

文本相似度匹配

- 数据集任务描述 问题相似度计算，即给定客服里用户描述的两句话，用算法来判断是否表示了相同的语义。示例：

1. “花呗如何还款” - “花呗怎么还款”：同义问句
2. “花呗如何还款” - “我怎么还我的花呗呢”：同义问句
3. “花呗分期后逾期了如何还款” - “花呗分期后逾期了哪里还款”：非同义问句

对于例子a，比较简单的方法就可以判定同义；对于例子b，包含了错别字、同义词、词序变换等问题，两个句子乍一看并不类似，想正确判断比较有挑战；对于例子c，两句话很类似，仅仅有一处细微的差别 “如何”和“哪里”，就导致语义不一致。

- 数据样本形式
 - 行号\t句1\t句2\t标注，举例：1 花呗如何还款 花呗怎么还款 1
 - 行号指当前问题对在训练集中的第几行；
 - 句1和句2分别表示问题句对的两个句子；
 - 标注指当前问题对的同义或不同义标注，同义为1，不同义为0。
- 评估指标
 - 评分以F1-score为准， $F1\text{-score} = 2 * \text{precision rate} * \text{recall rate} / (\text{precision rate} + \text{recall rate})$