2020MCM Problem C

文本评论量化

为了从用户的评论中提取对于商品评价的信息，我们需要用NLP（自然语言处理）来提取评论中于感情相关的词，并从这些词中判断用户的情感倾向，进一步将评论内容进行量化，我们将使用这个量化后的值作为衡量用户满意度的一个衡量标准

朴素贝叶斯法进行情感分析

首先我们想要得到的目标是，可以通过评论中的一个单词来预测这个评论的感情是积极的概率和消极的概率，从而来判断这个评论是积极的还是消极的。
朴素贝叶斯法对单词进行情感分类时：对输入的单词x，通过学习到的模型计算后验概率分布，将后验概率最大的类作为x的类输出。
假设$x \in X$，X是评论单词组成的集合，$c_k \in Y$，$Y = \{Pos, Neg\}$, 代表着评论的积极和消极。$Y = Pos$代表这个评论是积极评论，$Y = Neg$代表这个评论是消极评论。那么我们需要得到它的后验概率：
$P(Y = c_k|X = x)$，根据贝叶斯公式，我们有：

$$P(Y = c_k|X = x) = \frac{P(X = x|Y = c_k)P(Y = c_k)}{\sum_k P(X = x|Y = c_k)P(Y = c_k)}$$

因为
$\sum_k P(X = x|Y = c_k)P(Y = c_k) = P(X = x|Y = Pos)P(Y = Pos) + P(X = x|Y = Neg)P(Y = Neg)$是一个定值，所以我们求解的问题变为以下最优化问题：

$$y = argmax_{c_k} P(Y = c_k)P(X = x|Y = c_k)$$

其中$y \in Y$是x后验概率最大的类别，我们将这个类别作为该单词的类别输出。 我们在kaggle网站上寻找带有标签的亚马逊评论词料库作为训练集训练我们的贝叶斯分类器，然后用来分析评论的情感倾向。

