深度之眼
deepshare.net

Kaggle LECR信息检索大赛赛题分析和Baseline详解

导师：William

深度之眼
deepshare.net

目录

1 赛题分析

Competition Analysis

1 赛题分析

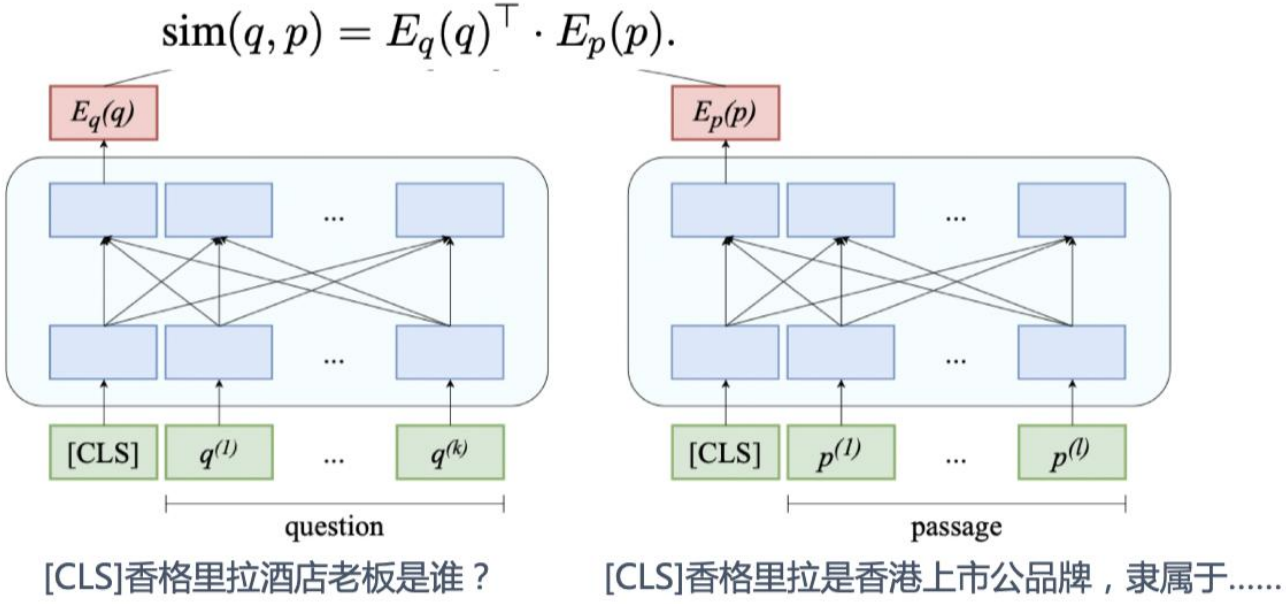
Competition Analysis

目标：本次大赛是NLP中的信息检索任务。本次比赛的目标是简化教育内容与课程中特定主题的匹配过程，将对应主题的内容进行匹配和分类

问题转化：典型的信息检索匹配问题

模型pipleline：title 内容 + doc 内

向量相似度计算



- 可用于快速索引
- 问题和段落之间难以交互

2 数据EDA

Data EDA

本次比赛提供了五份数据分别是content， topic，correlations， 分别存着content文本信息，topic的文本信息以及content和topic的匹配关系，sample_submission.csv提交成绩模版， 其中test, sample_submission为提交答案时用

