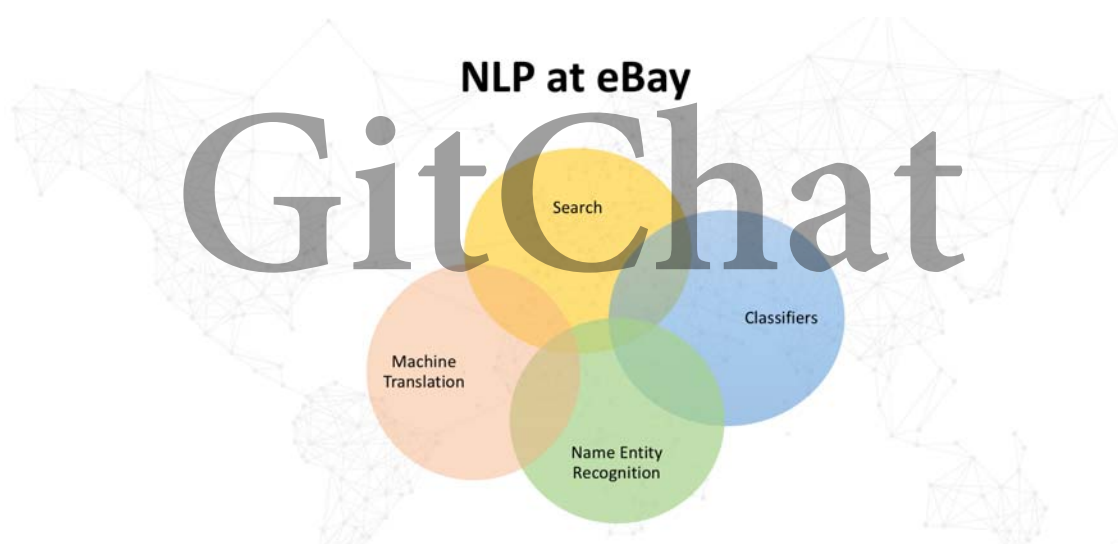


# 自然语言处理在电商的技术实践

这几年，电子商务发展得非常迅猛，无论是在国内还是国外，都大有赶超实体店的势头。作为一个在美国主要的电商网站之一的eBay混了n年的“数据科学家”，我给大家分享一下**自然语言处理**（**Natural Language Processing**，缩写为**NLP**）在eBay的技术实践。

## eBay用到自然语言处理的领域

在电商这个领域，我们处理的对象不外乎是卖家提供的商品和买家的需求。商品是由文字描述和图片构成的，而需求则通常是用文字表达的（搜索关键字）。eBay每天都有数亿的新商品上架和数亿次的搜索，产生的文字数据是海量的。因此，eBay对自然语言处理（NLP）的需求是不言而喻的。那么NLP都影响到了电商的哪些领域呢？



作为电商企业，**搜索**功能是其重中之重。这是买家进行购买的最便捷有效的途径。因此我们的搜索引擎也是公司最重要的产品（没有之一）。文档索引的祖师爷算法称为**TF-IDF**。这是NLP中一种用于信息检索与文本挖掘的常用加权技术。它是一种统计方法，用以描述一个词对于一个文件集的重要程度和对文件集中某篇文章的区分度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中不同文件中出现的频率成反比下降。传统的网页搜索把它作为网页与用户查询之间相关程度的度量或评级，从而推荐相关的网页和文章。那么在电商的应用场景中，我们会把它作为买家搜索关键字和商品相关程度的度量，从而推荐相关的商品。虽然，我们的搜索引擎针对电商的特殊应用已经进行了各种改进，但追根结底，核心还是TF-IDF这一NLP的重要算法。

NLP在电商中的另一个重大应用是**机器翻译**。eBay在全世界30多个国家都有网站，大多数的网站都支持跨境电商，也就是说我在美国卖，在俄国的买家也能买；或者我也可以从美国买俄国卖家的商品。但我们希望能够让他们用俄语进行搜索，并且看到我们的商

品用俄语描述。美国的网站每天有上亿个新商品上架，如果没有机器翻译，这是完全不可能实现的。

搜索引擎和机器翻译背后有各式各样的其它的NLP技术做支撑，比如**命名实体识别技术（Name Entity Recognition，缩写为NER）**和各式**文字分类器（Text Classifier）**。

