AI训练营-nlp项目内容

Week1：句子语义相似度预测

- 学习nltk，gensim等基本自然语言处理工具
- 学习机器学习基本工具Sklearn
- 相似度计算理论与方法学习

学习指引

Day1: 学习使用nltk,python，同时掌握nlp的一些基本处理技巧和方法，例如tokenize，stem等

今天我们主要把推荐书籍的第五章节前面的内容，选择性看一遍（根据自身的情况选择），同时把里面涉及的python代码敲一敲(做文中的example)。主要学习python的数据类型和使用，以及nlp的一些基本概念。

1. 安装nltk

   http://www.nltk.org/install.html

2. 学习 《Natural Language Processing with Python》一书前6章节，进行文本tokenize，stem等操作

   http://www.nltk.org/book/

Day 2: 安装使用 gensim

今天我们主要学习使用nlp中比较流行的一种词向量生成方法word2vec，gensim是其在python环境下的实现。除此之外，gensim本身还带有其他的一些例如特征提取，主题模型计算等功能

1. 安装gensim

   https://radimrehurek.com/gensim/install.html

2. Gensim 主要功能介绍，学习tf-idf提取等功能

   https://radimrehurek.com/gensim/tutorial.html

3. Gensim 工具例子: 学习使用gensim来提取，词向量等特征

   https://radimrehurek.com/gensim/models/word2vec.html

Day3: scikit-learn 安装和例子学习，学习使用工具进行拟合

scikit-learn是机器学习领域广泛被使用的一个python工具，里面集成了大量的机器学习算法。今天我们学习其中的Ridge Regression 和 svr方法，Ridge Regression使用比较简单，svr是深度学习前的主流方法(里面现在也集成了深度学习的工具，有兴趣的同学可以尝试)。后面可以选择使用这些方法对相似度数据进行拟合。

1. 官网主页
   http://scikit-learn.org/stable/
2. 安装 scikit-learn
   http://scikit-learn.org/stable/install.html
3. 学习使用 scikit-learn 进行拟合
   Ridge Regression
   http://scikit-learn.org/stable/modules/linear_model.html#ridge-regression
   Svr
   http://scikit-learn.org/stable/modules/svm.html#regression

   说明：
   1. 若由于某些原因链接打不开，可以从主页点进去试试
   Ridge Regression：主页 -> Regression -> 1.1.2 Ridge Regression
   svr: 主页 -> Regression -> 1.4.2. Regression
   2. 对拟合概念不太理解的同学可参考：
   https://zh.wikipedia.org/wiki/%E6%9B%B2%E7%B7%9A%E6%93%AC%E5%90%88
   Andrew Ng课程第三章：http://open.163.com/special/opencourse/machinelearning.html

Day4&5:阅读论文，学习语义相似度计算的方法

前面几天我们已经学习了一些基本的工具和方法，接下来需要开始思考如何使用自己所学习到的内容（可以不在这几天的课程之类）来完成相似度预测的任务。以下几篇论文作为参考方法，可以采用提到的任意一种方法作为最终的实现方案，有余力的同学也可以实现多个方案。

参考阅读材料：

http://nlp.arizona.edu/SemEval-2017/pdf/SemEval025.pdf
http://www.aclweb.org/anthology/S/S14/S14-2039.pdf
http://nlp.arizona.edu/SemEval-2017/pdf/SemEval001.pdf

Day6: 参考方案，进行相似度预测

如果大家已经完成自己系统的搭建并提交预测数据，那么恭喜。如果没有，可以参考下面提供的链接中的做法，计算两个句子的相似度后再和人工的语义打分进行拟合。最后提交结果。

参考链接：

http://www.52nlp.cn/%E5%A6%82%E4%BD%95%E8%AE%A1%E7%AE%97%E4%B8%A4%E4%B8%AA%E6%96%87%E6%A1%A3%E7%9A%84%E7%9B%B8%E4%BC%BC%E5%BA%A6%E4%BA%8C