2 数据EDA

topic 训练文件

id - 一个unique 标记符来表示topic id

title- topic 的文本信息

description - topic 的描述

channel- topic的channel

category- topic的来源

language- 语言

parent- parent id

level- topic数的depth

has_content - 有没有内容

Number of rows in topic data: 76972

Number of columns in topic data: 9

2 数据EDA

content 训练文件

id - 一个unique 标记符来表示content id

title- content 的文本信息

description - content 的描述

language- 语言

kind - 信息format

text- additional 文本信息

has_content - 有没有内容

copyright_holder-版权拥有方

license- license号

Number of rows in content data: 154047

Number of columns in content data: 8

2 数据EDA

correlations训练文件

topic_id - 一个unique 标记符来表示title id

content_ids - 每个topic下对应的content

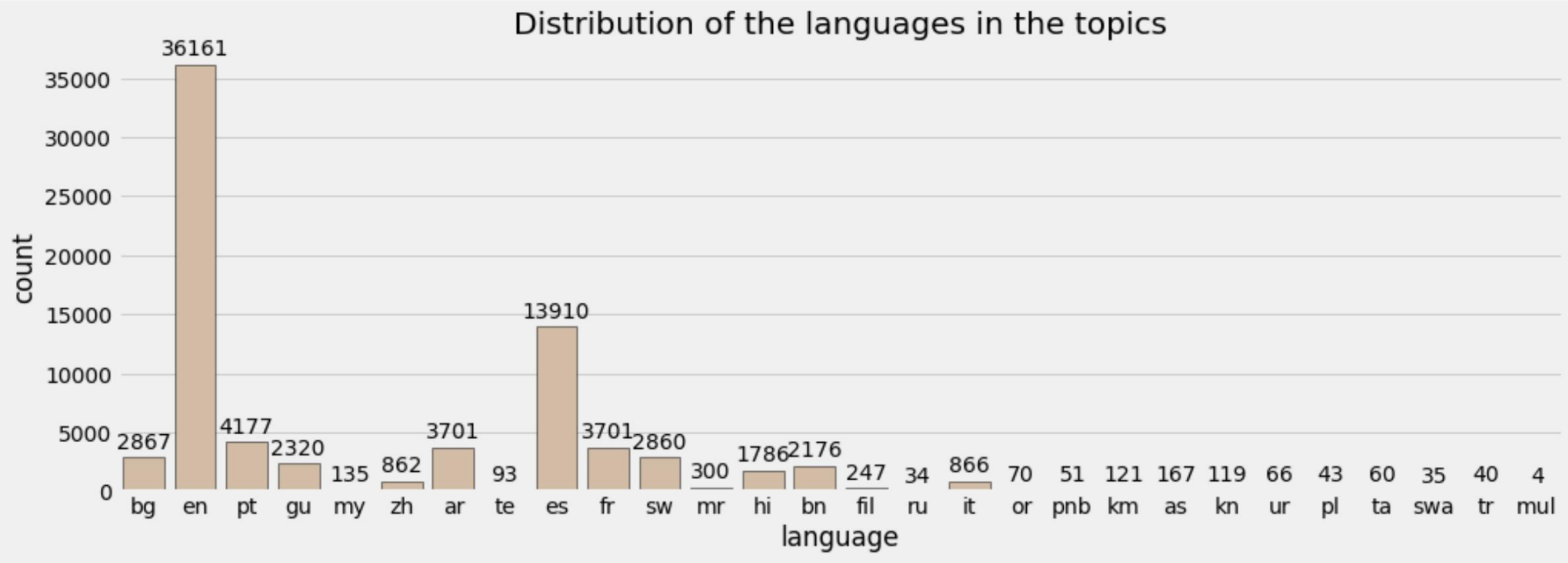
Number of rows in topic data: 61517

Number of columns in topic data: 2

2 数据EDA

深度之眼
deepshare.net

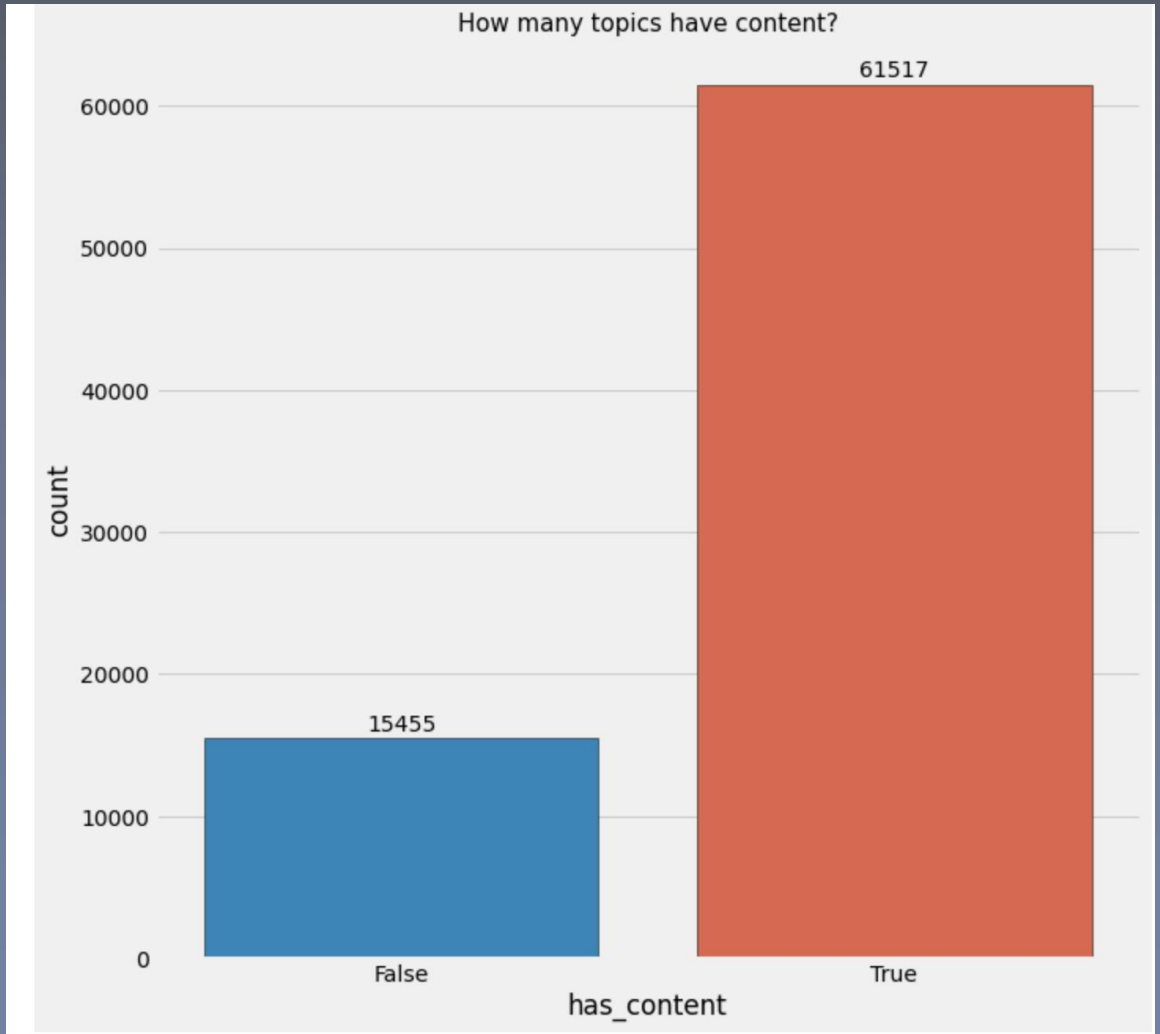
数据分布

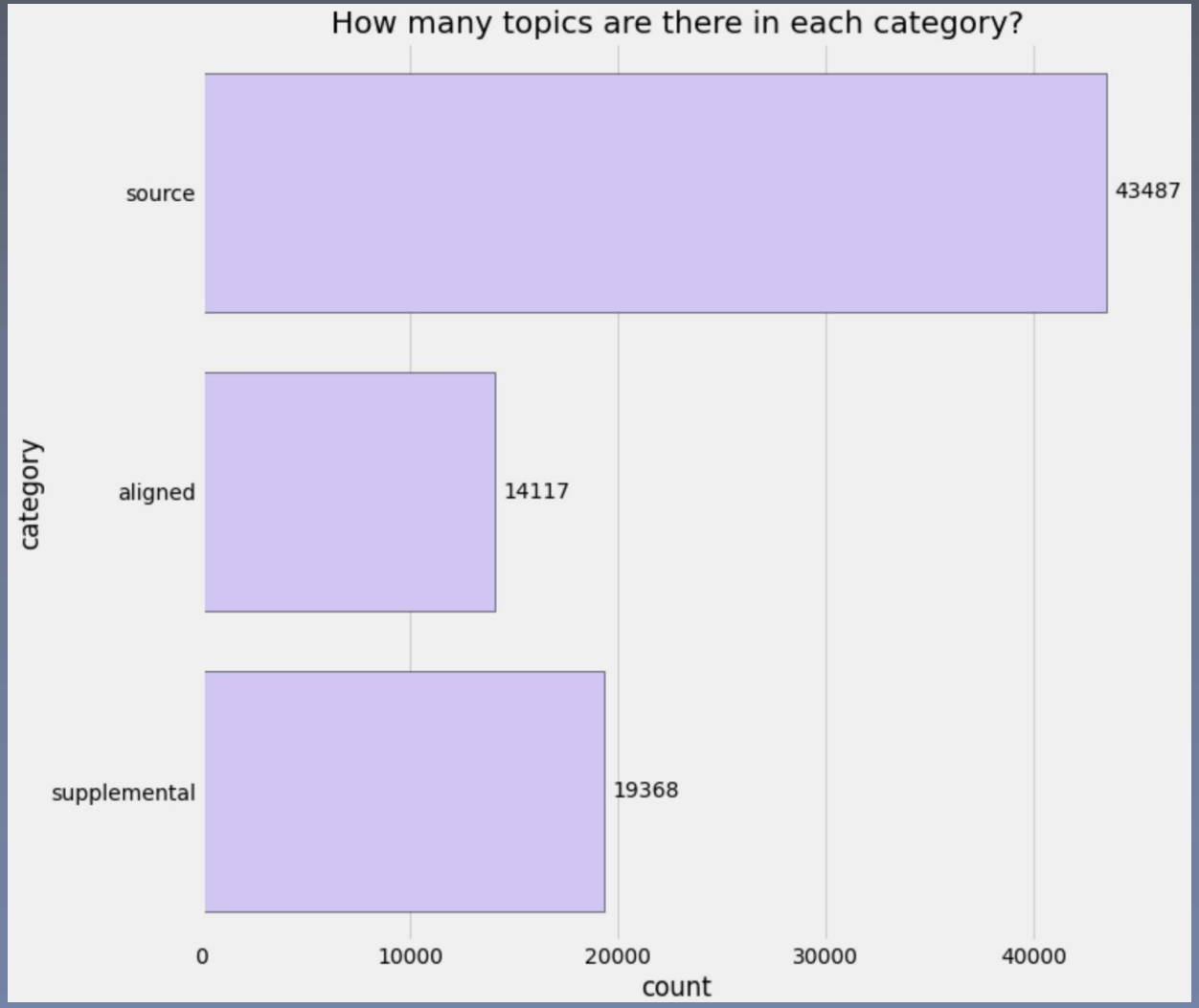


2 数据EDA

深度之眼
