Computer Science and Application 计算机科学与应用, 2018, 8(8), 1292-1299
Published Online August 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.88139

Hans汉斯

Research on Personalized Recommendation of Ancient Poetry Based on Word2vec Model

Zongliang Liu, Hao Huang

College of Information, University of International Business and Economics, Beijing
Email: Alex6687@163.com

Received: Aug. 5th, 2018; accepted: Aug. 20th, 2018; published: Aug. 28th, 2018

Abstract

Ancient poetry is the jewel of China's outstanding traditional culture. For two thousand years, China's outstanding poets have come forth in large numbers, and their poems are full of stars and stars, rich in content and far-reaching. With the continuous development of computer technology, the recommendation system is everywhere in our lives, providing convenience for more and more users. However, at present, there is a lack of personalized intelligent recommendation system for ancient poetry. Most poetry websites are only a simple display of poetry content, not a recommendation. Therefore, research on ancient poetry recommendation promotes the spread of Chinese excellent traditional culture. It is of great significance. Based on the Word2vec model, this paper realizes the personalized recommendation of ancient poetry by using the ancient poetry data crawled on the network.

Keywords

Recommended System, Ancient Poetry, Word2vec Model

基于Word2vec模型进行古诗词个性化推荐的研究

刘宗亮，黄　浩

对外经济贸易大学信息学院，北京
Email: Alex6687@163.com

收稿日期：2018年8月5日；录用日期：2018年8月20日；发布日期：2018年8月28日

摘　要

古诗词是中华优秀传统文化上璀璨的明珠，两千年来，我国优秀诗人辈出，其诗作若满天繁星，内容丰富，

文章引用：刘宗亮，黄浩．基于 Word2vec 模型进行古诗词个性化推荐的研究[J]．计算机科学与应用，2018，8(8): 1292-1299. DOI: 10.12677/csa.2018.88139

影响深远。随着计算机技术的不断发展，推荐系统在我们的生活中处处可见，为越来越多的用户提供了便利。然而，目前对于古诗词的个性化智能推荐系统比较匮乏，绝大多数的诗词网站也只是对于诗词内容的简单展示，而非推荐，所以进行古诗词推荐方面的研究对于促进中华优秀传统文化的传播具有重要意义。本文基于Word2vec模型，通过利用网络上爬取的古诗词数据进行训练，实现了古诗词的个性化推荐。

关键词

推荐系统，古诗词，Word2vec模型

1. 引言

习近平总书记在十九大上提出，我们要“坚定文化自信，推动社会主义文化繁荣昌盛” [1]。而古诗词作为中华文化之瑰宝，情感交流之载体，其传承和弘扬的重要性不言而喻。两千年来，我国诗人辈出，佳作如林，内容极为丰富，绚丽多彩。在互联网、大数据等技术蓬勃发展的今天，各种个性化智能推荐系统越来越多地应用于生活的方方面面，例如个性化视频推荐系统、个性化音乐推荐系统以及餐饮推荐系统等等，这赢得了越来越多用户的青睐。本文利用 Word2vec 模型，实现古诗词的个性化智能推荐，以期促进古诗词的传播，弘扬中华优秀传统文化。

本文首先利用开源爬虫框架 Requests 从互联网上爬取需要的古诗词数据，然后将爬取到的每一首古诗词的译文和赏析利用 Python 中的 jieba 分词库进行分词、去噪，生成古诗词语料库，进而利用该语料库训练 Word2vec 模型，并通过该模型得到每一首古诗词在该语料库中的向量表示，最后使用余弦距离计算出古诗词之间的相似度，选取相似度最高的十首古诗进行推荐。

2. 研究现状与相关技术

目前，绝大多数古诗词网站都只是对古诗词内容的简单展示，缺乏个性化推荐功能。本文利用深度学习的思想，从浩如烟海的古诗词语料库中训练出 Word2vec 模型，实现古诗词智能个性化推荐。相比于仅仅利用古诗词标签进行分类，Word2vec 模型能够更加准确的为用户推荐感兴趣的古诗词，从而使用户得到更好的体验效果。

