1. Vendor VS Us

1.1 话题归类 --- 已处于领先水平，话题覆盖度从行业通行算法的60-70% 上升到了80+%, 准确度增加10%

* 话题归类不够准确，导致情感分析结果有较大的偏差（例子，pegasus）
* 话题归类不够定制化，鞋子中篮球鞋（包裹性，防侧翻），跑步鞋（提速性能，反光功能等），休闲鞋没有足够的区分度和重点差异
* method
  + LDA(Latent Dirichlet Allocation) Topic modeling to discover topics hidden in text using sen2vec features trained by Bert Chinese model;
  + clustering model to group all senences that have similar topics;
  + built CNN(convolutional neural network) models to classify topics for comments

1.2 情感判断 --- 开发中，额外提升了20%+ 准确性

* 所用情感分析词库准确度偏低：例如不捂脚，不硌脚
* 一词多场景/一次多义问题问题没有很好的解决：
  + 例如快递包裹和鞋的包裹性，需要拆开
  + 例如鞋底薄和鞋面薄涉及相似语境相反情感
* 百度api，科大讯飞的词库@rf，说明这些词库已经是行业最好的了，但是对我们的问题依然无法保证足够的准确性和覆盖度 baidu sentiment analysis/boson nlp/snownlp common nlp packages cannot give accurate sentiment judgement on specific dimensions like 不卡脚不硌脚；Hownet sentiment dictionary does not fit oral language good but more for written situation
  1. 断句 --- 开发中，expected提升了10%+ 准确性
* 断句质量有问题（纯靠标点符号对于口语化情况很普遍的评论分析来说效力很差），正在开发，如果开发出来就是行业领先

行业内用的jieba分词（tokenizer）不能很好的处理很多口头语，会分错，jieba是最主流的切词软件

* 评论语境的特点
  + 口语化（潮流词汇更新迭代快）
  + 标点符号使用不规范
  + 繁体字，错别字： 行业研究都在初期阶段
  + 语料库相对于大字典说比较精炼，对于内部孵化来说工作量会比较小
* @rf 方法论跟第一个是类似的，部分产出，然后叠加了额外的关键词作进一步分类和清理
  1. 评论挖掘度 --- Done，与覆盖度指标挂钩
* 由于方法论问题，我们的聚类是可以sensitive to 还未被挖掘的和新的话题
* 人物，场景，identification, 复购率

1. 目前进度

2.1 完善customized 维度

2.2 对于1里面的问题，我们都有了进度，目前准确率（massive comments data）85~90%