必知必会	9
2.1 什么是数据科学?	9
2.2 为什么学习数据科学?	10
2.3 如何学习数据科学?	10
2.4 数据科学包含的知识点	11
2.5 数据竞赛平台	13
2.6 竞赛实例讲解	13
2.6.1 Rong360-用户贷款风险预测	15
2.6.2 Planet-Understanding the Amazon from Space	15

《竞赛入门讲义》

阿水独门秘籍，关注微信公众号
后台回复【讲义】免费获取




Coggle数据科学

微信扫描二维码，关注我的公众号

天池PAI DSW 交流群

389 人



 扫一扫群二维码，立刻加入该群。

天池PAI DSW

免费CPU/GPU资源：



一个专注于AI领域的开源组织

