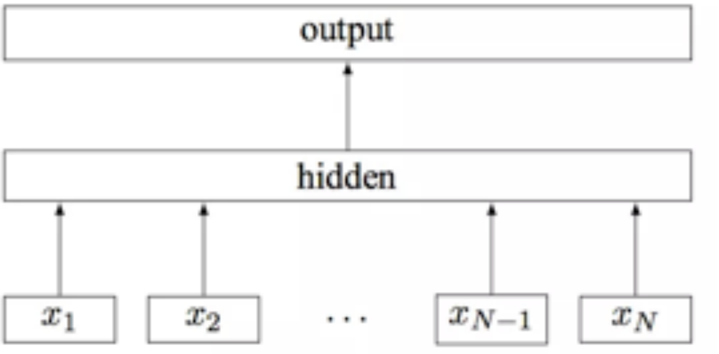
Text Classification

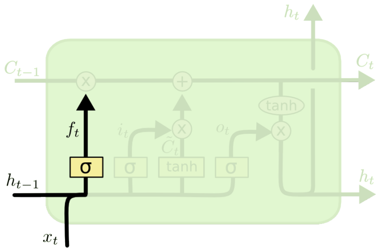
任务

- 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

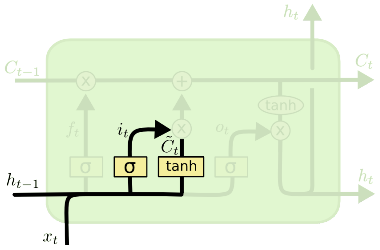
- fastText
  - 通过片段中词向量（x1, x2, ⋯, xn）的数值预测类别，原理和 word2vec 的 cbow 相似，cbow 用上下文预测中心词，fasttext 用全部的 n-gram 预测类别。
  - 

  - 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



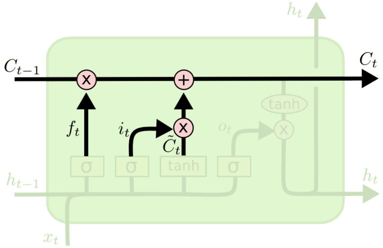
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

- 是否储存



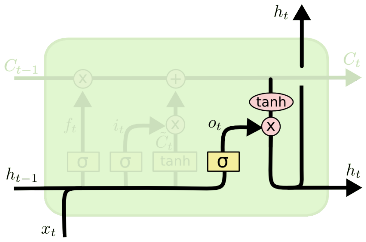
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

- 是否更新



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

- 是否输出



$$o_t = \sigma\left(W_o \ [h_{t-1}, x_t] \ + \ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

○

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 .

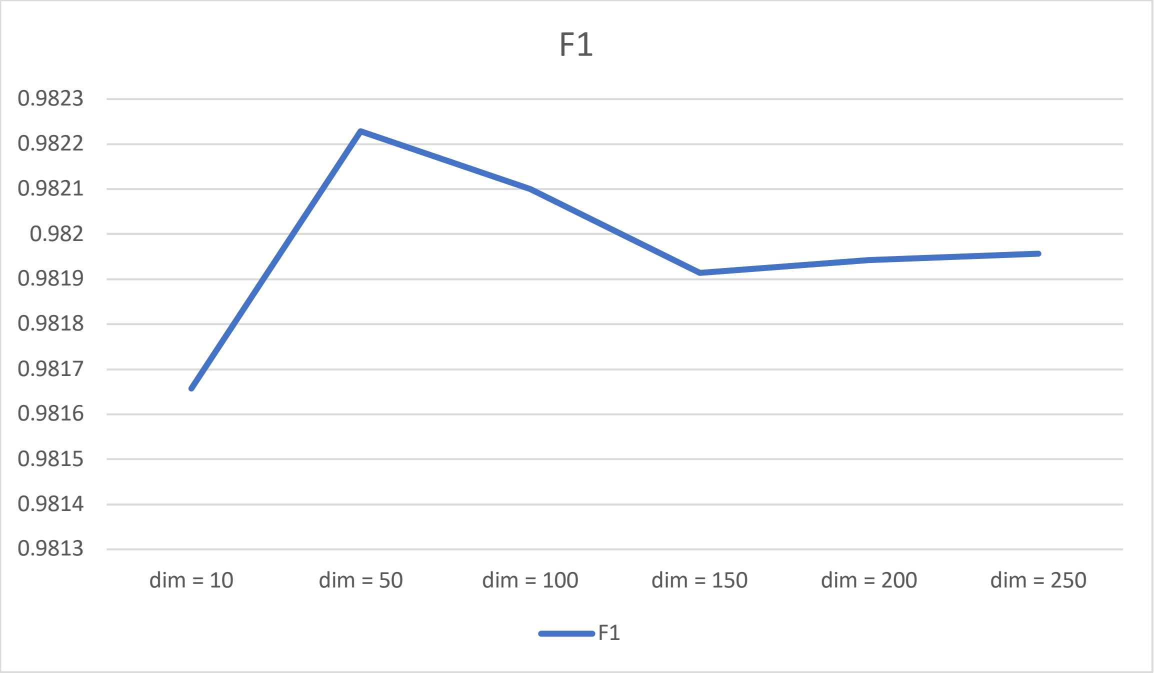
__label__13 the horse of pride the horse of pride is a 1980 film directed by claude chabrol . its title in french is le cheval d ' orgueil . it is based on le cheval d ' orgueil an autobiography by pêr-jakez helias . the film takes place in the bigouden area south of quimper .
上述例子中__label__7，表示文本 caddo 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。