2.1. 数据爬取

模型训练所使用的古诗数据是通过 Python 语言从网络上爬取所得，在此过程中主要用到了 Python 提供的两个库——Requests [2]和 BeautifulSoup [3]，其中 Requests 用于模拟 session/cookie 的存储和设置，BeautifulSoup 用于在进行网页抓取后的处理工作中，通过简短的代码完成过滤 html 标签，提取文本的工作。

Requests 是一款基于 Python 语言的开源爬虫框架，覆盖了典型爬虫的几大核心功能：页面下载、链接抓取、URL 管理和内容分析与持久化。Requests 是用 Python 语言编写，基于 urllib，采用 Apache2 Licensed 开源协议的 HTTP 库。它完全满足 HTTP 测试需求，比 urllib 更加方便，可以大大节省工作时间。Requests 主要方法及说明如表 1 所示。

BeautifulSoup 库是解析、遍历、维护“标签树”的功能库，BeautifulSoup 对应一个 HTML/XML 文档的全部内容。BeautifulSoup 类的基本元素如表 2 所示。

Table 1. Main methods and descriptions of Requests
表 1. Requests 主要方法及说明

方法	说明
Requests.request()	构造一个请求，支撑以下各方法的基础方法
Requests.get()	获取 HTML 网页的主要方法，对应于 HTTP 的 GET
Requests.head()	获取 HTML 网页头信息的方法，对应于 HTTP 的 HEAD
Requests.post()	向 HTML 网页提交 POST 请求的方法，对应于 HTTP 的 POST
Requests.put()	向 HTML 网页提交 PUT 请求的方法，对应于 HTTP 的 PUT
Requests.patch()	向 HTML 网页提交局部修改请求，对应于 HTTP 的 PATCH
Requests.delete()	向 HTML 页面提交删除请求，对应于 HTTP 的 DELETE

Table 2. Basic elements of the BeautifulSoup class
表 2. BeautifulSoup 类的基本元素

基本元素	说明
Tag	标签，最基本的信息组织单元，分别用<>和</>表明开头和结尾
Name	标签的名字， … 的名字是‘p’，格式：<tag>.name
Attributes	标签的属性，字典形式组织，格式：<tag>.attrs
NavigableString	标签内非属性字符串，<>…</>中字符串，格式：<tag>.string
Comment	标签内字符串的注释部分，一种特殊的 Comment 类型

2.2. 中文分词

中文分词(Chinese Word Segmentation)指的是将一个汉字序列切分成一个个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。

在获取数据后，需要对数据进行分词、去噪，生成古诗词语料库，此过程主要采用了 Jieba 分词工具。

Jieba 分词[4]是一款基于 Python 语言的简单易用的中文分词工具，支持三种分词模式：

① 精确模式，将句子精确分开，适合文本分析；

② 全模式，把句子所有成词的词语都扫描出来，但是不能解决歧义；

③ 搜索引擎模式，对长词进行切分，提高召回率。

Jieba 分词用到的算法：

基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图(DAG)采用了动态规划查找最大概率路径，找出基于词频的最大切分组合对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

2.3. Word2vec 模型

Word2vec 是 Google 在 2013 年提出的用于快速有效地训练词向量的模型。作者的目标是要从海量的文档数据中学习高质量的词向量，该词向量在语义和句法上都有很好地表现，现已广泛应用于自然语言处理的各种任务中。Word2vec 通过训练，可以把对文本内容的处理简化为 K 维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度[5]。因此，word2vec 输出的词向量可以被用来做很多 NLP 相关的工作，比如聚类、找同义词、词性分析等。