## 命名实体识别（NER）

在电商中，命名实体识别有着非常特殊的地位。从定义上讲，它是从自然语句中提取属于某些特定类别的词汇和短语并加以分类，这是**把无结构文字（unstructured text）转变为有结构文字（structured text）**的一个重要步骤。

下面这张图就是一个电商中NER的例子：



应用：

- 提取非结构化商品属性 → 帮助自动翻译，搜索，自动补充属性辞典

这是一个简单的商品标题“**New Apple iPhone 6s 16GB AT&T Locked Space Gray Smartphone**”。NER会干什么呢？它要尽可能地把这个标题上每一个单词都分成某一个类别（或者打上一个标签）：New是一种商品的**新旧状态（condition）**；Apple是一个**品牌（brand）**，iPhone 6s是一个**型号（model）**，16GB为一个**内存容量（storage capacity）**，AT&T unlocked为**手机供应商（carrier）**，Space Gray是**颜色（color）**，Smartphone是**产品类别（type）**。

目标明确了，那我们来谈一谈怎么做？NER本质上是一个分类问题，是一个从单词到有限类别的分类问题。分类问题首先要提取单词的**特征矢量（feature vectors）**。传统的NER通常会提取以下的特征：这个词有几个字母，里面含有一些什么样的字母组合，它的前后是否有某些特定的介词，它是否出现在句子的开头或者结尾，等等。而现在新的NER会用开源词向量（比如Google的word2vec），或者用专业语料训练的专业词向量（比如用大量eBay自己的商品标题自行训练的）作为特征矢量。定义好这些特征矢量之后，就可以用训练样本训练一个分类器。那么分类器可以是简单的模型，比如逻辑回归甚至线性回归，也可以是复杂的模型，比如现在很火的神经网络。

NER的最大应用就是帮助机器翻译。当NER对上面这个标题进行了理解之后，会根据一些规则进行翻译，比如品牌“Apple”和型号“iPhone 6s”就不需要翻译，而“Space Gray”这个


描述颜色的短语就会被翻译。

NER还能够提高搜索准确度。举个例子，在英国的网站，有人输入搜索关键字“Orange iPhone”，但其实他并不是要搜索一个橘色的iPhone，而是要搜索一个被无线供应商Orange所绑定的iPhone。那么NER就能在这个场合帮助我们正确理解每个单词，从而能够把Orange作为无线供应商来进行匹配，而不是匹配橙色的手机。

为了更好地描述每类产品，我们每个类别下都会有一些辞典，它们定义了这个类别的关键属性。比如手机类别里就有品牌（苹果，三星，华为等），型号（iPhone7，S8，P10等），颜色（白，黑，玫瑰金等），内存容量（16GB，32GB等）等。这些词典过去都是人手工编辑的，这是一个效率很低，而且容易出错的方式。那么怎么样能够高效准确地增加词典里面的内容？这就是NER可以施展拳脚的地方。打个比方，iPhone7去年出了一款红色的。红色原本是不在手机类别的颜色词典中的，而NER却从很多新商品标题中提取了这个颜色。加入它到颜色之后，我们能够给卖红色iPhone的卖家提供红色的选项，从而加快他们打广告的流程。同时我们也能够更加方便买家对产品进行细分，找到自己心仪红色手机。

## 分类器

正如前面提到的，搜索引擎和机器翻译是被各式分类器支撑的。下面这张图里介绍了一些电商中用到的典型分类器：

- 
- Classify items to categories – category recommendation
  - Classify items to product/accessory/bundle
  - Classify items to product types
  - Classify reviews to real product reviews/others
  - ...

