

# SMP CUP 2017 用户画像技术评测



团队: ELP

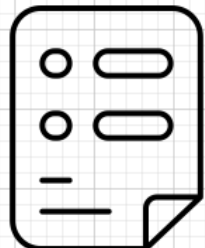
汇报人: 陆俊如

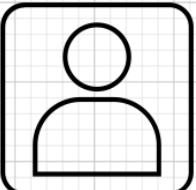
院校: 国际关系学院

指导老师: 李斌阳副教授

# 目录

 任务简介

 任务1：博文关键词提取

 任务2：用户兴趣标注

 任务3：用户成长值预测

 总结与展望

# 任务简介



# 任务简介

数据类型	数据类别	数据内容
静态属性	用户ID	U00296783
	博文ID	D00034623
	博文内容	[转]使用TextRank算法为文本生成关键字和摘要/TextRank算法基于PageRank...
动态行为	发表记录	U00296783 / D00034623 / 20160408 12:35:49
	浏览记录	D09983742 / 20160410 08:30:40
	评论记录	D09983742 / 20160410 08:49:02
	点赞记录	D00234899 / 20160410 09:40:24
	点踩记录	D00098183 / 20160501 15:11:00
	私信记录	U00296783 / U02748273 / 20160501 15:30:36
	收藏记录	D00234899 / 20160410 09:40:44
	关注信息	U00296783 / U02666623 / 20161119 10:30:44
	成长值	0.0367

# 任务I: 博文关键词提取



# 任务1: Supervised-TFIDF(S-TFIDF)

统计性特征:

- TF词频 / IDF逆文档频率: 在Jieba自带词典的基础上构建了基于本数据集的IDF词典, 增大特殊词的权重(原本为1)
- LDA主题词: 根据任务2的标签空间, 将文本分成**42类**, 并为每一类抽取**Top100**作为关键词; 当文本中出现这些词时, 其权重将增加

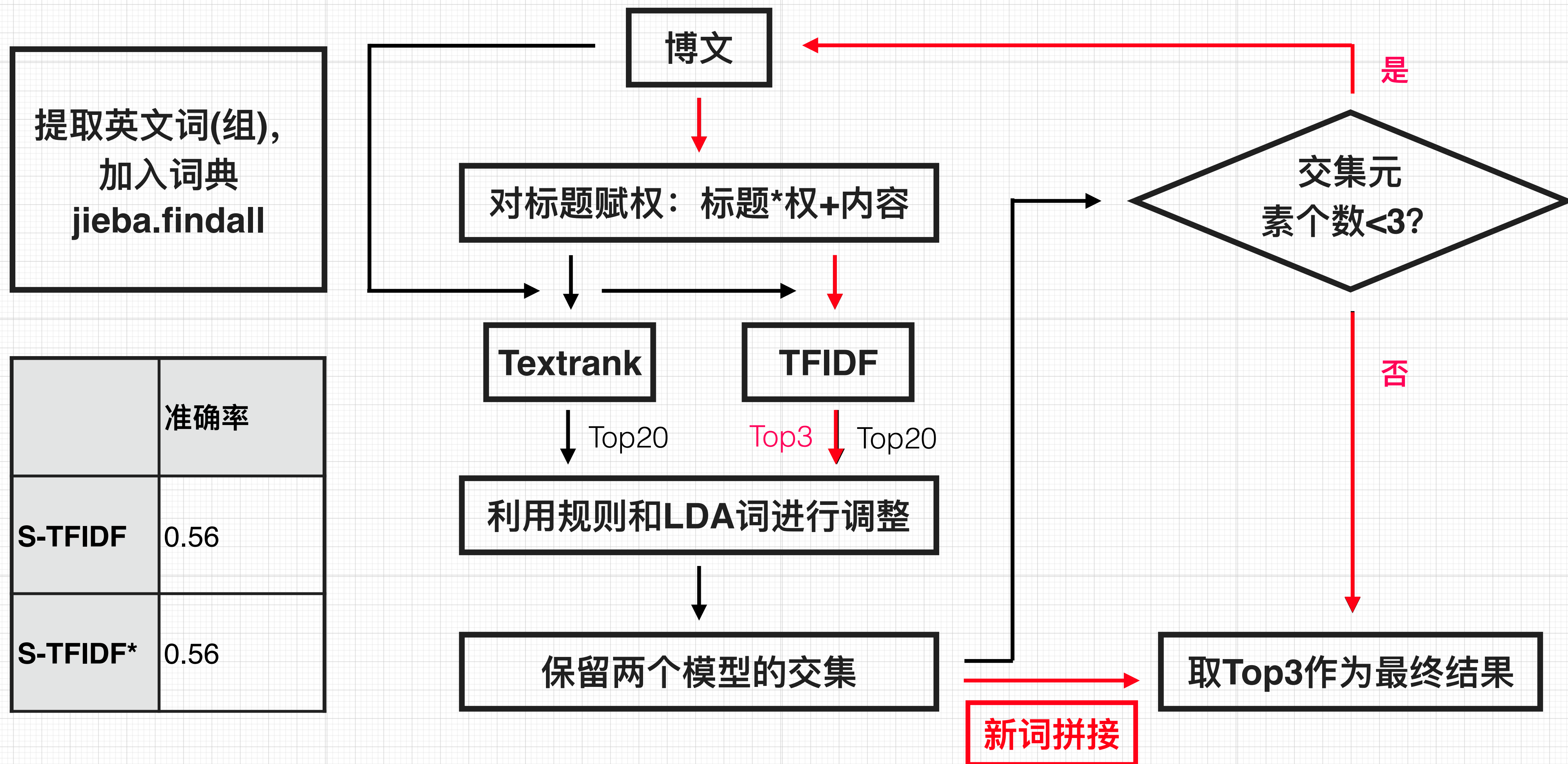
规则:

