



# SMP CUP 2017 CSDN 用户画像评测



# 目录

---



Southeast University

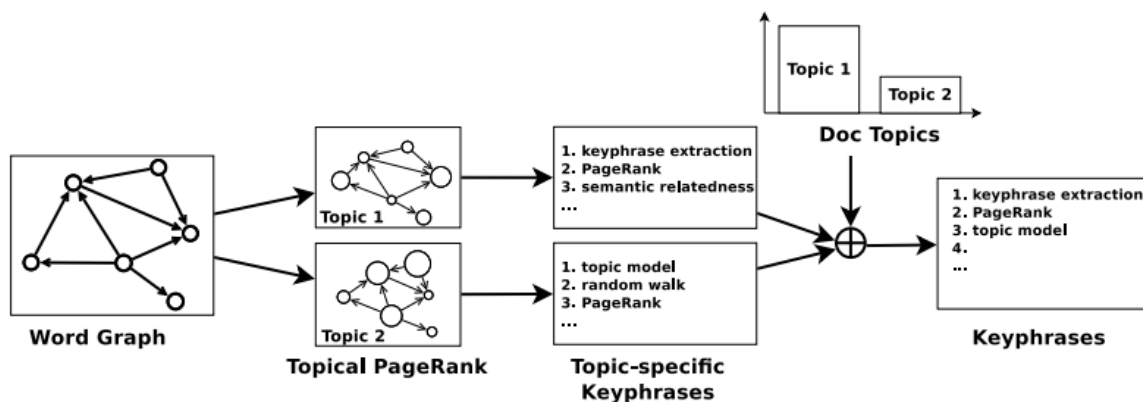
- 简介
- Task1文档关键词抽取
- Task2用户兴趣标注
- Task3用户成长预测
- 总结

## ■ 文档关键词抽取

### Unsupervised methods

Tfidf / TextRank / Topical Page Rank

### Supervised methods ( classification )



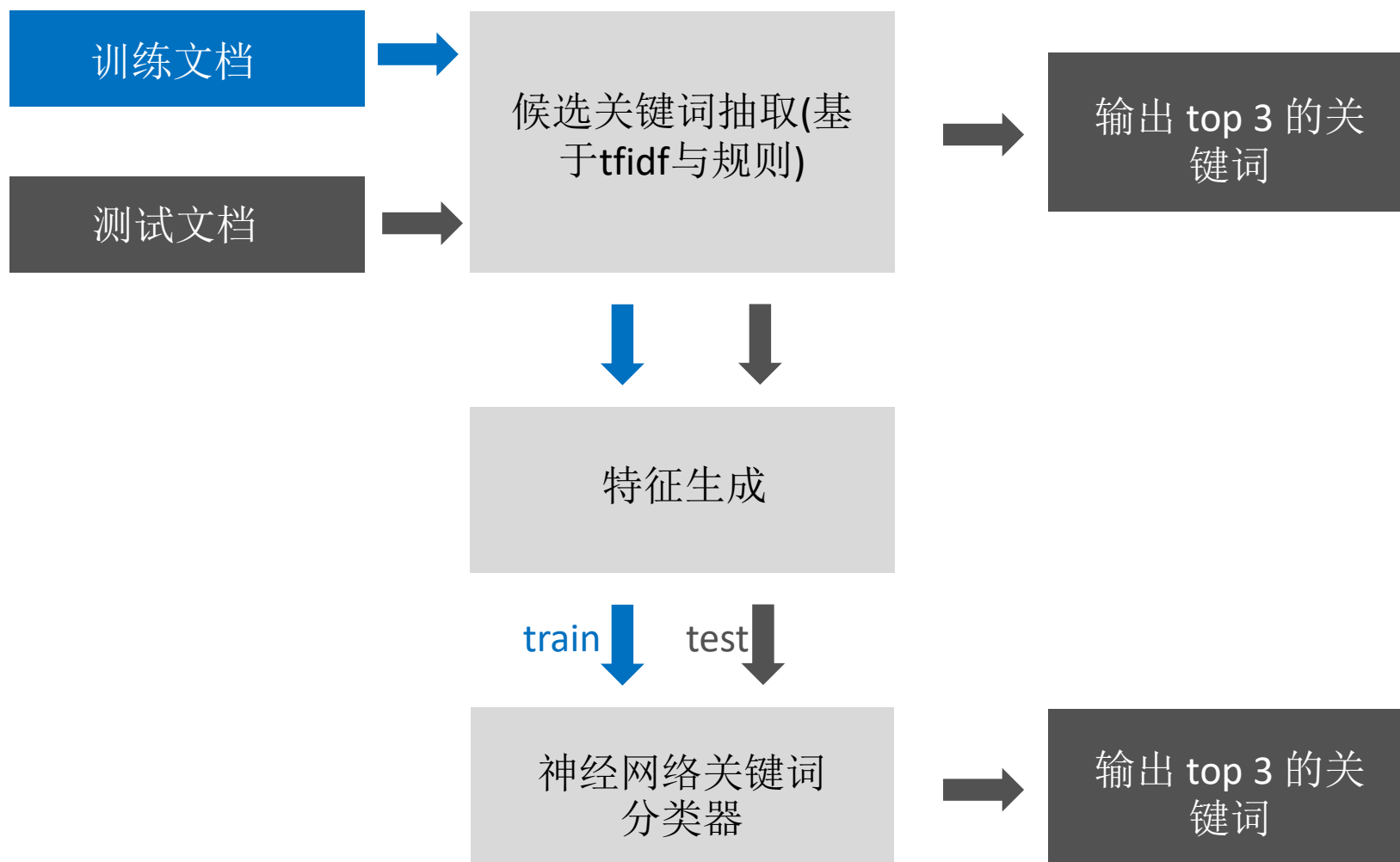
## ■ 用户兴趣标注 (classification)

