Word2vec应用案例

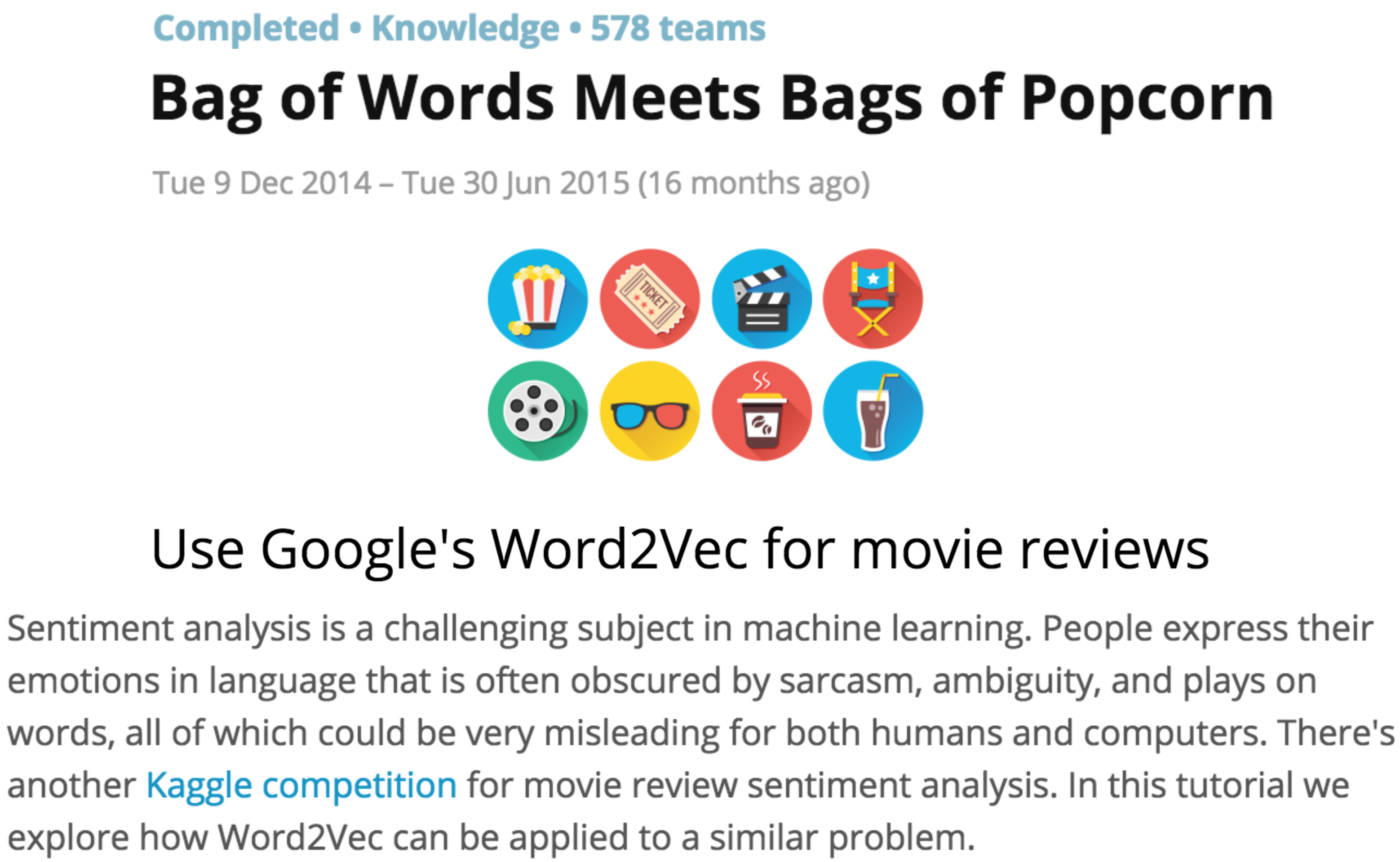
七月在线 寒小阳

2016年11月19日

1、文本情感分析
英文&&中文

情感
接受
惊异
哀伤
快乐

文本情感分析



Completed • Knowledge • 578 teams

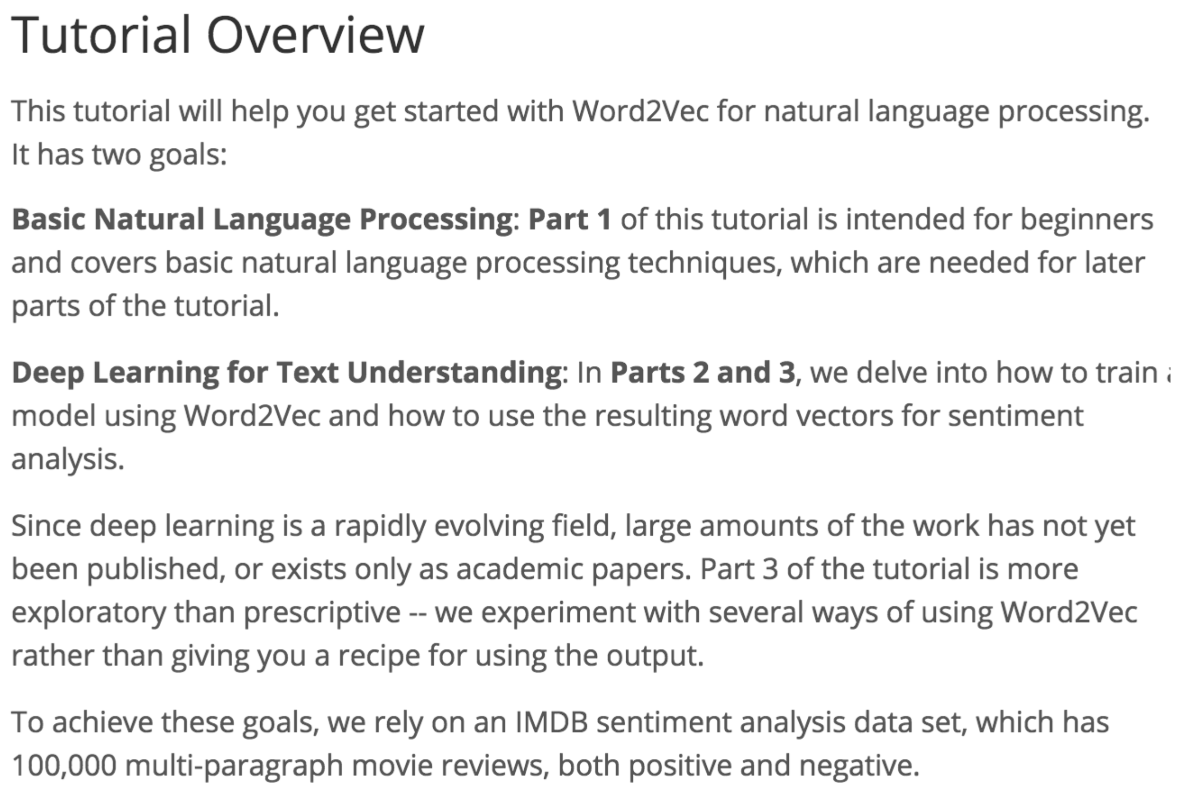
Bag of Words Meets Bags of Popcorn

Tue 9 Dec 2014 – Tue 30 Jun 2015 (16 months ago)

Use Google's Word2Vec for movie reviews

Sentiment analysis is a challenging subject in machine learning. People express their emotions in language that is often obscured by sarcasm, ambiguity, and plays on words, all of which could be very misleading for both humans and computers. There's another Kaggle competition for movie review sentiment analysis. In this tutorial we explore how Word2Vec can be applied to a similar problem.

文本情感分析



Tutorial Overview

This tutorial will help you get started with Word2Vec for natural language processing. It has two goals:

Basic Natural Language Processing: **Part 1** of this tutorial is intended for beginners and covers basic natural language processing techniques, which are needed for later parts of the tutorial.