- 词长: 词的权重乘以**词长**, 提高长词和英语词(组)的权重
- 英语词(组): **英语词(组)**权重增加
- 标题: 根据写作习惯, **标题**相较正文权重增加
- 训练集优先: 以**训练集**构建字典, 字典中词(组)权重增加





# 任务1: Supervised-TFIDF(S-TFIDF)



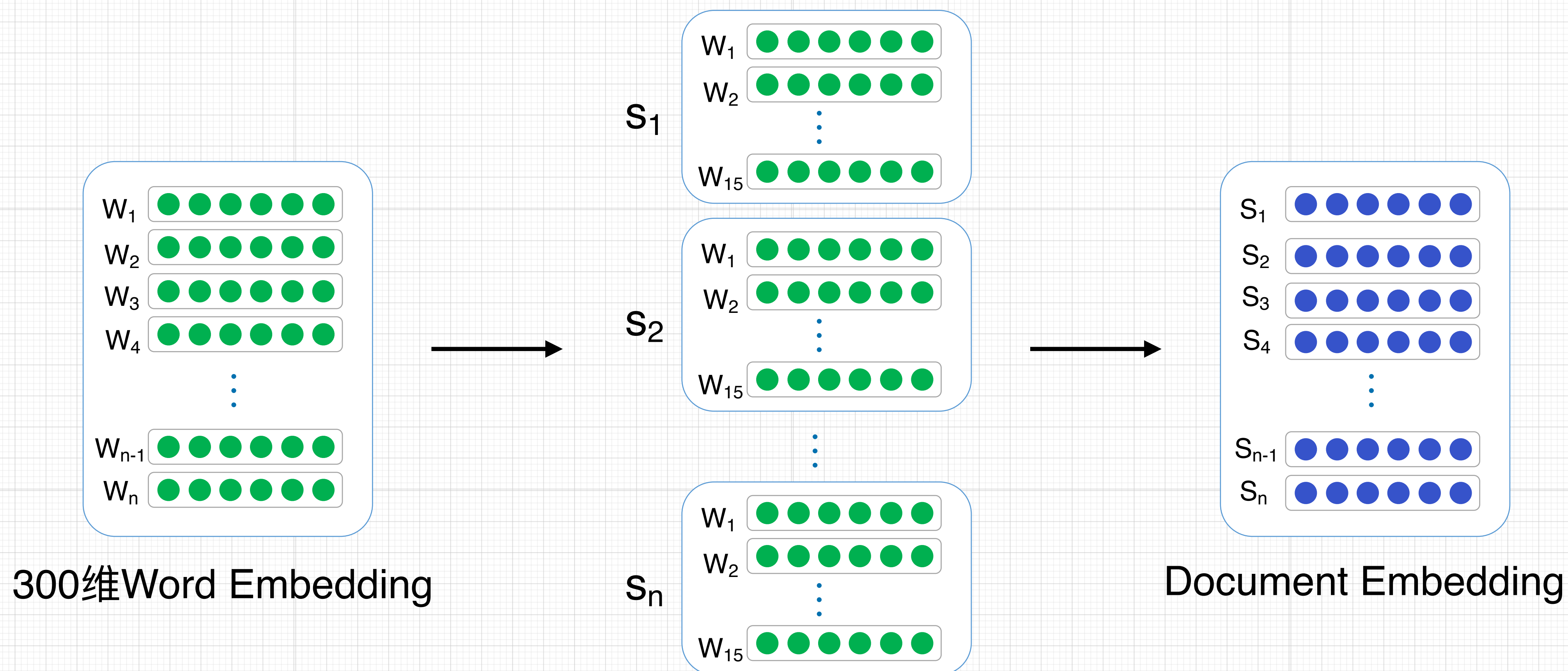