Word2vec 的基本思想是利用上下文信息，即与一个词前后相邻的若干个词，来提取这个词的特征向

量。为了利用这种上下文信息，Word2vec 采用了两种具体的模型：CBOW 和 Skip-Gram。这两种模型的本质是一样的，都是利用句子中相邻的词，训练一个神经网络。它们各有优劣，因此各自实现的 Word2vec 的效果也各有千秋。从图 1 可以看出两种模型均包含输入层、投影层和输出层。其中，CBOW 是一种根据上下文的词语预测当前词语出现概率的模型。而 Skip-Gram 则是逆转了 CBOW 的因果关系，即已知当前词语，预测上下文。

同时，Word2vec 提供了两套优化方法来提高词向量的效率[6]，分别是 Hierachy Softmax 和 Negative Sampling，将训练模型和优化方法进行组合可得到 4 种训练词向量的框架。

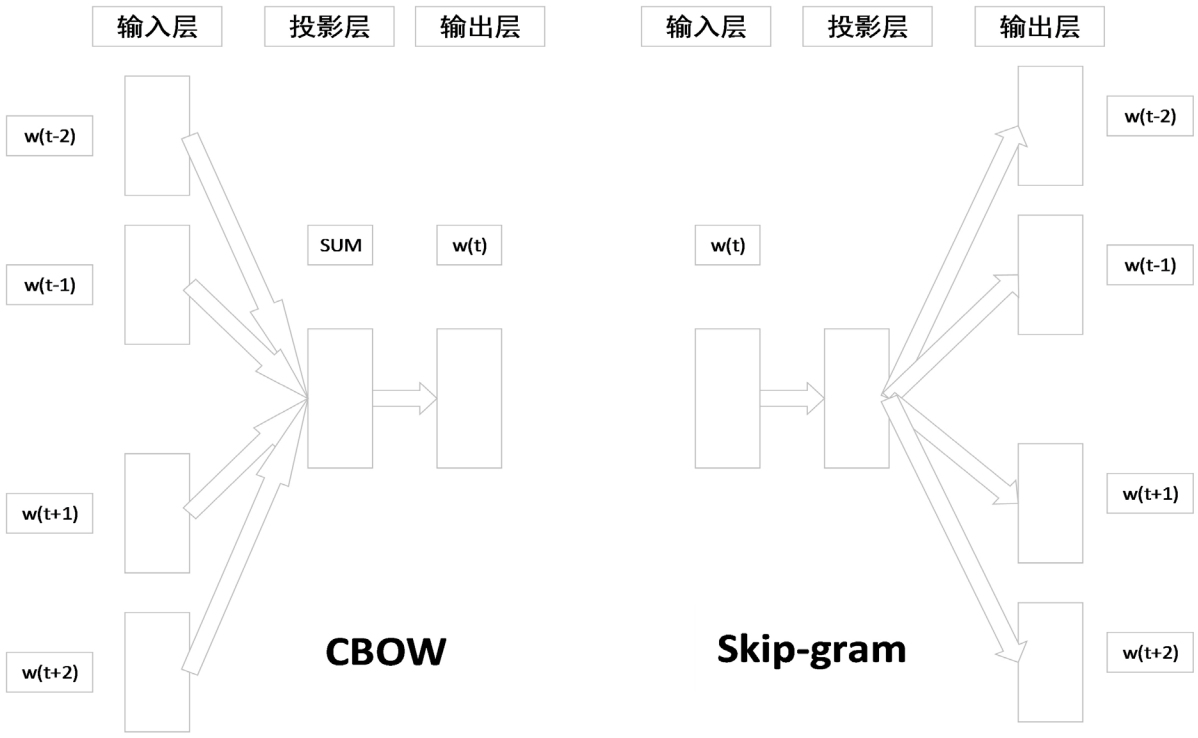
Hierarchical Softmax(HS)模型是利用一个二叉树去表示特征词典中的所有特征词[7]。特征词典中的 V 个特征词作为二叉树的叶子节点，所以非叶子节点有 V-1 个，对于每一个叶子节点，有且仅有一条从根节点到该叶子节点的路径，模型利用该路径去估计该叶子节点所代表的特征词的概率，如图 2 所示。