## ■ 用户成长预测 (regression)



## Task1 文档关键词抽取

## Task1文档关键词抽取



## Task1文档关键词抽取

### ■ 分词结果错误修正

k - means  $\rightarrow$  k-means

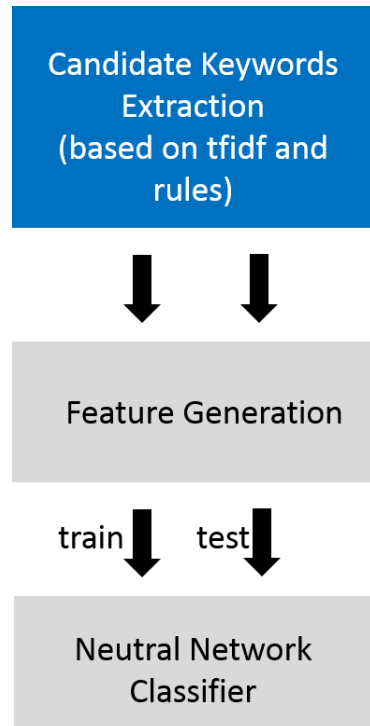
### ■ 偏向于长词的规则

$\text{len}(\text{'数据'}) < \text{len}(\text{'数据分析'})$

$\text{score}(\text{'数据分析'}) += \text{score}(\text{'数据'})$

### ■ 偏向于标题中的词的规则

出现在标题中的词累计词频时有较大的权重



# Task1文档关键词抽取

## ■ 形态特征

单词长度

是否包含数字字母

## ■ 位置特征

是否出现在标题中

首次出现在文档中的相对位置

## ■ 统计特征

词频

逆文档频率

Tfidf

## ■ 主题特征

单词主题分布与文档主题分布的余弦相似度

## ■ 语义语法特征

单词词向量与文档向量的余弦相似度

Candidate Keywords  
Extraction  
(based on tfidf and  
rules)



Feature Generation



Neutral Network  
Classifier

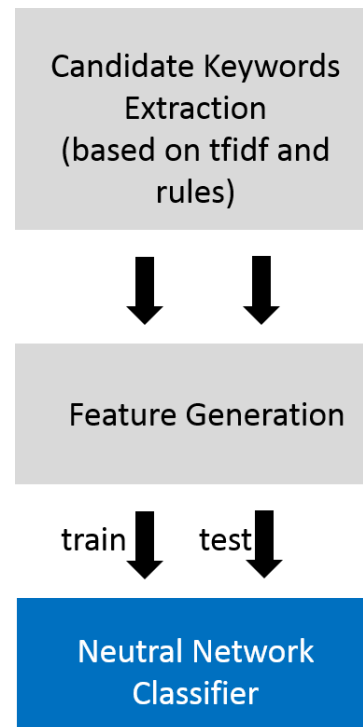


## Task1文档关键词抽取



Southeast University

- 三层神经网络（隐层128， batch\_size 32, epochs 200）
- 该分类器的instance是单词
- 对于某个文档，抽取出n个候选关键词，如果是5个正确关键词之一， label为1否则为0，这样就构建出了训练集。







