翻译进度：已翻译 翻译赏金：0 元 (?) ￥ 我要打赏

默认 原文 打印 双语对照

fastText 快速文本呈现和分类库 —— 来自 Facebook

参与翻译： CY2 (4), 負愚佾歸 (3)

fastText

fastText 是 Facebook 开发的一个用于高效学习单词呈现以及语句分类的开源库。

要求

fastText 使用 C++11 特性，因此需要一个对 C++11 支持良好的编译器，可以使用：

- (gcc-4.6.3 或者更新版本) 或者 (clang-3.3 或者更新版本)

我们使用 Makefile 进行编译，因此需要 make 工具。为了运行单词相似度演示脚本，我们需要如下工具：

- python 2.6 or newer
- numpy & scipy

构建 fastText

第 1 段（可获 1.04 积分）

CY2 1年前 0

邀请好友翻译

0 0 0 分享文章

使用如下命令来构建 fastText 库：

```
$ git clone git@github.com:facebookresearch/fastText.git
$ cd fastText
$ make
```


这将会为所有的类产生一堆文件，包括主二进制文件fasttext．如果你不打算用系统默认的编译器，在Makefile（CC 和 INCLUDES）的头部修改两个宏定义.

使用样例

这个包有两个主要功能：单词特征学习与文本分类．这都在以面两份论文[1] and [2]中有描述

单词特征学习

为了学习单词向量，就像[1]描述的那样：如下操作：

```
$ ./fasttext skipgram -input data.txt -output model
```

data.txt是一个训练文件，包含一些以utf-8编码的文本．默认的这些词向量将会划入字符(3致6个字符)帐目g-grams . 最后的分析程序会保存为两个文件：model.bin 和　model.vec．　model.vec是文本文件包含单词向量，每个单词一行．model.bin是二进制文件包含字典模型参数与所有的其它参数．　这个二进制文件可以用于计算单词向量或重新分析。

前期的训练模型可以从输出单词处计算词向量．假如你有一个文本文件queries.txt包含一些你想切分的单词向量，运用下面的命令：

```
ttext print-vectors model.bin < queries.txt
```

这会将单词向量输出到标准输出，一个向量一行．你也可以使用管道：

```
$ cat queries.txt | ./fasttext print-vectors model.bin
```

上面的脚本只是一个示例，为了更形像点运行：

```
$ ./word-vector-example.sh
```


这将会编译代码，下载数据，计算词向量，并可以测试那些由很少出现的词组成的数据集，测试它们的相似性[例如Thang 等等]．

文本分类

这个类库也可以用来监督文本分类训练，例如情绪分析．[2]里面描述可以用于训练文本分类, 使用：

```
$ ./fasttext supervised -input train.txt -output model
```

train.txt是包含训练语句的文本文件，每行都带有标签，默认情况下，我们假设标签为单词，用前后加下划线的单词表示　如__label__．这个命令将会生成两个文件：model.bin 和　model.vec．一旦模型被训练，你可以评价它，用第一部分来测试计算它的精度：

为了获得一段文本最相似的标签，可以使用如下命令：

```
$ ./fasttext predict model.bin test.txt
```

xt 包含一些文本用来根据每行进行分类。执行完毕将会输出每一行的近似标签。请看 classification-
ple.sh 来了解示例代码的使用场景。为了从论文 [2] 中重新生成结果，可以运行 classification-results.sh 脚本，这将下载所有的数据集并从表1中重新生成结果。

命令完整文档

```
The following arguments are mandatory:
  -input      training file path
  -output     output file path

The following arguments are optional:
  -lr         learning rate [0.05]
  -dim        size of word vectors [100]
  -ws         size of the context window [5]
  -epoch      number of epochs [5]
  -minCount   minimal number of word occurences [1]
  -neg        number of negatives sampled [5]
  -wordNgrams max length of word ngram [1]
  -loss       loss function {ns, hs, softmax} [ns]
  -bucket     number of buckets [2000000]
  -minn       min length of char ngram [3]
  -maxn       max length of char ngram [6]
  -thread     number of threads [12]
  -verbose    how often to print to stdout [1000]
  -t          sampling threshold [0.0001]
  -label      labels prefix [__label__]
```

参考资料

如果使用这些代码用于学习单词的呈现请引用 [1] ，如果用于文本分类请引用 [2]。

[1] P. Bojanowski*, E. Grave*, A. Joulin, T. Mikolov, *Enriching Word Vectors with Subword Information*

```
@article{bojanowski2016enriching,
  title={Enriching Word Vectors with Subword Information},
  author={Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas},
  journal={arXiv preprint arXiv:1607.04606},
  year={2016}
}
```

[2] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, *Bag of Tricks for Efficient Text Classification*

```
@article{joulin2016bag,
  title={Bag of Tricks for Efficient Text Classification},
  author={Joulin, Armand and Grave, Edouard and Bojanowski, Piotr and Mikolov, Tomas},
  journal={arXiv preprint arXiv:1607.01759},
  year={2016}
}
```

(* 这些作者贡献一样.)

加入 fastText 社区

- Facebook page: https://www.facebook.com/groups/1174547215919768
- Contact: egrave@fb.com, bojanowski@fb.com, ajoulin@fb.com, tmikolov@fb.com

请阅读 CONTRIBUTING 文件了解更详细信息。

许可证

fastText 使用 BSD 许可证，我们同时提供了一个附加的专利授权。

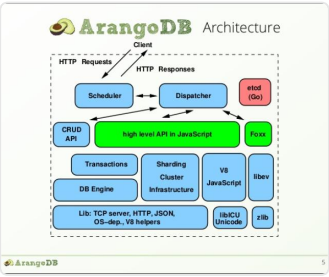
第 7 段 (可获 0.64 积分)

CY2
1年前

0

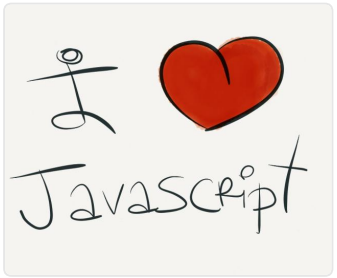
¥ 打赏译者

相关文章



ArangoDB 中集成的 RocksDB 存储引擎常见问题解答

How I Convinced Our CTO to Switch From CoffeeScript to ES6



10 个最终编译成 JavaScript 的脚本语言

怎样在中年时期建立韧性

原文：fastText: Library for fast text representation and classification. / fastText 快速文本呈现和分类库 —— 来自 Facebook

作者：facebookresearch

频道：计算机

发布：CY2 (2016-08-05)

标签： Facebook

文章评论

邀请好友翻译

0 0 0 分享文章

对此文有什么看法请在这里发表评论

请输入左图的验证码

发表评论