我们的贝叶斯分类器经过训练后，可以根据分类准确率分别得到一个积极评价的词云和消极评价的词云：：

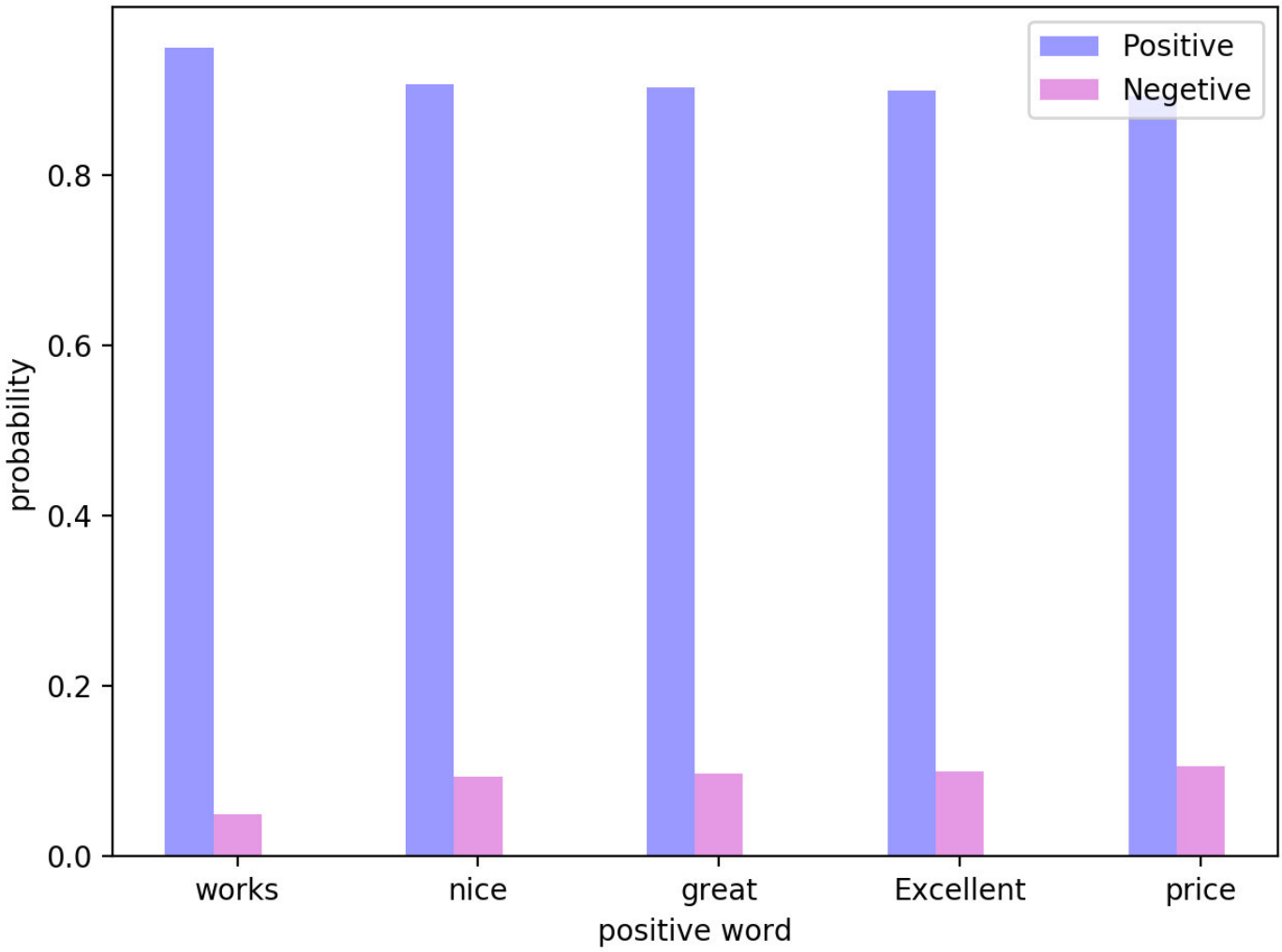
积极评价的词云：


free
without
ears
handas
fine
Excellent
Jabra
works
working
cell
best
years
lot
nice
high
highly
semm
great
price
fits
well
quick