deepshare.net

数据分布

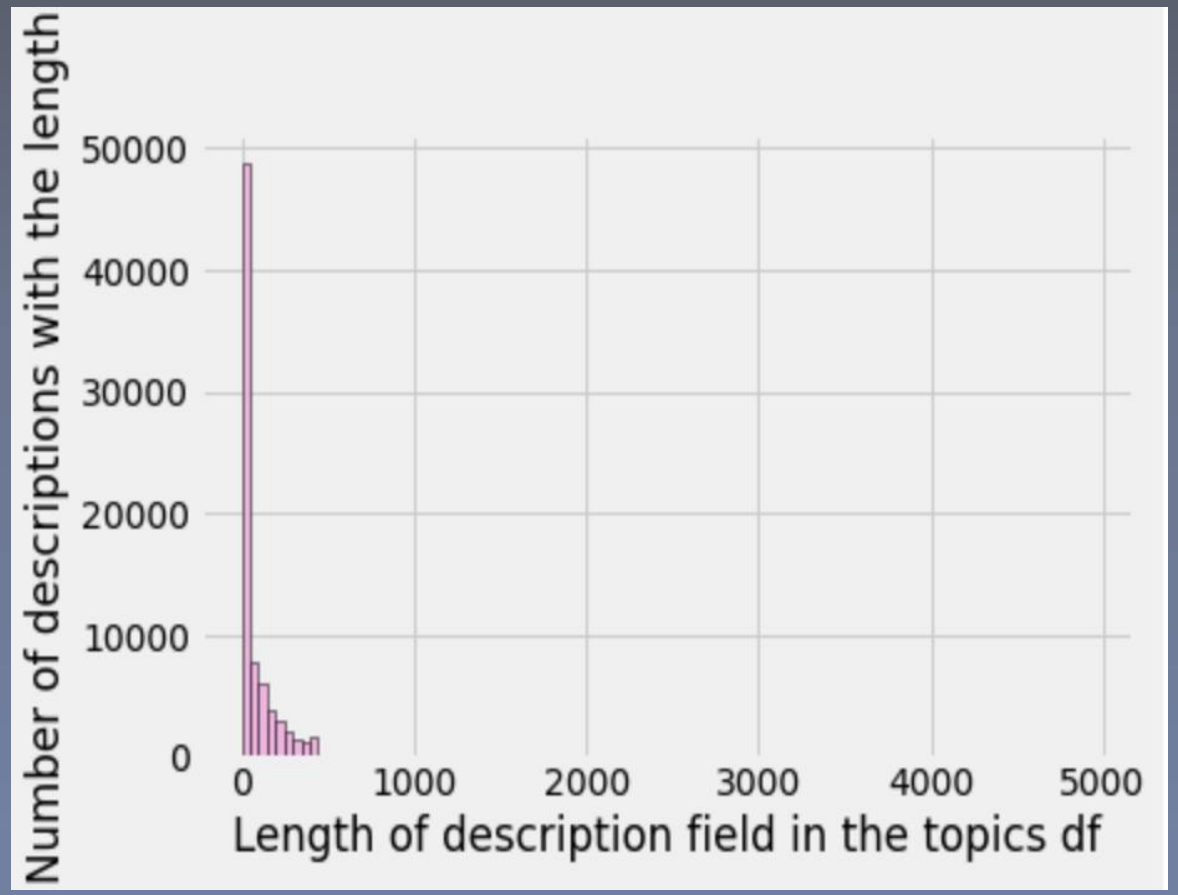


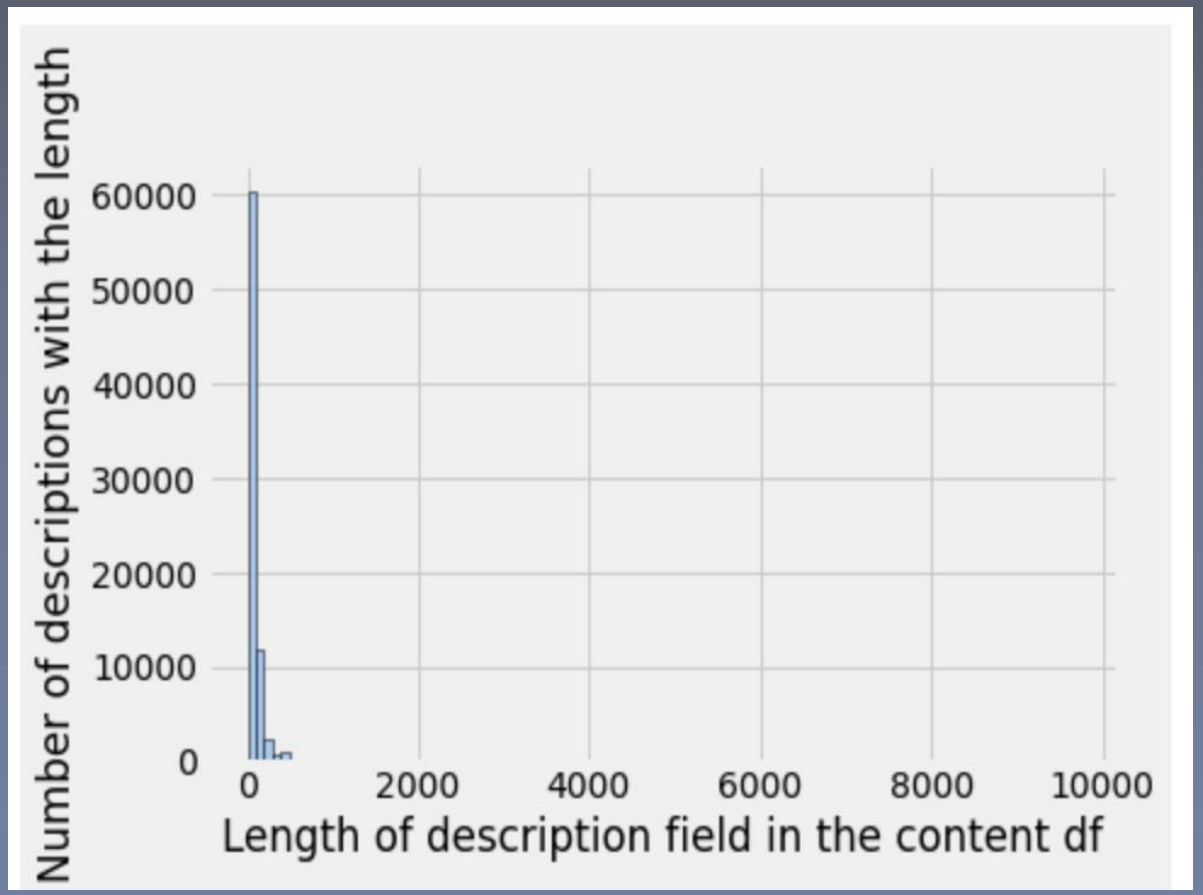


2 数据EDA

深度之眼
deepshare.net

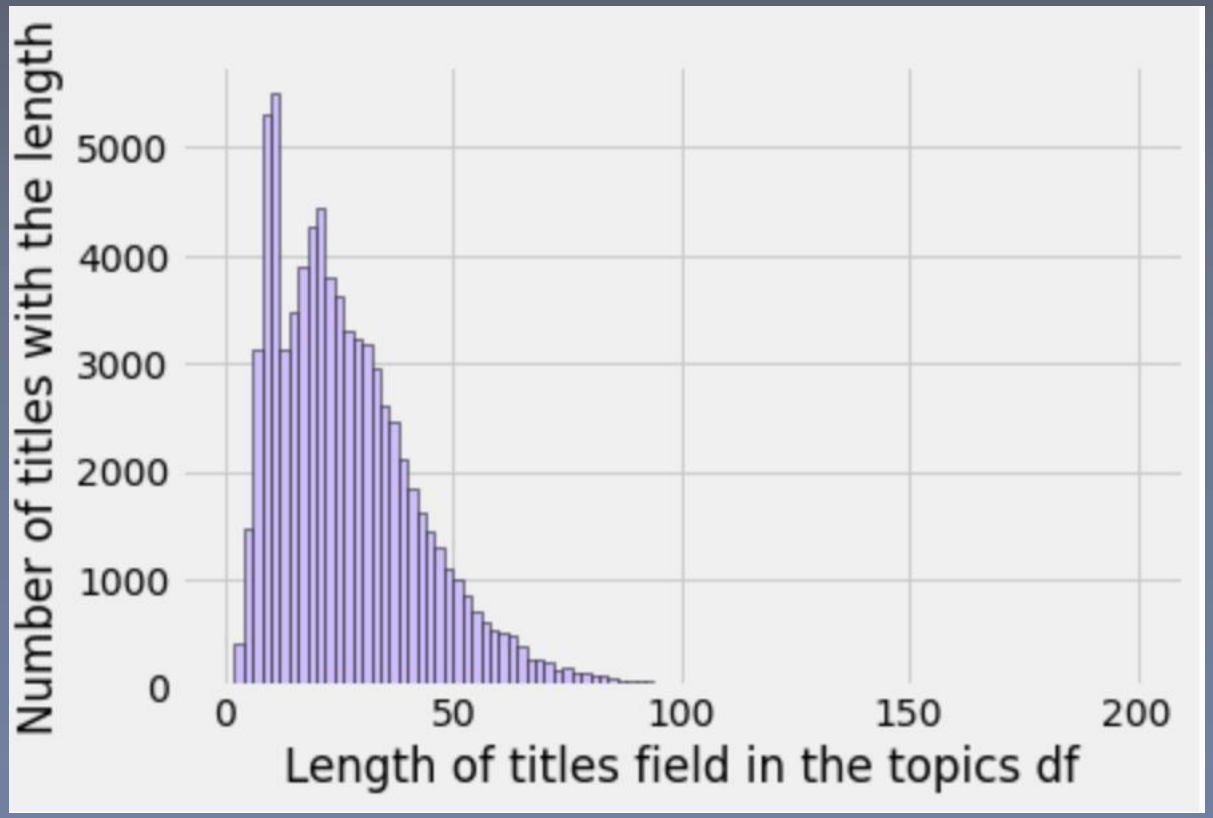
数据分布

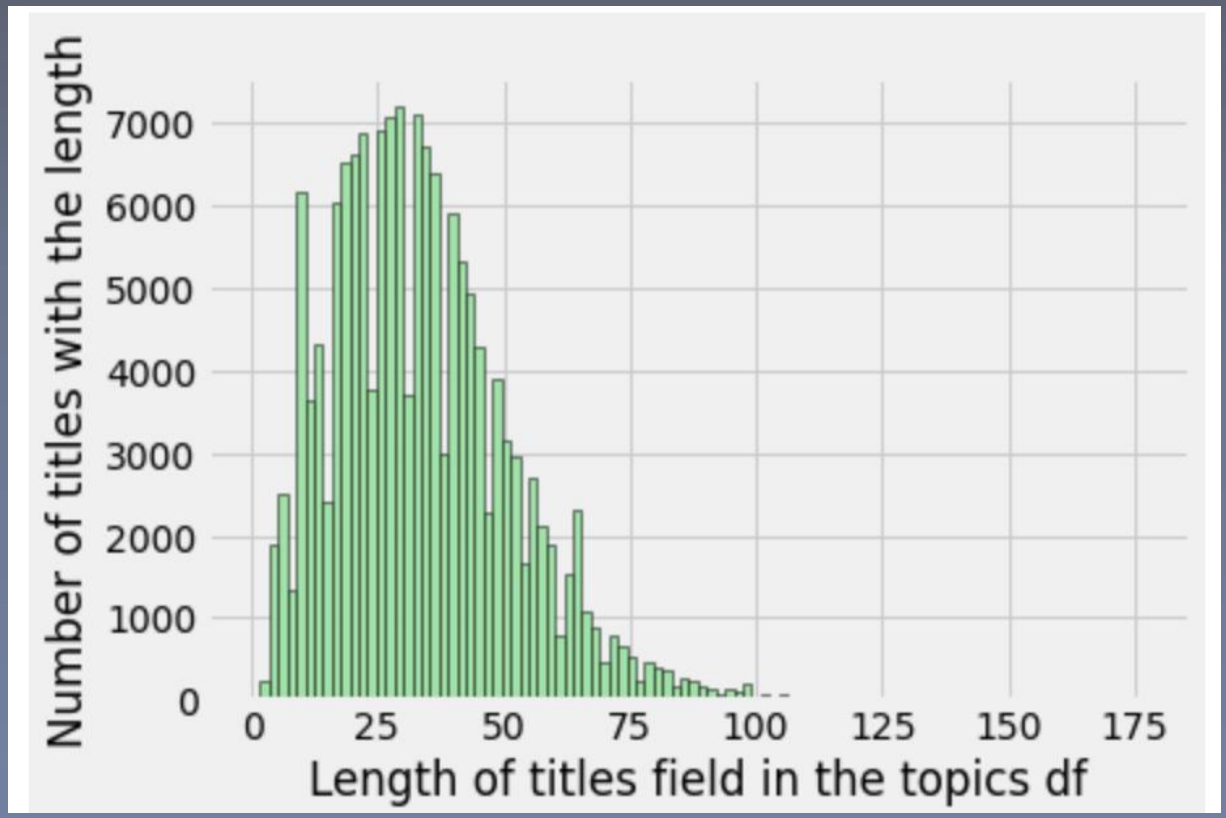




2 数据EDA

数据分布


Number of titles with the length
5000
4000
3000
2000
1000
0
0
50
100
150
200
Length of titles field in the topics df


Number of titles with the length
7000
6000
5000
4000
3000
2000
1000
0
0
25
50
75
100
125
150
175
Length of titles field in the topics df

3 模型选择及训练代码构建

模型选择和标签设计

信息检索任务：

模型需要对topic和content的语义进行抽取，变成一个稠密向量

模型设计上可以参考孪生网路模型设计

topic 和 content 信息一同输入模型中

根据bert的cls位置 + Dense Layer 做分类任务 更新模型

用训练好的模型作为特征抽取器再抽取topic的content存起来

4 提交成绩

有本地GPU算力

1. 本地运行完train 代码

2. 上传自己的weight到kaggle上

3. 线上完成inference

3. 提交自己的成绩

5 答疑互动

请让我们一起立一个flag！

我承诺：

4周努力上TOP100！

结语

再小的细节，也值得被认真对待

deepshare.net

深度之眼

联系我们：

电话：18001992849

邮箱：service@deepshare.net.net

Q Q： 2677693114

公众号

客服微信