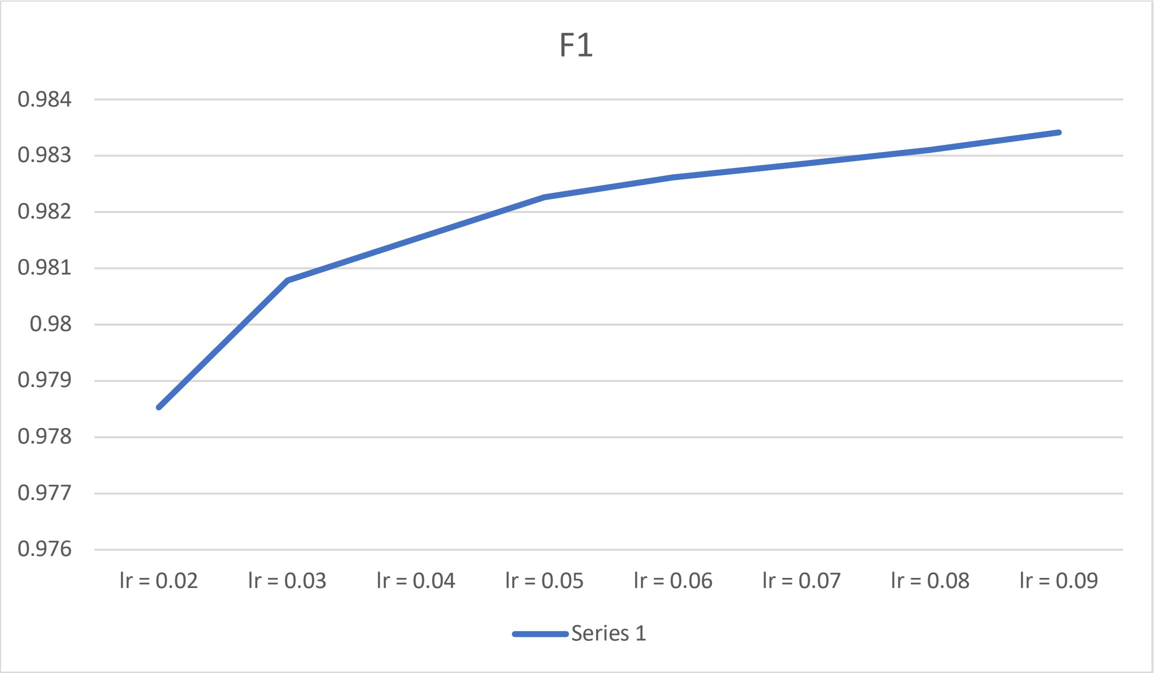
```
(70000, 0.9816571428571429, 0.9816571428571429) 0.9816571428571429 dim = 10
(70000, 0.9822285714285715, 0.9822285714285715) 0.9822285714285715 dim = 50
(70000, 0.9821, 0.9821) 0.9821 dim = 100
(70000, 0.9819142857142857, 0.9819142857142857) 0.9819142857142857 dim = 150
(70000, 0.9819428571428571, 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。

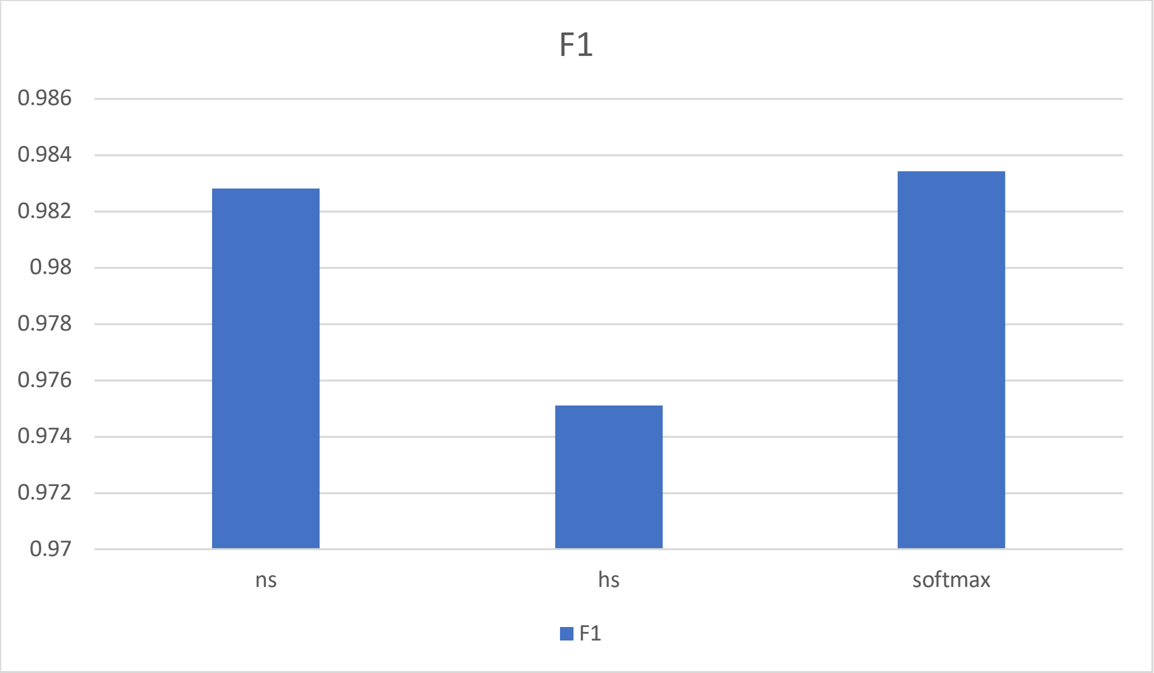
```
(70000, 0.9785285714285714, 0.9785285714285714) 0.9785285714285714 lr = 0.02
(70000, 0.9807857142857143, 0.9807857142857143) 0.9807857142857143 lr = 0.03
(70000, 0.9815285714285714, 0.9815285714285714) 0.9815285714285714 lr = 0.04
```