负抽样(Negative Sampling, NS)的思路比较简单直接：针对每一个训练样本，原始模型都要更新所有的“输出向量”，而 NS 模型仅仅只更新“输出向量”中的一部分。它的主要思想是：选择部分负样本(非目标特征词)协助正样本(目标特征词)的相关参数(“输出向量”)进行更新。

3. 模块设计与实现

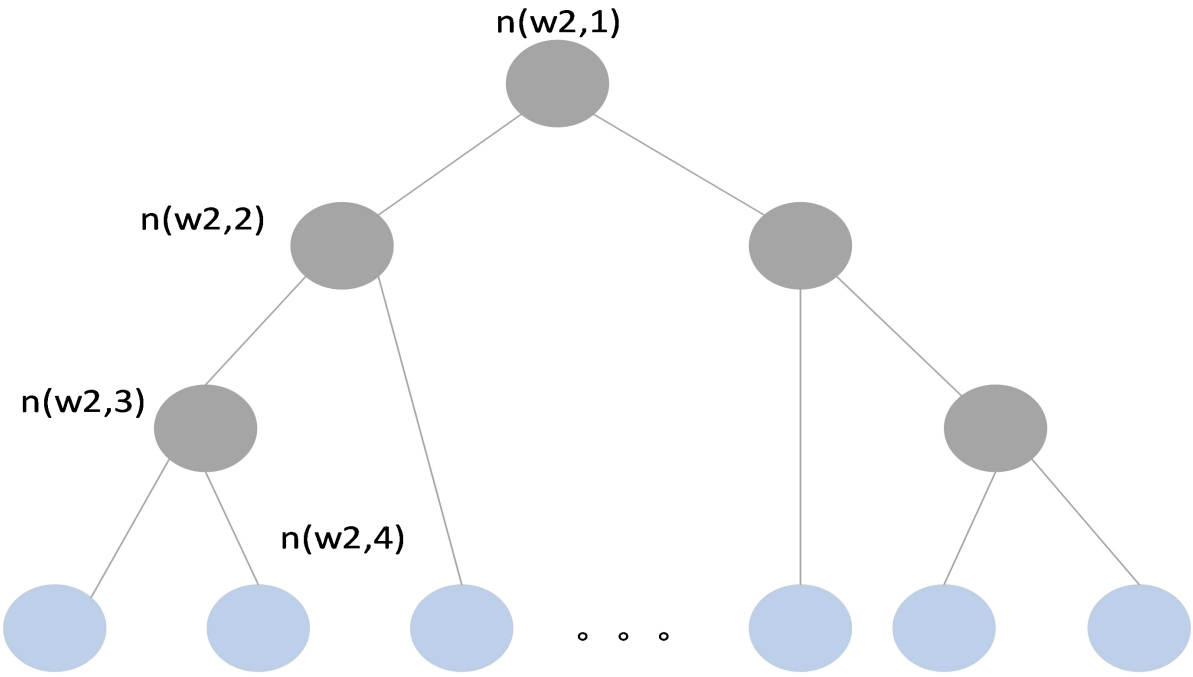
3.1. 流程图

本研究主要分为四个模块：数据爬取模块、中文分词模块、模型训练模块以及兴趣推荐模块，根据其模块分类及扩展可得整体流程图如图 3 所示。



注：其中 w(t)代表当前词语位于句子的位置 t，同理定义其他记号。在窗口内除了当前词语之外的其他词语共同构成上下文。

Figure 1. Network structure of CBOW model and Skip-gram model
图 1. CBOW 模型和 Skip-gram 模型的网络结构