**第一种称为产品类别推荐。** eBay在美国的网站上有上万六千个产品门类，有卖衣服，玩具，日用品的，也有卖电子产品，工具，建材的。它们就像超市里不同的走廊。每个走廊里面还有细分，比如到了手机的走廊，还分别有手机，充电器，保护膜的货架。这样的细分能帮助网站更好地管理商品，帮买家进行查找。这些门类分的非常细，对于一个新的卖家而言，他不知道放在哪一个门类下面，如果他放错了，他的东西就被买家找不到，进而卖不出去。我们的工作就是帮助这些卖家，帮他们把商品正确的放在正确的门类下面。这个称为类别推荐，也就是一个从商品标题到产品类别（上万类）的分类器。

**第二种是产品 / 附件 / 套装分类器。** 为什么会有这种需求呢？虽然我们给卖家推荐了一个门类，可是有些卖家还是会故意把商品放在错误的门类下。听上去有点奇怪，为什么呢？eBay（我听说淘宝也有类似的问题）会有一种常见的问题：你在手机的门类下面你去搜索iPhone，然后把搜到的商品按价格从低往高排序（大家都想买便宜货嘛），结果

发现排在前面的全是手机壳之类的附件。为了防止这种情况的发生，我们的任务就是用算法区分商品和附件，从而不让附件出现在手机的搜索结果里。这就引出了对产品 / 附件 / 套装的分类器的需求。从机器学习角度看，这就是一个从商品到3个类别的分类器。

除了对商品进行以上两种维度上的划分，有时还有需要在一些其他维度上的划分。这就引出了今天提到的**第三种分类器：商品类型分类器**。什么是商品类型呢？比如：翻页笔，麦克风，电脑，口红，等灯。那为什么要做这样的事情呢？假设你刚买了一个手机，那我是不是应该推荐你买个壳，买个保护膜呢？这就是个性化推荐。它就是通过把搜索关键字和商品分成不同的产品类型做到的。它还有一个应用：区分歧义的搜索关键字。比如，买家输入了搜索关键字“Lord of Rings”，那他到底是想买书还是想买电影光盘呢？这时我们就可以根据买家的购物历史判断他的兴趣，同时结合和搜索匹配的商的产品类型，选择和他兴趣相关的商品。比如这个买家过去总是买了很多电影光盘，那看来这次他也很有可能是想找“Lord of Ring”的电影，而不是书。这就是一个典型的分类器提升搜索引擎的例子。从机器学习的角度看，它是一个从商品到商品类型（几千类）的分类器。

今天要讲的**第四种分类器称为评价分类器**。这里的评价指的是产品评价。我们的网站上面有几百万，上千万的产品。

首先，我需要定义一下**什么是产品**。它和之前多次提到的商品是不同的概念。“iPhone 7 Red 32GB unlocked”对应了一个条码，也就是一件产品，而eBay上可能有10万个对应这个产品的商品。

我们希望能够提供一些用户的评价来指导买家对产品的选择。比如，对于我这个从来没买过华为手机的人来说，是买华为的P9还是P10呢？我最好的办法就是看产品评价。买家们的确在上面写了很多评价，可有些并不是真正对产品的评价。有些是针对物流的，比如说卖家出货太慢，或者包装摔坏了。有些是针对卖家的，比如说这个卖家包装得不好，或者这个卖家反馈不及时。这些都不是真正的产品评价。还有人甚至在评价里说脏话，这是违反规定的，需要被删除。从数据科学的角度看，这也是一个分类问题，是一个从评价到真正的产品评价 / 非产品评价 / 脏话的分类器。

大家可以想象类似的分类问题还有很多，这里我就不一一赘述。这时候，可能有些人要问了，这些分类器听起来很重要，可是它们和NLP有什么关系呢？你是不是跑题啦？今天难道不是讲“NLP在电商的实践”吗？答案是：没跑题，这些分类器的背后都运用了NLP的算法，**比如Bag of Words，统计语言模型（Statistic Language Model，SLM），词向量等**。下面我会就两个分类器深入地讲一下。

## 商品类别推荐详解

# Example: Item Category Recommendation

## Problem

- Classify item to leaf category (>16k categories)
- If doing randomly, accuracy < 0.00625%

