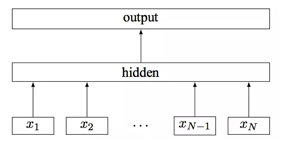
Text Classification

# 任务

* 对dbpedia.train进行训练，对dbpedia.test进行测试。
* 算法选择
  + fastText
  + CNN
  + LSTM
* 算法要求
  + 所有算法获得的准确率不低于85%，F1值不低于0.8。

# 算法介绍

* fastText
  + 通过片段中词向量（x1, x2, …, xn）的数值预测类别，原理和word2vec的cbow相似，cbow用上下文预测中心词，fasttext用全部的n-gram预测类别。
  + 
  + 通过softmax 对输出值进行归一化映射。
* CNN
  + 输入：把每个词变为k维的词向量，每个句子就是Nxk的矩阵（N是句子的长度）
  + 卷积层：卷积核进行一维的滑动，卷积核的款为k，长度为n-gram中的n
  + 池化：max-pool，减少模型参数
  + 全连接：softmax
* LSTM
  + RNN会出现long-term dependencies，而LSEM不会。
  + LSTM的cell中有几个gate可以用于增加和删除信息。
    - 是否遗忘



* + - 是否储存



* + - 是否更新



* + - 是否输出



# 数据情况

\_\_label\_\_7 caddo lake drawbridge the historic caddo lake drawbridge at mooringsport louisiana is a vertical-lift bridge that is listed on the u . s . national register of historic places . it was built in 1914 to replace a ferry by the midland bridge company of kansas city missouri under authority of the caddo parish police jury . the lift span has been inoperable since the 1940s . this vehicular bridge illustrates the vertical-lift design of john alexander low waddell of the firm of waddell & harrington .

\_\_label\_\_9 kolga tartu county kolga tartu county is a village in nõo parish tartu county in eastern estonia .

\_\_label\_\_13 the horse of pride the horse of pride is a 1980 film directed by claude chabrol . its title in french is le cheval d ' orgueil . it is based on le cheval d ' orgueil an autobiography by pêr-jakez helias . the film takes place in the bigouden area south of quimper .

上述例子中\_\_label\_\_7，表示文本caddo lake drawbridge the historic caddo lake drawbridge at mooringsport louisiana is a vertical-lift bridge that is listed on the u . s . national register of historic places .所属的类别标签。这里无需知晓label3的具体含义，对最终结果无任何影响。

训练集560000条，测试集70000条。

# 实验结果

## Fasttext

主要修改了词向量的向量数，学习率，softmax的方式和迭代次数。

首先，在默认softmax方式为负采样和默认迭代次数为5的前提下，和调整词向量维度。

(70000, 0.9816571428571429, 0.9816571428571429) 0.9816571428571429 dim = 10

(70000, 0.9822285714285715, 0.9822285714285715) 0.9822285714285715 dim = 50

(70000, 0.9821, 0.9821) 0.9821 dim = 100

(70000, 0.9819142857142857, 0.9819142857142857) 0.9819142857142857 dim = 150

(70000, 0.9819428571428571, 0.9819428571428571) 0.9819428571428571 dim = 200

(70000, 0.9819571428571429, 0.9819571428571429) 0.9819571428571429 dim = 250

可以发现，当词向量维度为50已经足够表示一个词汇。

在词向量维度为50后，调整学习率。

(70000, 0.9785285714285714, 0.9785285714285714) 0.9785285714285714 lr = 0.02

(70000, 0.9807857142857143, 0.9807857142857143) 0.9807857142857143 lr = 0.03

(70000, 0.9815285714285714, 0.9815285714285714) 0.9815285714285714 lr = 0.04

(70000, 0.9822571428571428, 0.9822571428571428) 0.9822571428571428 lr = 0.05

(70000, 0.9826142857142857, 0.9826142857142857) 0.9826142857142857 lr = 0.06

(70000, 0.9828571428571429, 0.9828571428571429) 0.9828571428571429 lr = 0.07

(70000, 0.9831, 0.9831) 0.9831 lr = 0.08

(70000, 0.9834142857142857, 0.9834142857142857) 0.9834142857142857 lr = 0.09

然后，在学习率为0.05后，调整softmax方式

(70000, 0.9828142857142858, 0.9828142857142858) 0.9828142857142858 loss = ns

(70000, 0.9751, 0.9751) 0.9751 loss = hs

(70000, 0.9834285714285714, 0.9834285714285714) 0.9834285714285714 loss = softmax

在选择默认softmax之后，调整迭代次数

(70000, 0.9815571428571429, 0.9815571428571429) 0.9815571428571429 epoch = 2

(70000, 0.9826714285714285, 0.9826714285714285) 0.9826714285714285 epoch = 3

(70000, 0.9831714285714286, 0.9831714285714286) 0.9831714285714286 epoch = 4

(70000, 0.9834857142857143, 0.9834857142857143) 0.9834857142857143 epoch = 5

(70000, 0.9834, 0.9834) 0.9834 epoch = 6

(70000, 0.9835714285714285, 0.9835714285714285) 0.9835714285714285 epoch = 7

(70000, 0.9833714285714286, 0.9833714285714286) 0.9833714285714286 epoch = 8

(70000, 0.9833142857142857, 0.9833142857142857) 0.9833142857142857 epoch = 9

在迭代次数为5后，F1值变化不大，故选择迭代次数为5

F1 = 0.9834857142857143

达到要求