# 任务2: 用户兴趣标注





## 任务2: S-TFIDF/DocumentEmbedding-SVC-Stacking(SDSS)

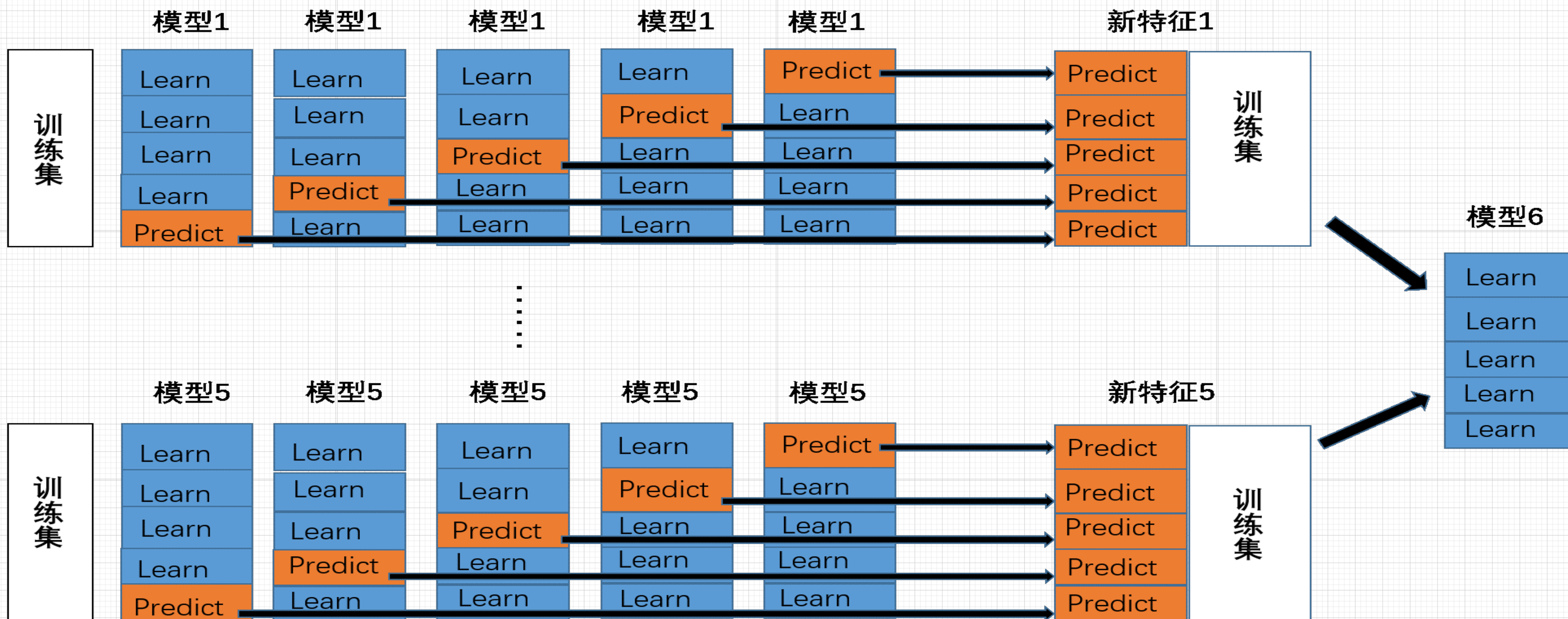
- Word2Vec: 对100W文档分词结果进行分布式表示, 生成300维词向量
- Document Embedding: 每篇文章取TFIDF的**Top15**来表示, 这些词对应W2V**加和平均**生成DE