## Progress over the past 10 years

- 2006 ~ 2011: Histogram + Naive Bayesian ~ **50%** top1 accuracy
- 2011 ~ 2015H1: KNN based on search engine () : ~**73%** top1 accuracy
- 2015H2: KNN + SLM : ~**81%** top1 accuracy
- 2017: DNN: ~**90%** top 1 accuracy

eBay美国网站有一万六千多个类别，如果说随机推荐的话，准确度是很低很低的（<0.00625%）。在过去10年，eBay一直在这个问题上不断的改进。

那最早的称为**Histogram + Naïve Bayesian**的版本是什么原理呢？我们虽然有一万六千个类别，可是这些类别的分布是非常不均匀的。可能有10%的商品都是手机，但却只有0.01%是邮票。这个算法就是利用了这个不均匀性。当我们根据实际商品分类画出类别柱状图（histogram）之后，每一个类别有会对应一个概率，于是，我们根据这个概率分布随机地把一个新的商品分到这些门类下去。比如，随机给10%的商品推荐手机类别，随机给0.01%的商品推荐邮票类别。这个算法听上去too simple too naive对不对？可它却上有差不多50%的准确度。

后来大家觉得这个办法实在是太土了，就做了一大步改进，用**K近邻法（K Nearest Neighbor，缩写为KNN）**。这也是业界很常用的一个算法。它的原理是什么呢？当你输入卖的产品的标题之后，它就被变成一个搜索关键词，然后利用搜索引擎找到跟你这个产品类似的标题。比如搜索引擎找到了最相似的一百个商品，然后看这些产品分属于哪些类别，占最大商品比例的类别就是推荐结果。这是一个很简单却非常行之有效的办法，它把分类准确度从50%一下提高到了73%。

这个版本的算法用了好多年，直到2015年的时候我们在这个基础上做了一个简单的改进。刚才讲到用KNN找到了一百个相似商品。假设它们分属于10个类别，那么面对这10个类别，我们不再简单地根据哪个类别的商品多就用哪个类别，而是增加了一层逻辑：用原本属于这些类别的大量商品训练出每类的统计语言模型，然后用输入关键字跟这些模型计算相似度，重新对这10个类别重新进行排序，推荐排名最高的类别。这一步简单的做法，就使得准确度从73%提高到了81%。统计语言模型是NLP的一个重要的技术，它能够简单有效地描述词语的上下文关系，时常被用在文字分类的应用中。一会儿我会讲统计语言模型的具体细节。

去年，在神经网络兴起之后，我们也想，要不要试一试神经网络呢？于是，我们做了一个卷积神经网络的模型，把准确度从81%提高到了90%。卷积这项技术最早是应用于图象的，是用来抽取图象特征的（均值，边缘等）。可同样的思路后来也可以被应用于文字上：Yoon Kim 2014年在EMNLP上发表了一片很好的一篇文章，具体算法大家可以参考这篇文章（见下图）。



# Convolutional Neural Networks for Sentence Classification

Yoon Kim

New York University  
yhk255@nyu.edu

## Abstract

We report on a series of experiments with convolutional neural networks (CNN) trained on top of pre-trained word vectors for sentence-level classification tasks. We show that a simple CNN with little hyperparameter tuning and static vectors achieves excellent results on multi-

local features (LeCun et al., 1998). Originally invented for computer vision, CNN models have subsequently been shown to be effective for NLP and have achieved excellent results in semantic parsing (Yih et al., 2014), search query retrieval (Shen et al., 2014), sentence modeling (Kalchbrenner et al., 2014), and other traditional NLP tasks (Collobert et al., 2011).

当我们回头看这四个版本的算法时，如果我们只关心准确度的话，大家一定会认为神经网络版的最好。可是这样大的准确度提升是有代价的。第一，它需要极大的训练样本，一万六千个类别每类都需要几千个商品作为训练样本。大量的训练样本的采集是很昂贵的。第二，神经网络的训练时间是很长的，经常需要好几个礼拜。第三，我们的类别结构每年要改好几次，这个算法没有办法推荐新的类别。而在这一点上，KNN有天然的优势，它完全能够适应每一次的类别结构修改，因为它是完全基于搜索引擎的。

