我们组题目是source/target belief and sentiment evolution。组员为徐冰冰，赵越，倪艺函。

这个周主要完成三部分任务，第一是题目理解，第二部分是题目的背景知识调查，第三是赛题现状了解和后期计划。

一 题目理解

目标

源与目标的情感与信任检测，who（源） —-» （情感态度）what（目标）

- 源：实体 两类 Geo-Political Entity (GPE) Organization
- 目标：实体（belief）、事件（belief）、关系（sentiment）

输入

ere.xml 实体，关系，事件，内部文档集合

输出

best.xml

评测

- ere.xml中有两种标注：人工，系统预测结果
- 两种评价方式：所有provenance实例都要找到，找到1个正确provenance（provenance指能表明情感的文本段落指针）
- 四种评测设置 评测文本：大部分英文，包括中文与西班牙文

PST 四元组

【source-entity , target-object , value , provenance-list 】

- value

1 sentiment ： positive，negative

2 belief : CB （确信，包括未来事件），NCB（不确信），ROB（本人态度不明朗）

- provenance-list：文本段落指针（表明情感）
- （s, t）组可能出现多次

1 s态度发生改变，不同PST情感可能冲突

2 s在不同文档中表明态度

输入数据

详见 A. Appendix: Details on RICH-ERE.XML File Format xml格式样例：http://volta.ccls.columbia.edu/~rambow/best-eval-2016/ .

- 训练数据来自论坛（大部分），新闻（少部分）
- 两个mention可能出现在同一object中，共引用coreference没有显示标明
- hopper指event（同一string可表示不同event，event mention 只属于一个event：hopper）

输出数据

详见 B.Appendix: Details on BEST.XML File Format xml格式样例：http://volta.ccls.columbia.edu/~rambow/best-eval-2016/

二 背景知识调查

赛题中已经指明待判断情感的实体，目的是找出体现这些实体的belief和sentiment的位置及所体现的情感。那么不再需要情感分析中的实体抽取步骤，重点在于情感分析。现在的情感分析有三种方法，基于词典的方法，基于传统机器学习的方法和基于深度学习的方法。

1. **基于词典的方法**

基于词典的方法主要通过制定一系列的情感词典和规则，对文本进行段落拆解，句法分析，情感值计算，最后通过情感值作为文本的情感倾向依据。词汇虽然可以表达情感信息，但是单一的词汇缺少对象，缺少关联程度，不同词汇组合的情感程度大不相同，因此表达情感的最基本单位为句子。基于词典的情感分析步骤如下：

- 分析文章段落
- 分解段落中的句子
- 分解句子中的词汇
- 搜索情感词标注和计数
- 搜索情感词前的程度词，根据程度大小，赋予不同的权值
- 搜索情感词前的否定词，赋予反转权值（为负值）
- 计算句子情感得分
- 计算段落情感得分
- 计算文章情感得分

上面的步骤对于积极和消极的情感词分开执行，得到两个分值，分别表示文本的正向情感值和负向情感值。

2. **基于传统机器学习的情感分析**

基于传统机器学习的情感分析将情感分析作为一个分类过程看待，移植到本题可以看成两个分类任务，其一是belief分类任务，分成三类，其二为sentiment分类，分为positive 和negitive两类。对文本内容进行结构化处理，输入到给定算法进行分类训练，对测试数据用模型预测结果。

- 文本结构化指将原始文本转化为分类器可理解的输入，一般用bigram作为最小语义粒度（N-gram,N太小表示的寓意信息太少，太大维度的增长是指数级的），确定最小语义粒度然后用词袋模型或者向量空间模型向量化，将原始数据转化为单词－文档而为矩阵，利用tf-idf或者textrank计算每个词的权值，过滤掉一部分不重要的词。最后进行特征选择，挑选出对目标类别贡献高的特征维度。
- 分类算法常用的是svm，可以选择朴素贝叶斯，crf等。
- 最后进行模型的评价和调参。

3. **基于深度学习的情感分析**

基于词典需要有人工标注的情感词典，基于传统机器学习需要大量标注数据，这些深度学习都不需要。

另外与传统机器学习相比，深度学习在情感分析上还具有很多优势，

• 首先是在特征处理的方式上：普通的机器学习核心在于特征工程，效果的好坏90%取决于特征是否有效，深度学习可以通过数据本身的各种的特征让机器自动去做特征提取。

• 深度学习是神经网络模型，如运用CNN、RNN等模型，能保留词序信息。而这部分信息在浅层模型如LR、SVM、决策树的应用中是需要丢弃的。这部分词序信息，如果在大量训练数据时能够提升效果。

• 深度学习因为Embeddin特征的学习，具有很强的扩展性。神经网络的模型，需要将词进行向量化，利用如W2V\GloVe等离线模型(特征挖掘)向量的接入后，其特征已具备语义信息，训练出的模型具有了很强的扩展性。而浅层学习的输入多为词级别的特征，对于实际训练集非常有限的情形下，效果大受制约。

• 情感分析是语义级别的文本任务，首先深度学习是多层的网络结构，天生具备了学习深层次语义信息的架构，更符合人类思考的方式。

深度学习在文本情感分析上主要是运用输入词的Embedding向量接入，利用如CNN、RNN(LSTM)等神经网络结构，通过SGD算法在训练数据上进行训练。因为模型本身Embedding有包含语义信息和网络结构保留了词序信息，大量数据的训练结果会使得效果上是要好传统的机器学习。

斯坦福大学利用深度学习和语义树进行细粒度情感分析，在情感pos/neg识别上80.7%的准确率。

三 比赛现状和后期安排

这个赛题是2016年的新题目，从2016年2月持续到2017年2月，上一期的论文集还没发布，本次2017年比赛的具体赛程还没出，依照往年，可能是4月出训练数据，10月出测试数据并进行测评，最终到2018年2月论文提交。

我们计划精读部分相关论文，采用情感词典和深度学习结合的方式进行尝试。