## 任务2：S-TFIDF/DocumentEmbedding-SVC-Stacking(SDSS)

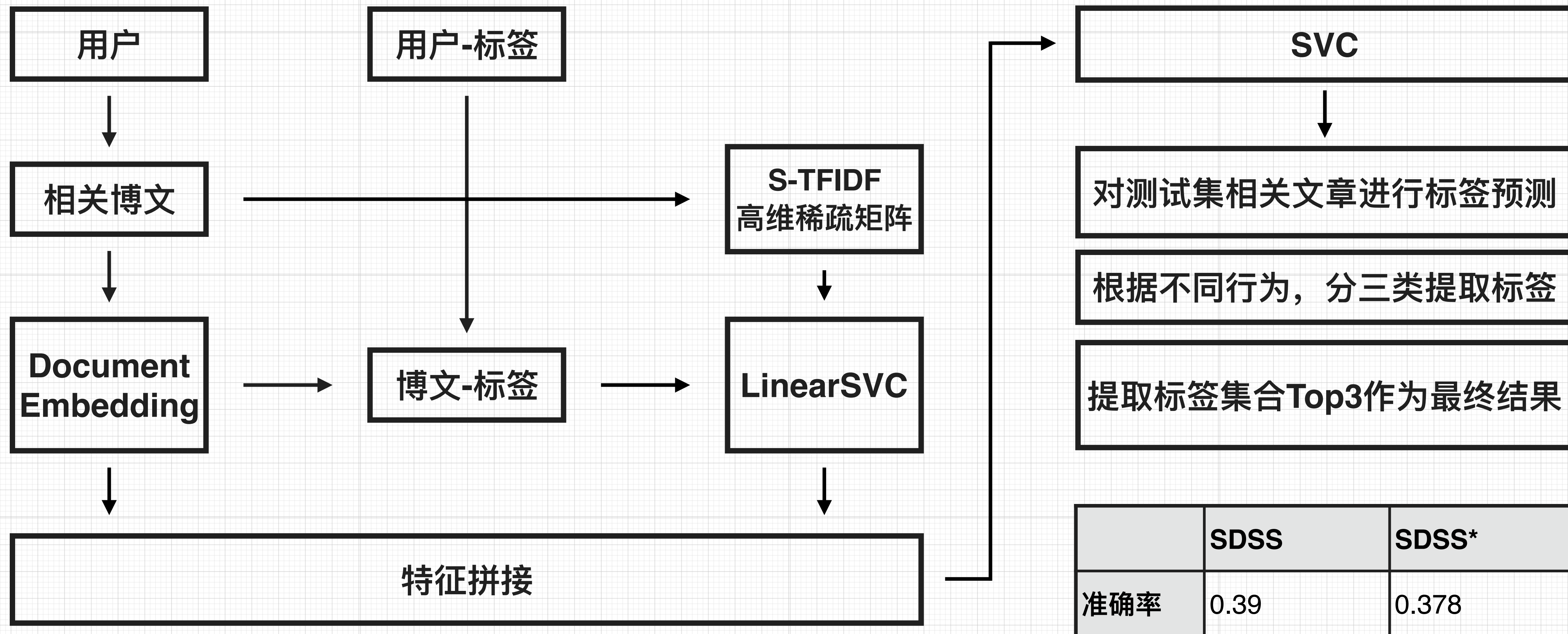
- 支持向量机 SVM
- 集成学习Ensemble Learning：博采众长、集思广益——以Stacking为例







## 任务2: S-TFIDF/DocumentEmbedding-SVC-Stacking(SDSS)



# 任务3: 用户成长值预测



# 任务3： PAR/GDR-NuSVR-Stacking(PGNS)

行为计数： 按月统计用户7种行为， 共84维； **经相关性分析， 剔除负相关的第38列**

U0002438	76	82	136	114	111	143	141	129	149	149	155	139
U0002524	0	0	0	0	0	0	0	0	0	0	0	0
U0002619	0	0	0	0	0	0	0	0	0	0	0	0
U0002837	0	0	0	0	0	0	0	0	0	0	0	0

计数统计量：  
1. 12个月行为均值  
2. log值(+1取log)