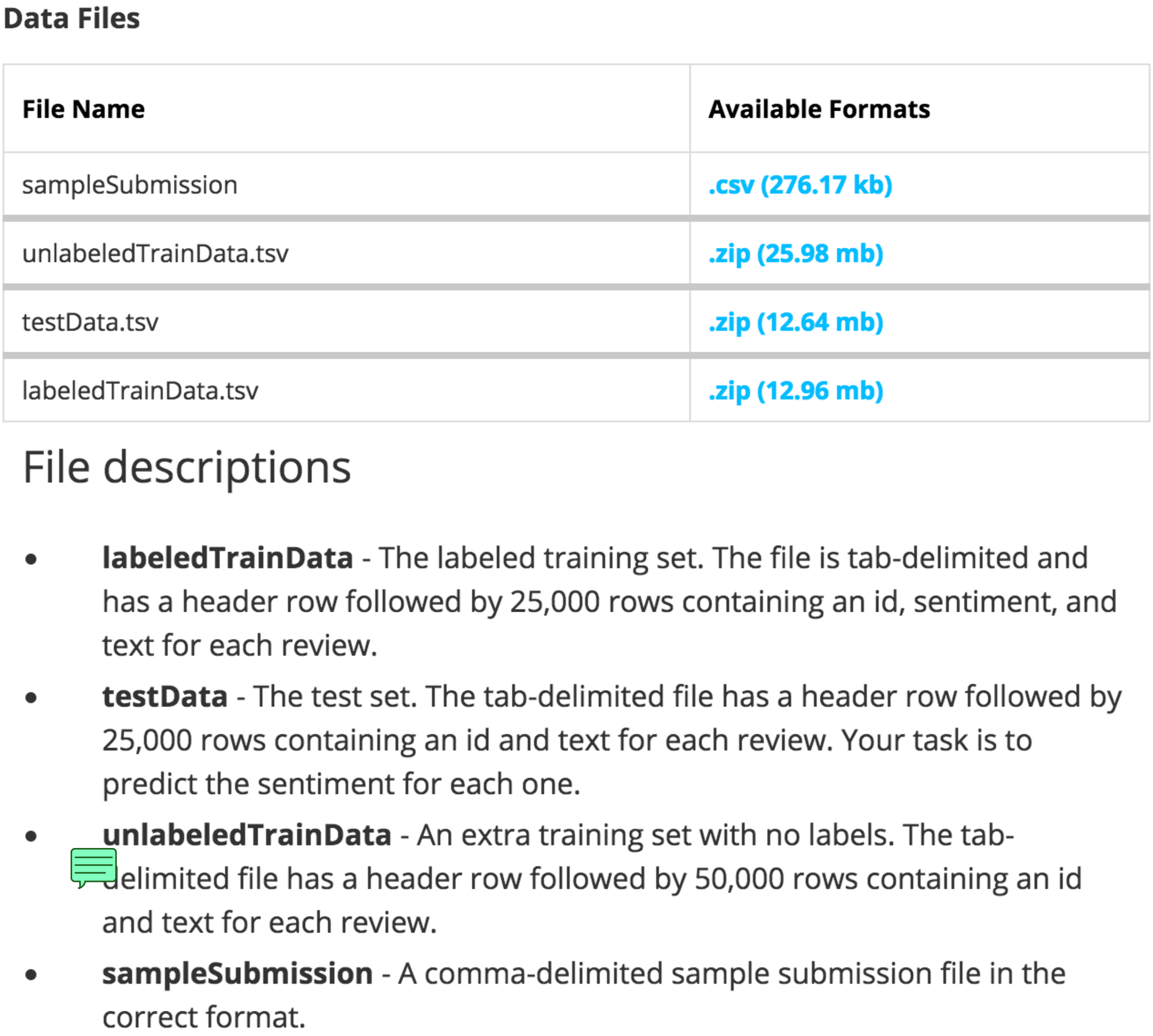
Deep Learning for Text Understanding: In **Parts 2 and 3**, we delve into how to train model using Word2Vec and how to use the resulting word vectors for sentiment analysis.

Since deep learning is a rapidly evolving field, large amounts of the work has not yet been published, or exists only as academic papers. Part 3 of the tutorial is more exploratory than prescriptive -- we experiment with several ways of using Word2Vec rather than giving you a recipe for using the output.

To achieve these goals, we rely on an IMDB sentiment analysis data set, which has 100,000 multi-paragraph movie reviews, both positive and negative.

1.基本的文本预处理技术（网页解析，文本抽取，正则表达式等）
2.word2vec词向量编码与机器学习建模情感分析

数据



Data Files

File Name	Available Formats
sampleSubmission	.csv (276.17 kb)
unlabeledTrainData.tsv	.zip (25.98 mb)
testData.tsv	.zip (12.64 mb)
labeledTrainData.tsv	.zip (12.96 mb)