**我做模型的理念是：能用简单的就不要用复杂的。**尤其在工业界里面越复杂系统越容易出错，除非你的性能有非常大的提升，值得付出额外的时间和维护成本。

## 产品 / 附件 / 套装分类器详解

我刚才提到的第二种分类器是如何把这个产品分辨出它是产品本身，附件，还是套装。从下面的图中，大家就能够清楚地这个为什么要做这个分类器。

### Example: Product/Accessory/Bundle Classifier

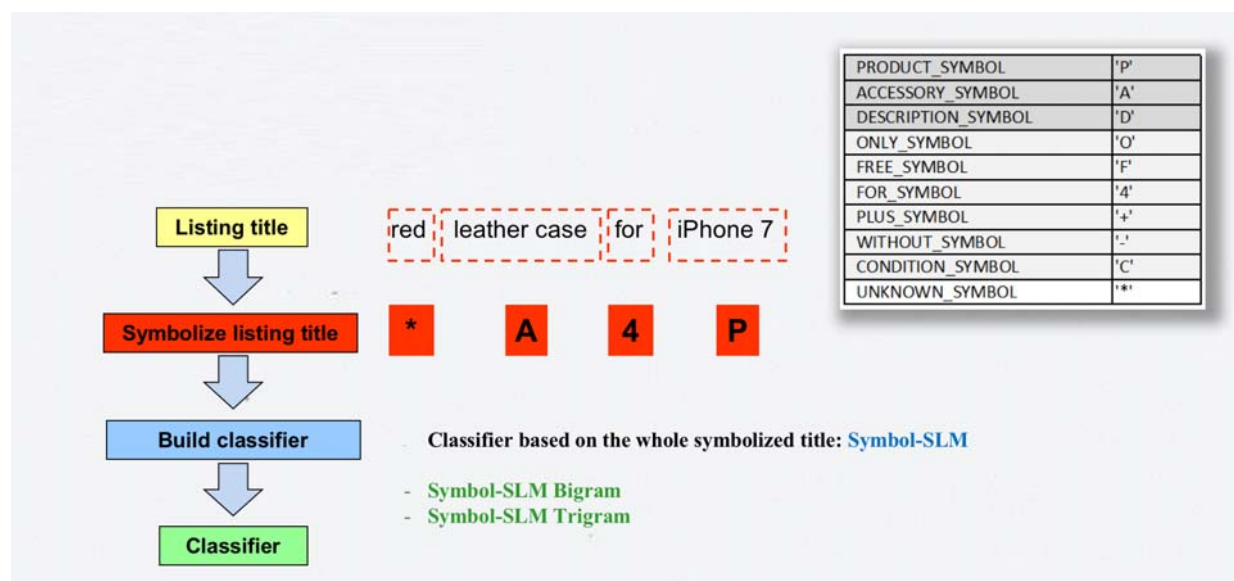
The image shows a screenshot of an Amazon product listing for a camera bundle. The listing is divided into three sections, each with a product image, title, price, and shipping information. Annotations are used to identify the components of the bundle:

- Bundle:** The top section shows the "C Olympus OM-4 E-M5 Black Body Camera with 14-42mm Black Lens (M.Zuiko ED M. 155 15h left 11/10, 10AM) \$1,099.99 Buy It Now Free shipping". An orange starburst labeled "Bundle" points to this section.
- Accessory:** The middle section shows the "C Olympus M.Zuiko 14-42mm Micro Lens E-P3 E-PL2 E-PL3 E-PM1 OM-4 camera lens 28d 15h left 11/23, 10AM \$114.95 Buy It Now Free shipping". An orange starburst labeled "Accessory" points to this section.
- Product:** The bottom section shows the "C Olympus OM-4 E-M5 16MP Live MOS Interchangeable Lens Digital Camera Black Body 7d 14h left 11/2, 10AM \$999.99 Buy It Now Free shipping". An orange starburst labeled "Product" points to this section.