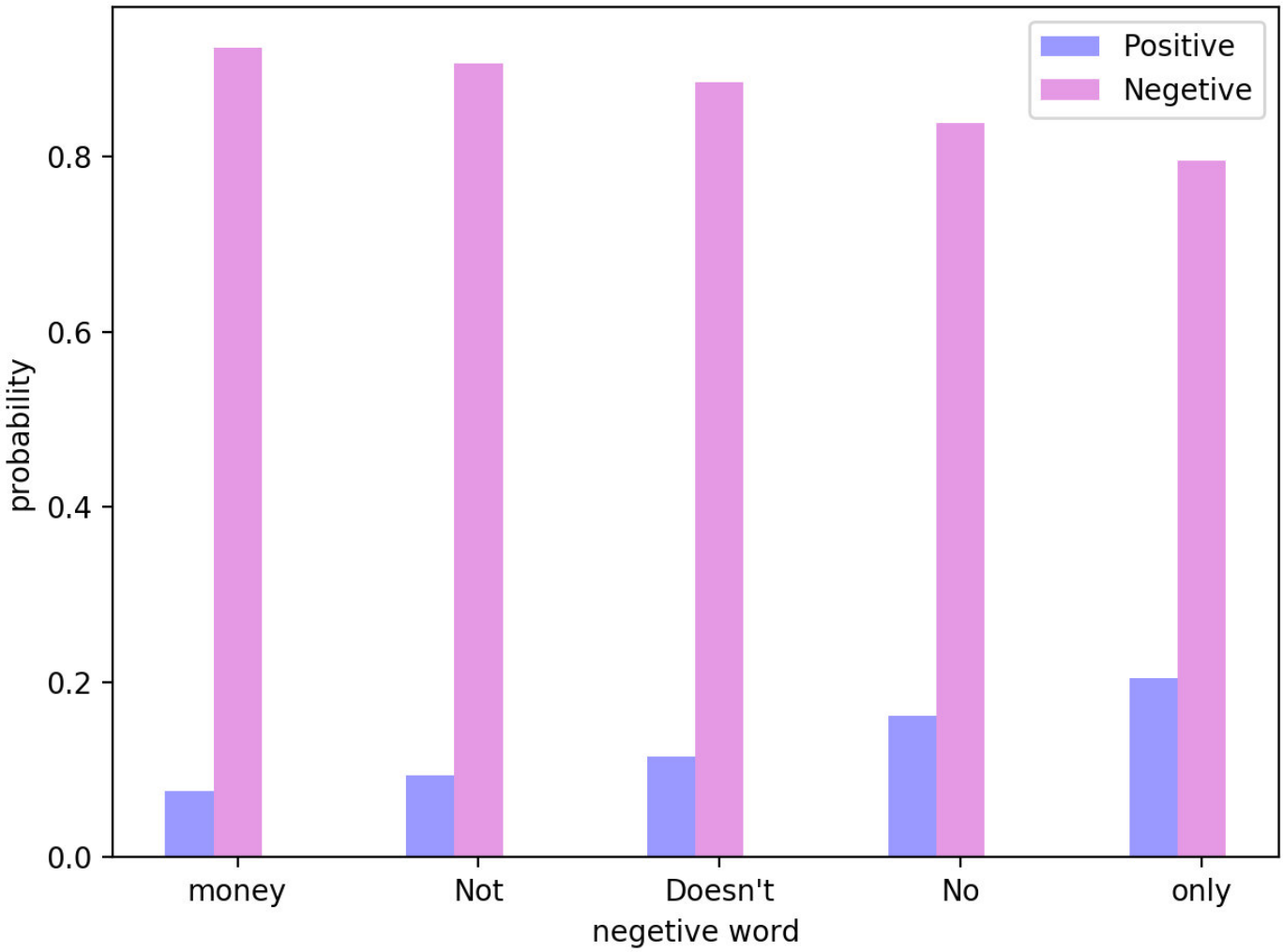
消极评价的词云：

其中单词的体积越大，说明这个单词越具有分类效果，根据以上这些分类效果好的单词，我们可以对其和对应的后验概率进行可视化：

积极单词：


Positive
Negetive
probability
0.0
0.2
0.4
0.6
0.8
works
nice
great
Excellent
price
positive word

消极单词：



基于这些具有显著分类效果的单词，我们可以对评论的感情基调进行量化。

对评论感情进行量化

我们在对评论进行量化的时候，对于评论中的每一个单词，我们都会计算在它存在于评论中的情况下，该评论是积极的和评论是消极的概率。然后我们根据以下公式来赋予评论一个“用户满意度”分数。

假设y_i代表评论的分数，P_{ij}代表着第i个评论中第j个单词的后验概率，C是一个常数，那么量化公式为：

$$y_i = log(\sum_{j}^{n} P_{ij} + C) - log(C)$$

其中

$$P_{ij} = \begin{cases} P(Y = Pos|X = x_j) & P(Y = Pos|X = x_j) \geq 0.5 \\ -P(Y = Neg|X = x_j) & P(Y = Pos|X = x_j) < 0.5 \end{cases}$$

$$C = min_i \sum_{j} P_{ij}$$

我们提出这个量化公式主要基于以下考虑和假设：

- 虽然评论中并不是每一个词都对评论情感判断有着很显著的帮助，但与此同时根据这些词计算出来的后验概率一般会接近0.5，比根据显著效果的词计算得到的后验概率（一般接近0.9或0.1）要小，在整体上这些无关紧要的词不会对整体量化的值影响很大。
- 对于情感为积极的词，我们施加一个“奖励”；对于情感为消极的词，我们施加一个“惩罚”。这便是我们为什么当判断为积极的后验概率大于0.5的时候整体分数加上这个概率，当这个概率小于0.5的时候减去对应的判断为消极的后验概率。
- 为了让这个分数在整体上更加平缓，我们用对数函数来减少极端值对整体带来的影响，也可以减少评论整体上的方差，因为对文本情感进行精确量化是很困难的。为了实现这个操作，我们在对评论中单词的后验概率进行求和之后加上一个常数C，使得所有的求和都大于0，之后再减去这个常数的对数，为了让这个分数在$x = 0$上下波动，如果分数>0，代表着该评论倾向于积极，如果分数<0，代表着该评论倾向于消极。