行为增长： 统计每个用户7种行为的月间增长率(+1求增长率)，  
增长率可以反映行为的变化情况， 共76维

U0002438	0.0779220779220779	0.650602409638554	-0.160583941605839	-0.026086956521
U0002524	0	0	0	0
U0002619	0	0	0	0
U0002837	0	0	0	0
U0003009	0	0	0	0

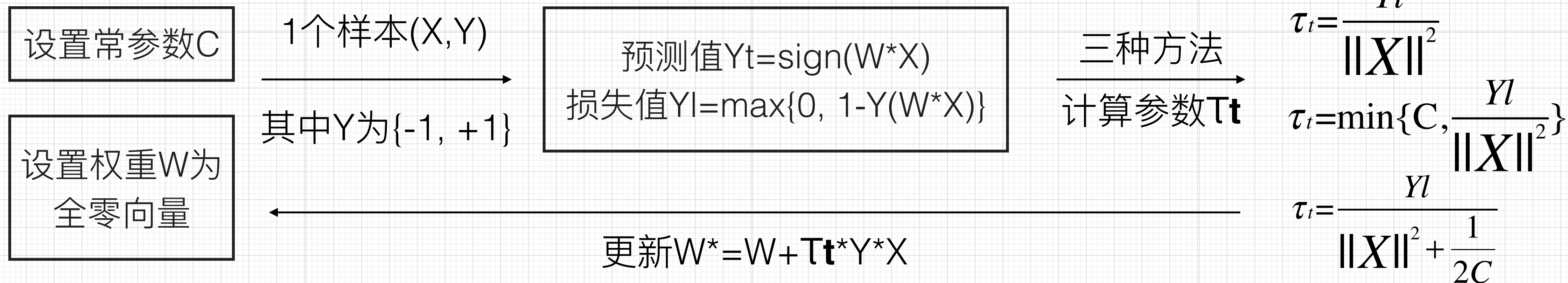




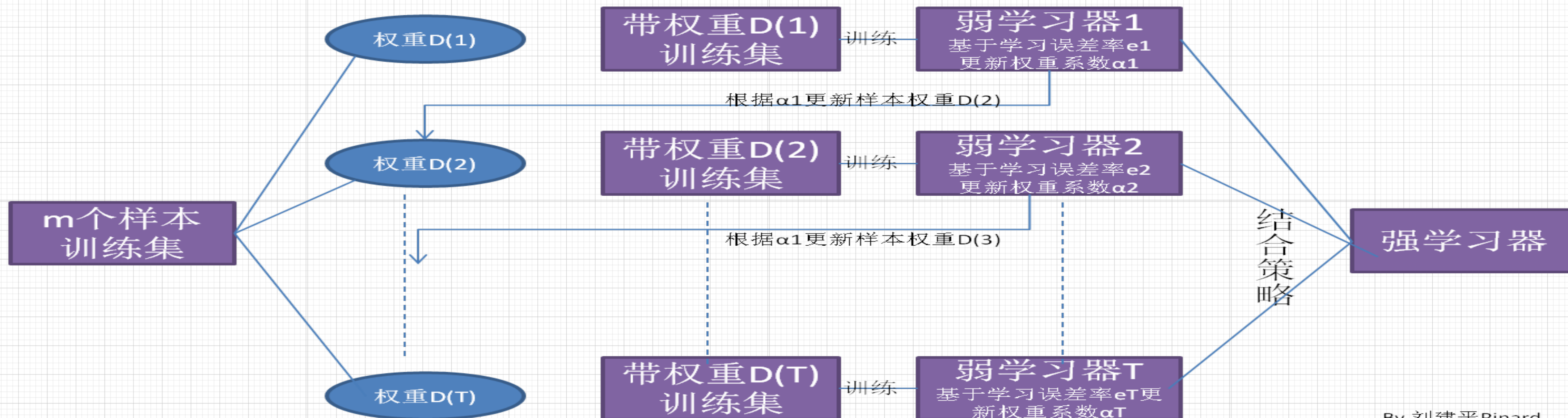
# 任务3: PAR/GDR-NuSVR-Stacking(PGNS)