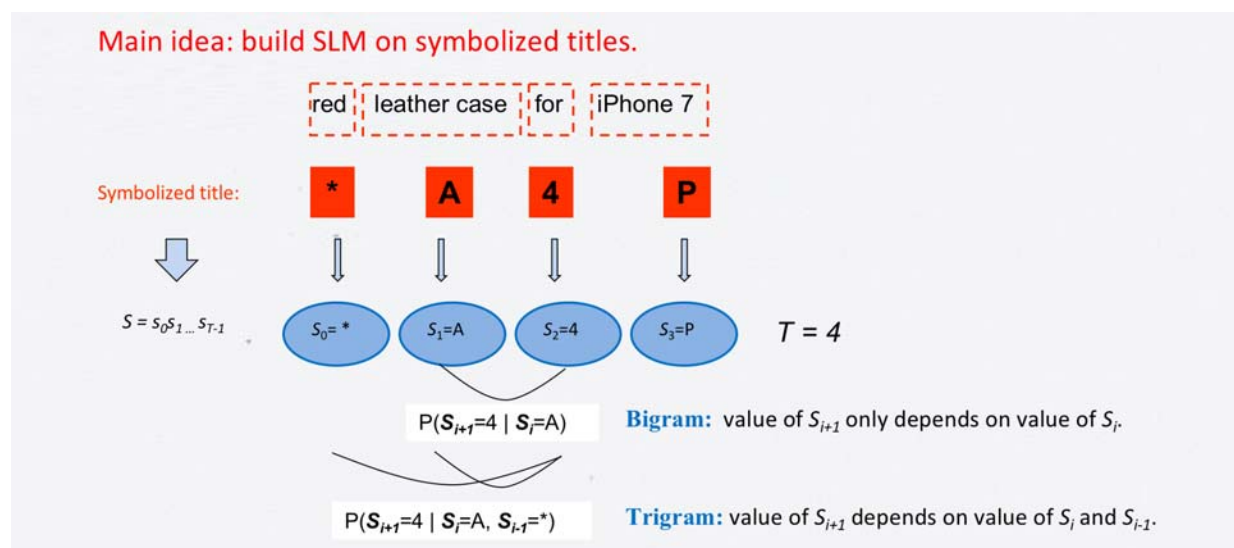
这里三个商品。如果我们看用长方形标出的文字，我们看到这三个商品都有共同的关键词：奥林巴斯，E-M5（型号），lens（镜头）。可是它们的价格却差很多，有1099块

的，有100多块的，还有999块的。为什么呢？因为第一个是套装：机身加镜头，第二个是镜头本身，第三个仅仅是机身。不难想象，完全不同的东西，价格是一定会有差别的。

那我要怎么去用我们的分类器识别出他们分别是商品本身，套装，还是附件呢？现在我来给大家具体讲一下细节。



面对一个标题，我首先把它符号化。比如这个标题叫做“Red leather case for iPhone 7”。符号化是参照一个人工定义的词典进行的。比如Leather case被符号化为A；for是一个特殊连接词，被符号化为数字4；iPhone 7是一个型号，算是一个产品的描述，被符号化为P。而red不在这个词典里面，我们就把它符号化成\*（星号）。当我们把词语都变成简单的符号后，我们就可以构造一个基于这些符号统计语言模型的分器（Symbol-SLM）。这个模型可以是二阶（bigram）的，也可以是三阶(trigram)的。下面这张图解释了什么是二阶，什么是三阶。



这是一个统计语言模型在文字分类中的典型应用。这时有些同学会问：那为什么不用单词本身建模呢？干嘛要多此一举？原因是这样的。不同类别（手机，电脑，照相机，等）都需要这样的模型，它们每个类别的训练样本都不多，训练不出很好的语言模型。而符号化之后，它们的语言结构就变得类似，这样就可以把不同类别的训练样本合在一起训练一个统一的模型，既简单又有效。

# 总结

刚才我讲到了命名实体识别，各种基于NLP的分类器在eBay的实践，希望能起到一个抛砖引玉的作用。

我相信这些方法不仅可以用在电商领域，也应该可以拓展到其他的领域，比如说在电子支付，银行信贷，保险这些行业。比如反欺诈其实就是一个简单的分类问题，比如某人的用户的行为是正常还是不正常行为。另外还有机器翻译，也是完全可以应用于非电商的领域，比如说前段时间碰到新华社的朋友，我很好奇他为什么对NLP感兴趣，他说新华社需每天要把大量的外文稿件翻译成中文。可见NLP技术已经渗透到很多传统非技术领域了。

但凡有文字处理需要自动化的地方，都需要用到NLP技术，所以掌握了些基本技术，你就可以应用在各行各业里边，解决各种实际问题。

# GitChat