信息熵

信息熵是表示随机变量不确定性的度量，设X是一个取有限个值的离散随机变量，其概率分布为：

$$P(X = x_i) = p_i, i = 1, 2, \ldots, n$$

则随机变量X的信息熵定义为：

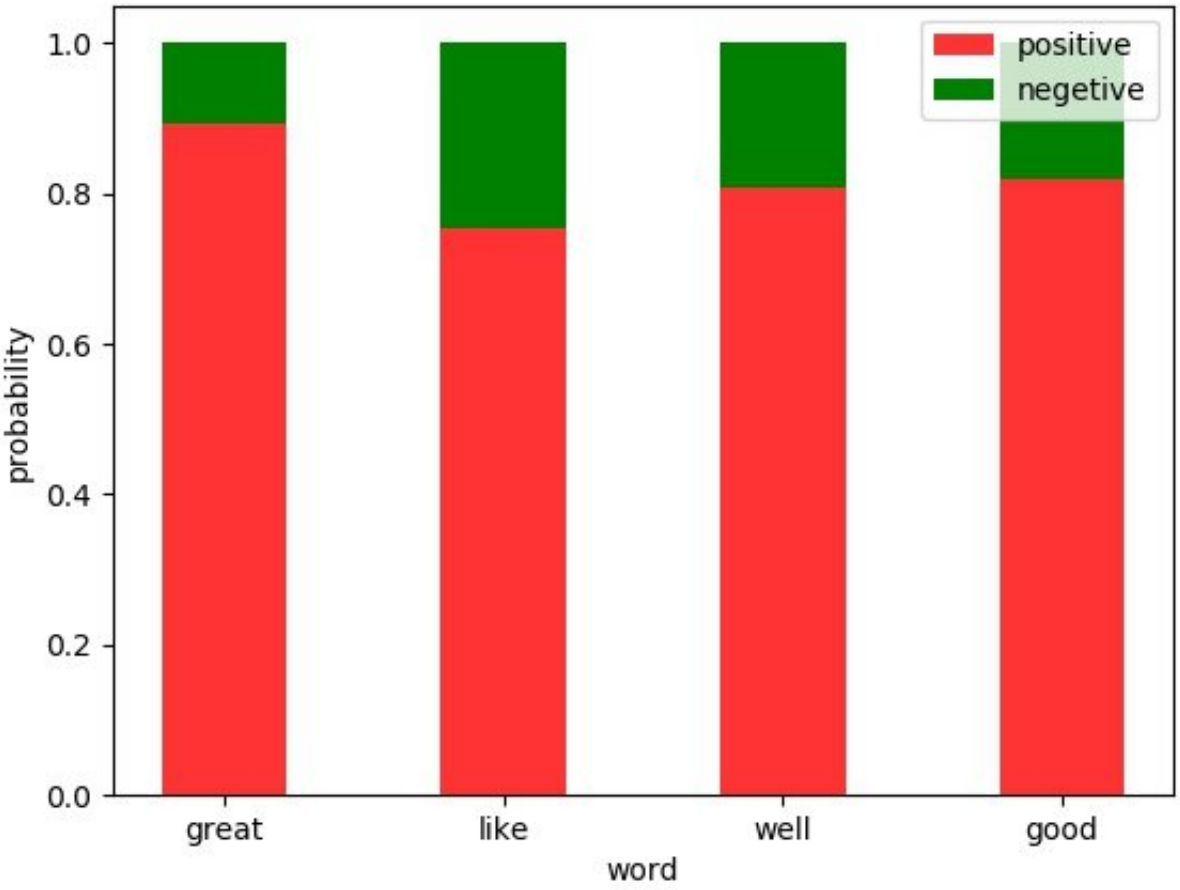
$$H(X) = -\sum_{i=1}^{n} P_i log P_i$$

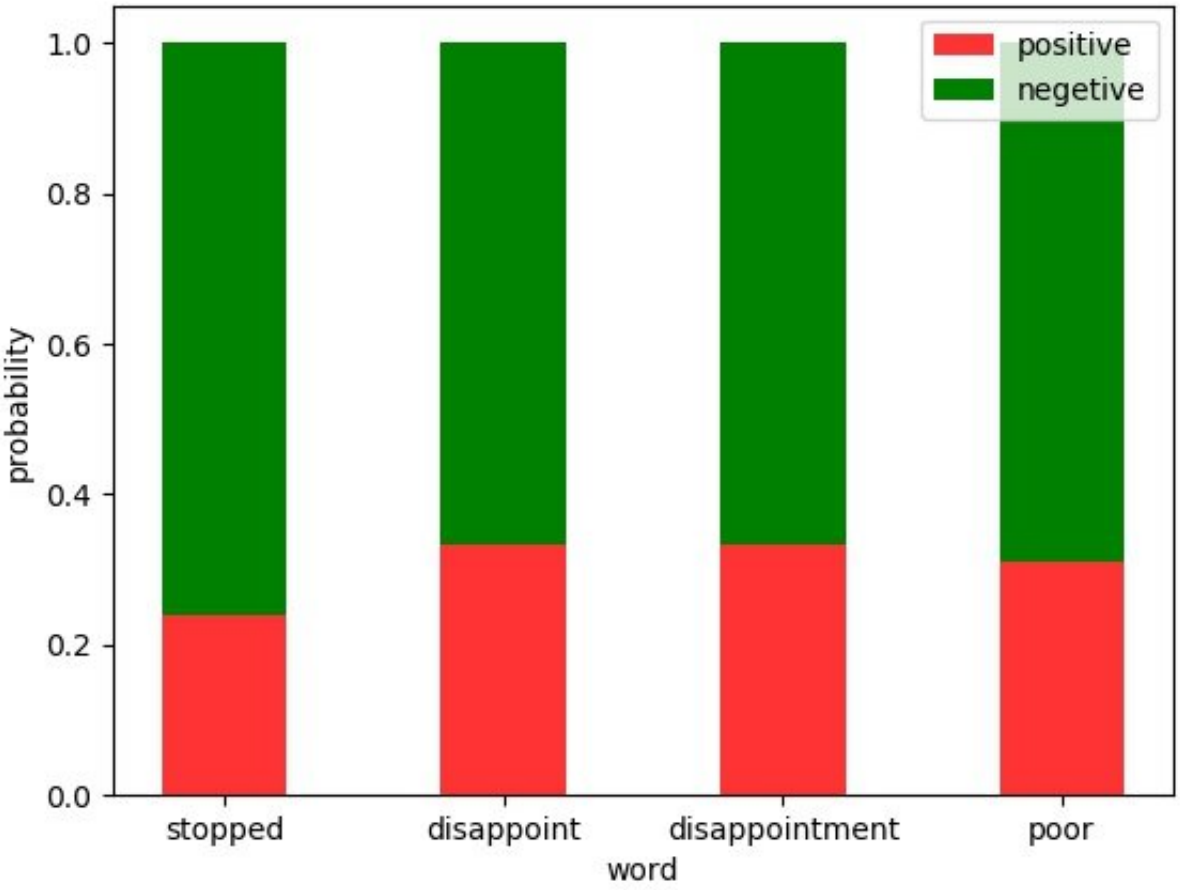
信息熵的值越大，随机变量的不确定性越大。而一个随机变量所包含的信息量与其不确定性有着直接的关系。比如说我们要搞清楚意见非常不明确的事，或者我们一无所知的事情，就需要了解大量的信息。所以从这个角度来看，可以认为信息量等于不确定性的多少。因为我们可以用信息熵来衡量一个随机变量所包含的信息。

特定质量描述与评级水平的关系

我们对于一些带有感情色彩的词（比如great, like, well, good, awesome），统计了当评论中出现这些单词的时候评级为优（4-5星）和劣（1-3星）的概率，我们发现，对于一些带有强烈感情色彩的词，评级水平往往与这些单词有着紧密的联系。

我们对于几个带有积极感情色彩的词和消极感情色彩的词，统计了评论中出现这些词其评分质量为高和低的概率：


positive
negetive
1.0
0.8
0.6
0.4
0.2
0.0
probability
great
like
well
good
word



从图中我们可以看到，这些单词对于评分的区分度比较大，如果在评论中出现这些单词，那么相对应的评分往往会与其积极性或消极性挂钩。但并不是所有的词语都具有类似的区分度，比如一些与情感不相关的单词（price），或者一些中性词语（much），这些单词对于评分的区分度并不高，因此我们认为这些单词与评级水平并没有直接的关系。