## Task1文档关键词抽取



Southeast University

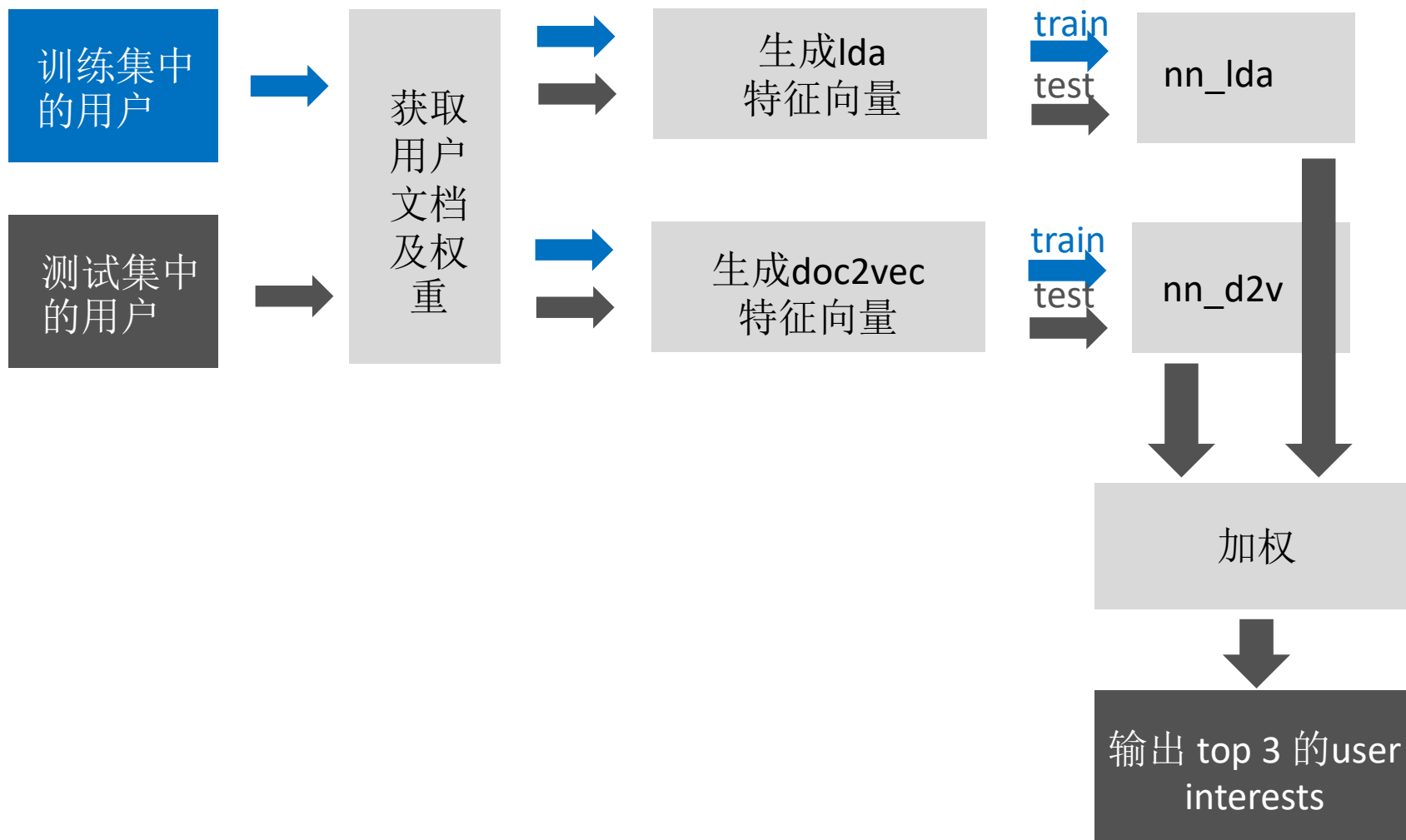
	Task1
Validation set	0.6216
Test set	0.6046



## Task2用户兴趣标注



## Task2用户兴趣标注





文档类型	权重
Post	20
Browse	5
Comment	10
Vote up	10
Favorite	20

- 一个用户的重复的文档只考虑一次，取最大的权重
- 用户级别的特征向量由文档级别的特征向量按权重加权得来



### nn\_lda

- 三层神经网络（隐层64， batch\_size 5, epochs 100）

### nn\_d2v

- 三层神经网络（隐层30， batch\_size 5, epochs 50）

### 预测结果

- 两个神经网络分别输出一个42维的向量，对应42个user interests，经过在训练集中划分出的验证集上的验证，使用的权重为nn\_lda占0.9， nn\_d2v占0.1。



