Bert可以提取关键词了：KeyBERT的介绍与使用

深度学习自然语言处理 2月20日

点击上方，选择**星标**，每天给你送干货！

来源：ChallengeHub

KeyBert

简介

官方文档：https://maartengr.github.io/KeyBERT/

> KeyBERT是一种小型且容易上手使用的关键字提取技术，它利用BERT嵌入来创建与文档最相似的关键词和关键字短语。

尽管我们已经有许多可用于关键字生成的方法（例如，Rake、YAKE!、TF-IDF等），但是我们还是需要创建一种非常高效并且功能强大的方法来提取关键字和关键字。这就是KeyBERT诞生的初衷！它使用BERT嵌入和简单的余弦相似性来查找文档中与文档本身最相似的子短语。

首先，使用BERT提取文档向量(嵌入)以获取文档级表示。然后，针对N元语法词/短语提取词向量。最后，我们使用余弦相似度来查找与文档最相似的词/短语。然后，可以将最

相似的词识定义为最能描述整个文档的词。

KeyBERT可能不是唯一的提取关键词的方法，它的定位主要是一种用于创建关键字和关键词的快速简便的方法。尽管有很多出色的论文和解决方案都使用BERT嵌入（例如1、2、3），但是很少有直接基于BERT的解决方案，该工具无需从头开始进行训练模型，初学者也可直接使用 :`pip install keybert`

安装

可以直接通过pip安装:

```
pip install keybert
```

使用教程

下面是提取关键词的一个小例子

```
from keybert import KeyBERT

doc = """
        Supervised learning is the machine learning task o
        maps an input to an output based on example inpu
        function from labeled training data consisting of
        In supervised learning, each example is a pair co
        (typically a vector) and a desired output value (
        A supervised learning algorithm analyzes the train
        which can be used for mapping new examples. An o
        algorithm to correctly determine the class labels
        the learning algorithm to generalize from the trai
        'reasonable' way (see inductive bias).
    """
model = KeyBERT('distilbert-base-nli-mean-tokens')
```

我们可以设置`keyphrase_length`来设置生成的keyphrase的长度:

```
>>> model.extract_keywords(doc, keyphrase_ngram_range=(1, 1))
[('learning', 0.4604),
 ('algorithm', 0.4556),
 ('training', 0.4487),
```

```
 ('class',  0.4086),
 ('mapping',  0.3700)]
```

要提取关键字短语，只需将关键字短语_ngram_range设置为（1，2）或更高，具体取决于我们希望在生成的关键字短语中使用的单词数：

```
>>>  model.extract_keywords(doc,  keyphrase_ngram_range=(1,  2))
[('learning  algorithm',  0.6978),
 ('machine  learning',  0.6305),
 ('supervised  learning',  0.5985),
 ('algorithm  analyzes',  0.5860),
 ('learning  function',  0.5850)]
```

更多材料

- https://github.com/thunlp/BERT-KPE
- https://github.com/ibatra/BERT-Keyword-Extractor
- https://github.com/pranav-ust/BERT-keyphrase-extraction
- https://github.com/swisscom/ai-research-keyphrase-extraction

说个正事哈

由于微信平台算法改版，公号内容将不再以时间排序展示，如果大家想第一时间看到我们的推送，强烈建议星标我们和给我们多点点【在看】。星标具体步骤为：

（1）点击页面**最上方**“**深度学习自然语言处理**”，进入公众号主页。

（2）点击**右上角的小点点**，在弹出页面点击“**设为星标**”，就可以啦。

感谢支持，比心❤️。

投稿或交流学习，备注：**昵称-学校（公司）-方向**，进入DL&NLP交流群。

方向有很多：机器学习、深度学习，python，情感分析、意见挖掘、句法分析、机器翻译、人机对话、知识图谱、语音识别等。