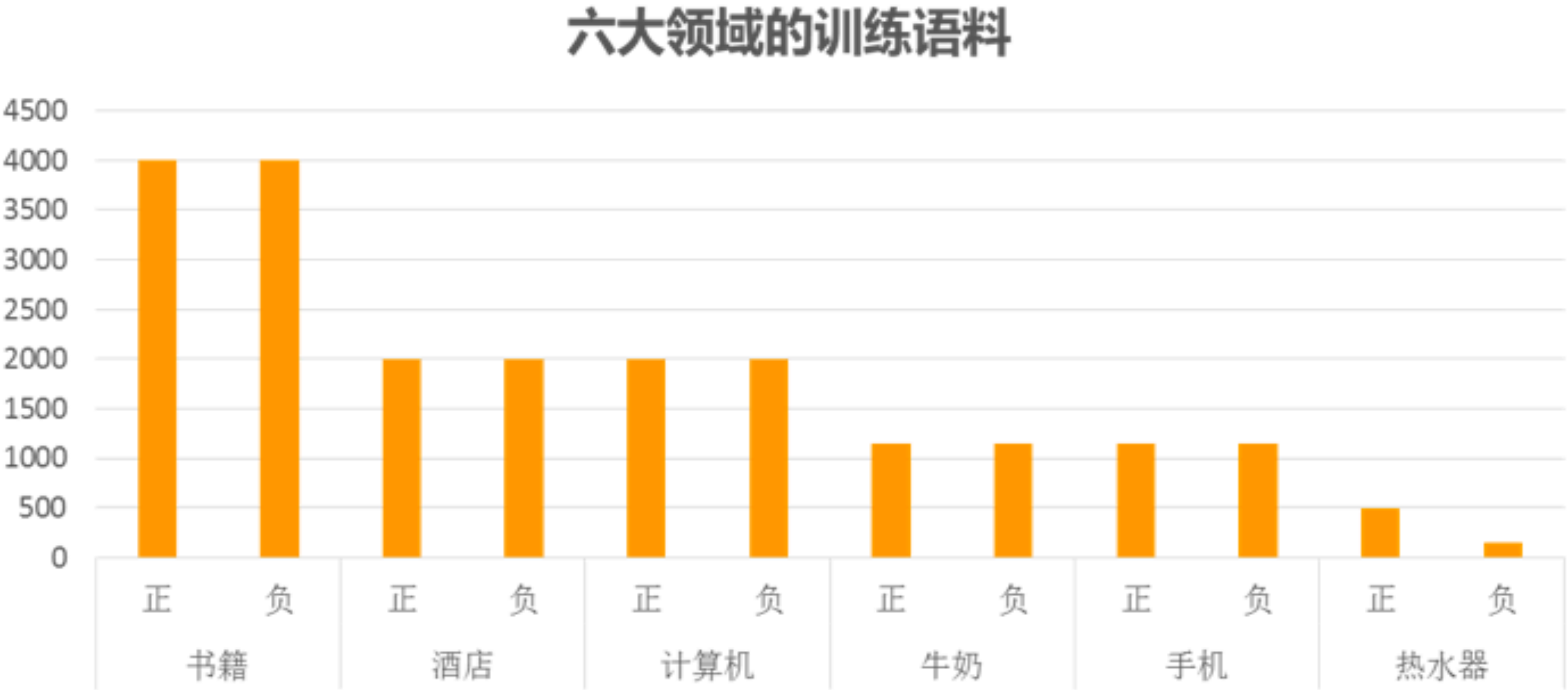
File descriptions

- **labeledTrainData** - The labeled training set. The file is tab-delimited and has a header row followed by 25,000 rows containing an id, sentiment, and text for each review.
- **testData** - The test set. The tab-delimited file has a header row followed by 25,000 rows containing an id and text for each review. Your task is to predict the sentiment for each one.
- **unlabeledTrainData** - An extra training set with no labels. The tab-delimited file has a header row followed by 50,000 rows containing an id and text for each review.
- **sampleSubmission** - A comma-delimited sample submission file in the correct format.