## Task2用户兴趣标注



Southeast University

	Task2
Validation set	0.4712
Test set	0.4579



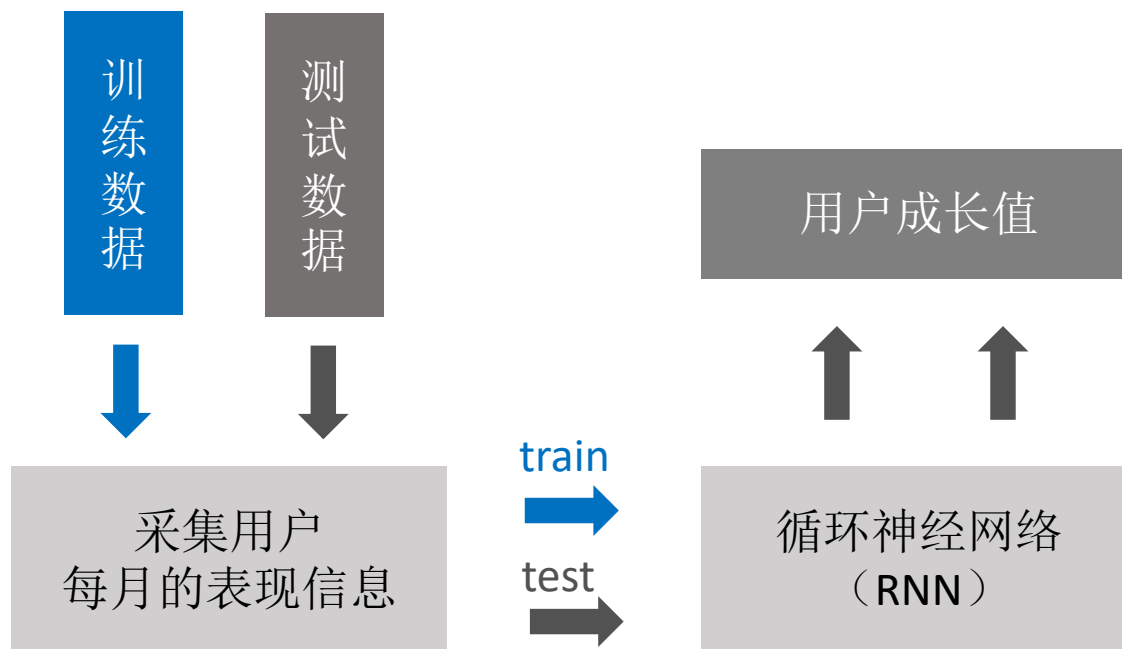
## Task3用户成长值预测

# Task3成长值预测



Southeast University

利用循环神经网络考虑用户的表现随时间的变化





# Task3成长值预测

特征组成（每月）：

## ■ 用户的活跃程度特征：

发表博客数量、浏览博客数量、评论博客数量、点赞博客数量、点踩博客数量、收藏博客数量、私信他人次数、被私信次数。  
（8维特征）

## ■ 用户发表博客的质量：

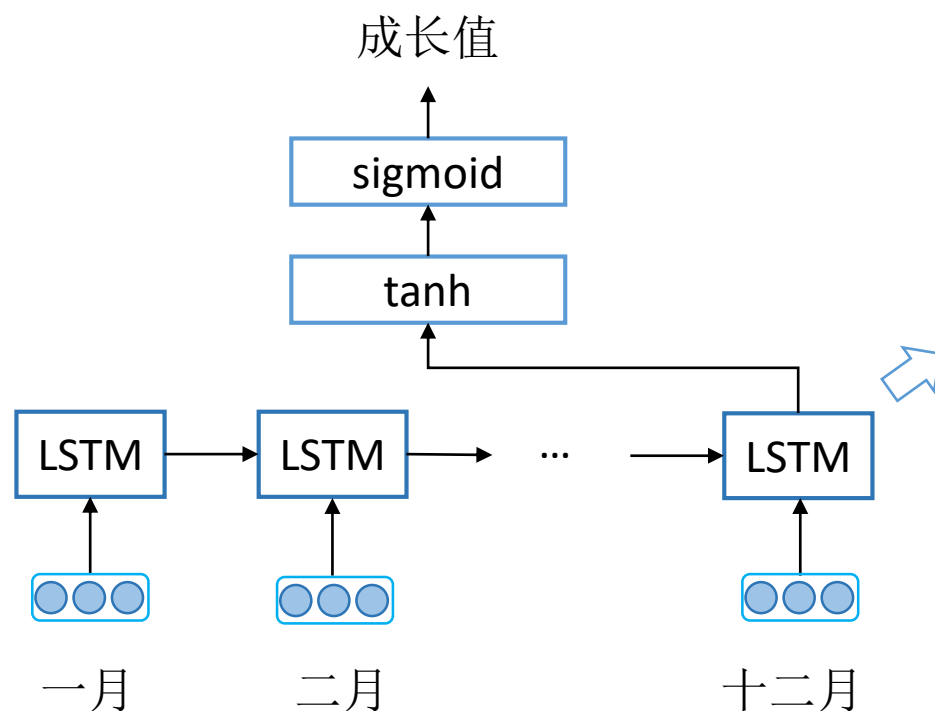
被浏览次数、被评论次数、被点赞次数、被点踩次数、被收藏次数以及前述所有频数的平均值。（10维特征）