\*以二分类为例

- PassiveAggressiveRegressor: Online Passive Aggressive Method 逐个样本进行权重更新



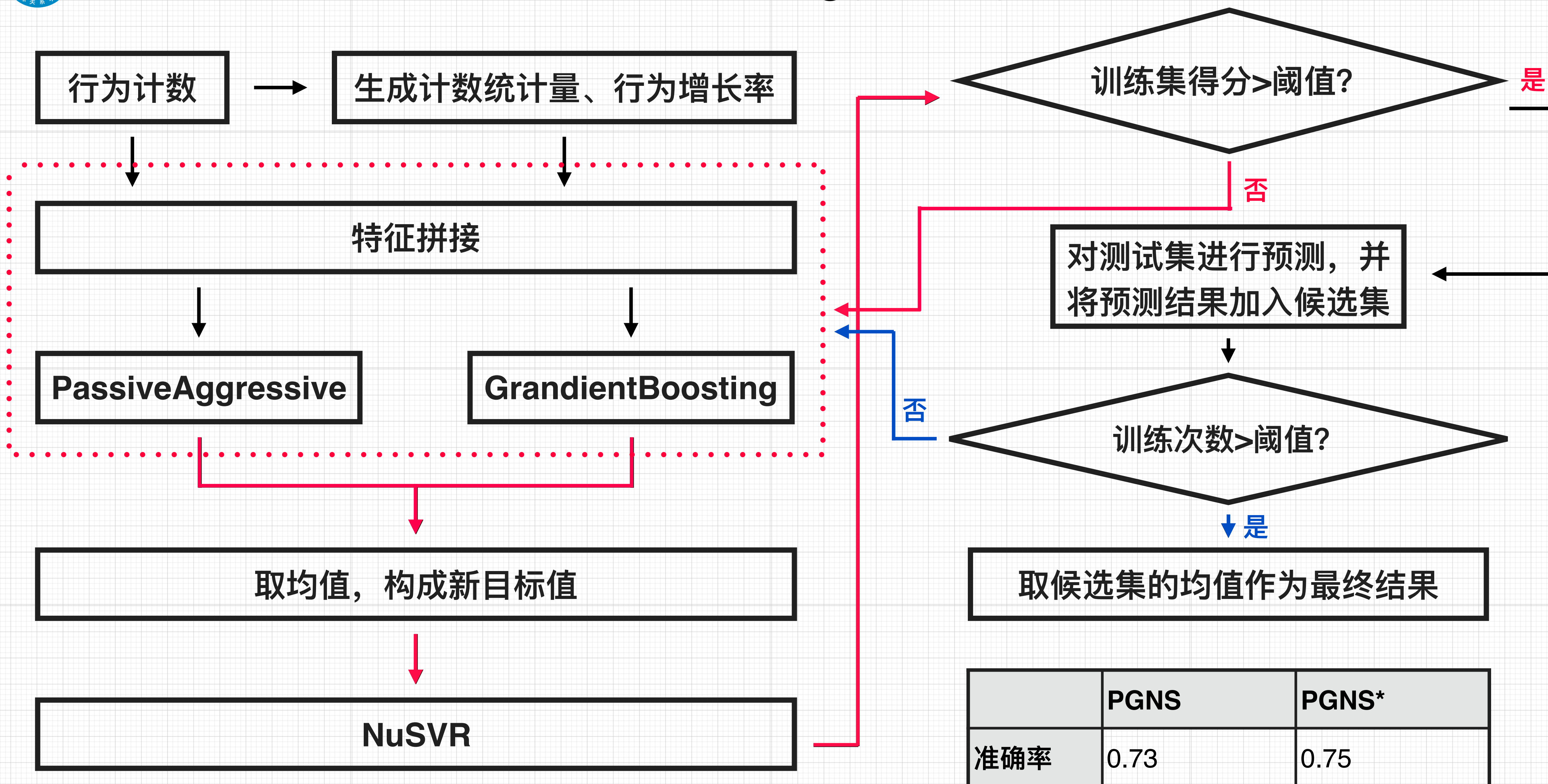
- GrandientBoostingRegressor:







# 任务3：PAR/GDR-NuSVR-Stacking(PGNS)



# 总结与展望

# 总结与展望

	S-TFIDF*	SDSS*	PGNS*	合计
准确率	0.563	0.378	0.751	1.692

- 用户之间的网络关系尚未被应用，可以考虑邻近用户的标签来对用户进行分类
- 使用深度学习和XGBoost等方法进行尝试

# 感谢聆听

开源代码及更多信息：[https://github.com/LuJunru/SMPCUP2017\\_ELP](https://github.com/LuJunru/SMPCUP2017_ELP)