```
(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
```



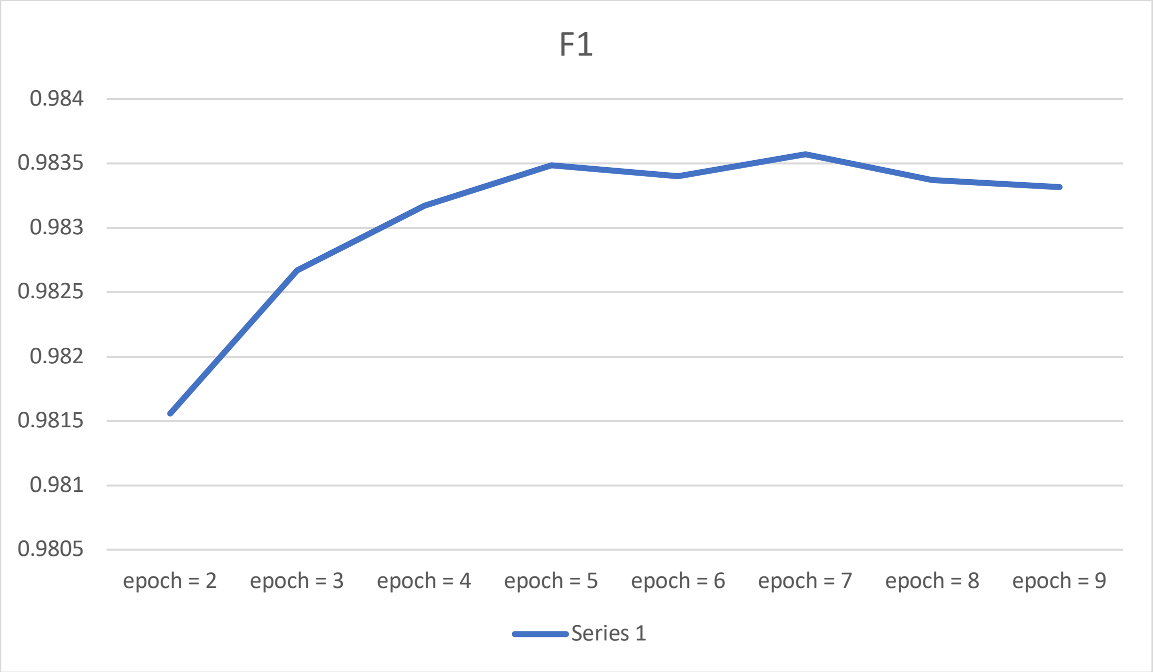
然后，在学习率为 0.05 后，调整 softmax 方式

```
(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