## ■ 用户发表博客的特征极值：

被浏览次数、被评论次数、被点赞次数、被点踩次数、被收藏次数皆取最大值。（5维特征）

# Task3成长值预测

模型框架：

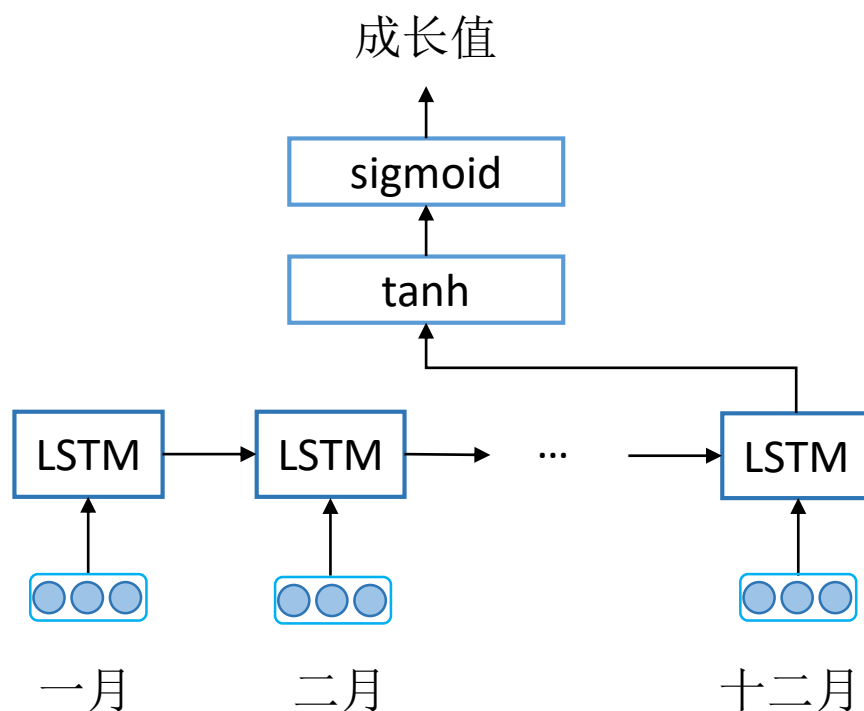


## LSTM

$$\begin{aligned}
 \mathbf{i}_t &= \sigma(\mathbf{W}^{(i)} \mathbf{x}_t + \mathbf{U}^{(i)} \mathbf{h}_{t-1} + \mathbf{b}^{(i)}) , \\
 \mathbf{o}_t &= \sigma(\mathbf{W}^{(o)} \mathbf{x}_t + \mathbf{U}^{(o)} \mathbf{h}_{t-1} + \mathbf{b}^{(o)}) , \\
 \mathbf{f}_t &= \sigma(\mathbf{W}^{(f)} \mathbf{x}_t + \mathbf{U}^{(f)} \mathbf{h}_{t-1} + \mathbf{b}^{(f)}) , \\
 \mathbf{u}_t &= \tanh(\mathbf{W}^{(u)} \mathbf{x}_t + \mathbf{U}^{(u)} \mathbf{h}_{t-1} + \mathbf{b}^{(u)}) , \\
 \mathbf{c}_t &= \mathbf{i}_t \odot \mathbf{u}_t + \mathbf{f}_t \odot \mathbf{c}_{t-1} , \\
 \mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t) ,
 \end{aligned}$$

# Task3成长值预测

模型框架：



batch size: 15  
epoch: max 200

tanh层输出: 100维

LSTM隐藏层: 150维  
(dropout: 0.1)

特征向量: 12\*23维

# Task3成长值预测



Southeast University

	Task3
Validation set	0.7475
Test set	0.7495



# 总结



Southeast University

一些尝试:

## ■ Task1

一些其他特征;

## ■ Task2

当做multi-label问题优化ranking损失;

上采样、下采样解决标记不平衡;

## ■ Task3

特征标准化;

CNN+RNN;



# 总结

---



Southeast University

- 简单的模型
- 用户关系未考虑
- Task1、Task2的分类器可替换

**Thank you !**