语料处理与模型构建过程

- 详见课程ipython notebook

中文数据构建与工具



- □ 训练集：测试集 = 8 : 2
- □ Sklearn => SVM, gensim => word2vec

神经网络:LSTM(一个有记忆的模型)可以沿着时间轴一个词就输出一个结果

语料处理与模型构建过程

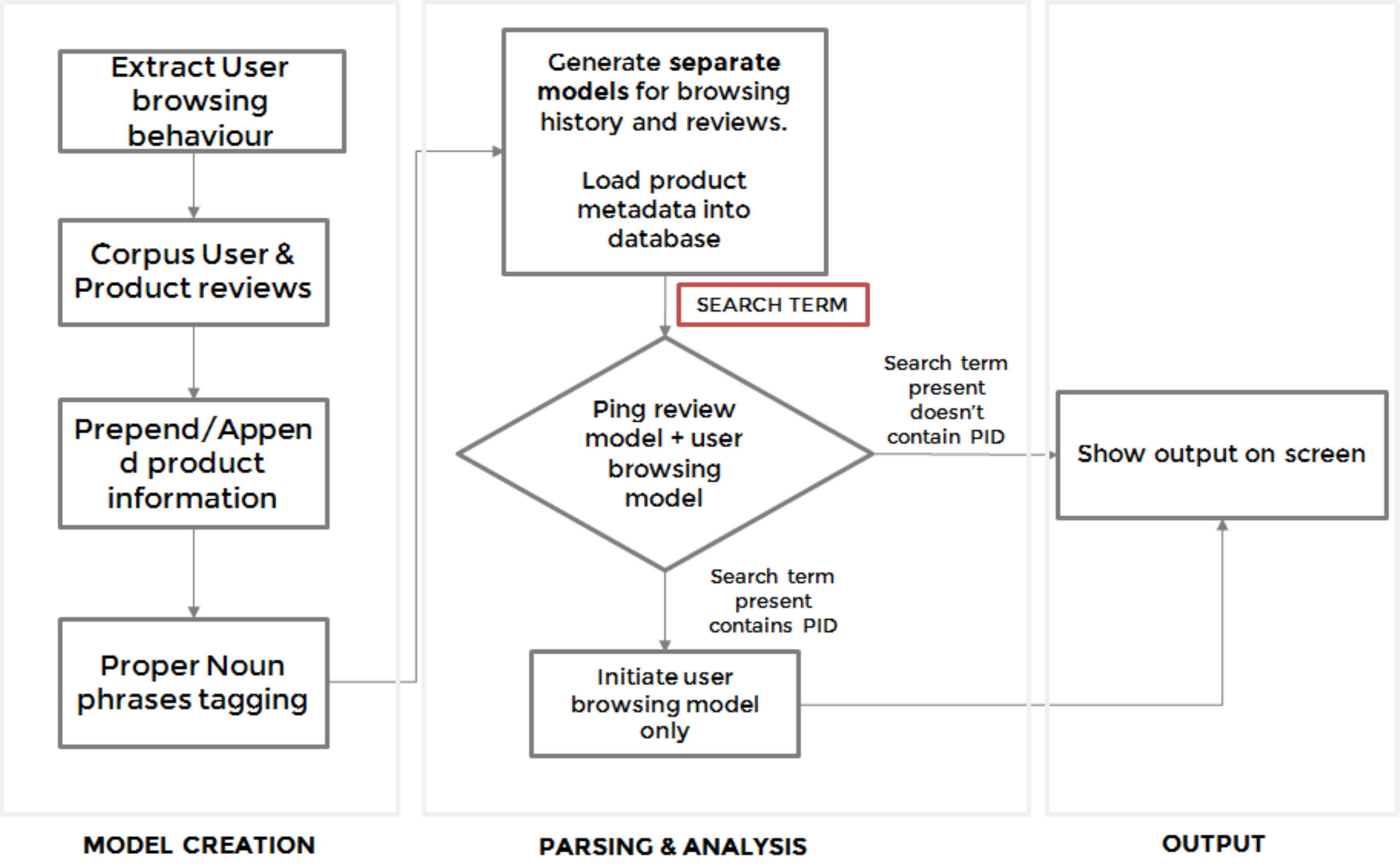
- 详见课程ipython notebook

有兴趣的同学可以试试gensim的doc2vec
同时使用LSTM神经网络分类
效果会比SVM更好

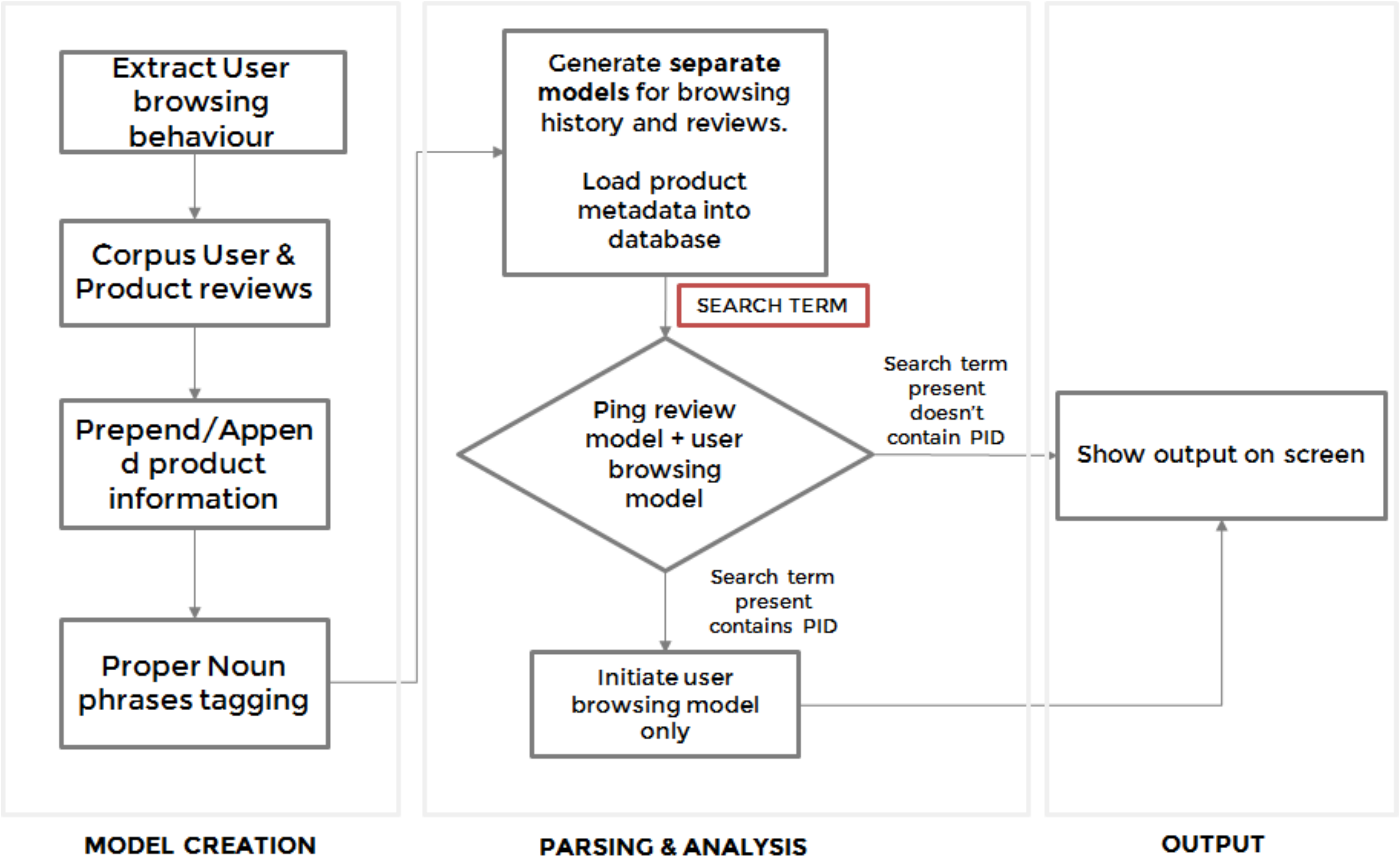
2、文本之外的应用

如何在亚马逊的用户行为和评论数据上构建推荐系统

怎么同时使用浏览 && 评论数据？


Extract User browsing behaviour
Corpus User & Product reviews
Prepend/Appen d product information
Proper Noun phrases tagging
Generate separate models for browsing history and reviews.
Load product metadata into database
SEARCH TERM
Ping review model + user browsing model
Search term present doesn't contain PID
Show output on screen
Search term present contains PID
Initiate user browsing model only
MODEL CREATION
PARSING & ANALYSIS
OUTPUT

同时使用用户浏览行为 && 评论数据？



Extract User browsing behaviour

Corpus User & Product reviews

Prepend/Append product information

Proper Noun phrases tagging

Generate **separate models** for browsing history and reviews.

Load product metadata into database

SEARCH TERM

Ping review model + user browsing model

Search term present doesn't contain PID

Search term present contains PID

Initiate user browsing model only

Show output on screen

MODEL CREATION

PARSING & ANALYSIS

OUTPUT

构建推荐系统

- 详见课程代码与讲解

感谢大家！

恳请大家批评指正！