注：其中，n(w, j)表示从根节点通往特征词 w 所代表的叶子节点中的第 j 个节点，其路径长度表示未 L(w)，故 L(w2) = 4。

Figure 2. A simple example of the HS model
图 2. HS 模型的一个简单示例

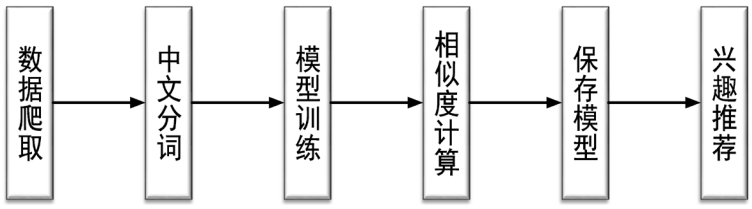


Figure 3. Module design flow chart
图 3. 模块设计流程图

3.2. 数据爬取模块

数据爬取模块实现的功能是通过爬虫技术从互联网获取古诗词数据，作为整个系统的原始数据。在该模块中，本文采用的是 Requests 爬虫框架，编写 Python 程序，从古诗文网爬取古诗词数据，包括古诗词的标题、作者、朝代、原文、翻译、赏析、链接等内容，并将爬取结果保存为 CSV 文件。

古诗文网是静态网页，网站的首页包含唐诗三百、宋词三百、古诗三百、诗经、楚辞等诸多门类的古诗词，在进行数据爬取时分别针对不同的类型进行先后爬取，完成某一类型的爬取之后再进行另一类型的爬取，直到爬取完所有类型的古诗。以爬取唐诗三百为例，具体处理流程如图 4 所示。

3.3. 中文分词模块

中文分词模块实现的功能是将数据爬取模块中爬取到的每一首古诗词的翻译和赏析序列切分成一个个单独的词，作为训练 Word2vec 模型的语料库。在该模块中，本文采用的是 Jieba 分词组件，首先从 CSV 文件中读取每一首古诗词的相关数据，然后使用 Jieba 分词组件对翻译和赏析进行分词，最后与标题联系起来，构成上下文关系，并保存为 txt 文件。

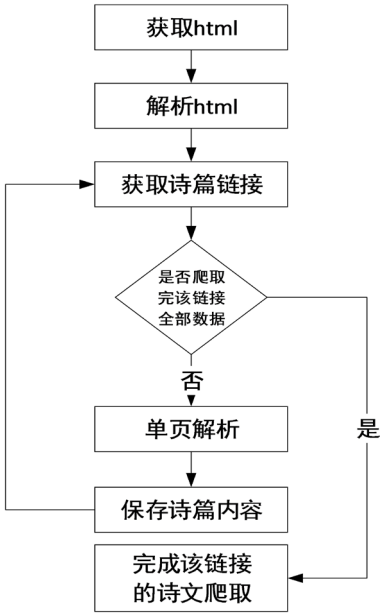


Figure 4. Ancient poetry crawling flow chart
图 4 古诗词爬取流程图

具体来说，本文采用 Jieba 分词的精确模式，将爬虫爬取到的每一首古诗词的翻译和赏析序列切分成一个个单独的词，作为训练 Word2vec 模型的语料库。例如“数风流人物，还看今朝”经过 Jieba 分词处理之后可以得到“数”、“风流”、“人物”、“还”、“看”、“今朝”等词汇。同时，因为在网络上爬取到的数据并不是非常标准的，其内容包含标点符号等无关的信息，所以在分词结束之后，需要对数据再进行去噪处理，去除标点等无关的信息，最终得到模型训练所需要的数据。

3.4. 模型训练模块

模型训练模块实现的功能是训练出一个模型，通过该模型能够比较准确地计算出任意两首古诗词的相似度。在本模块中，采用 Word2vec 对中文分词模块中得到的古诗词语料库进行建模，通过训练将词表征为 K 维实数向量，然后根据每一首古诗词的翻译和赏析内容计算它们的向量表示，最后通过古诗词之间的距离(如余弦相似度、欧氏距离等)来判断它们之间的语义相似度。

在 word2vec 的训练过程中，主要涉及以下几个重要参数[8]，如表 3 所示。

表 3 word2vec 模型的主要训练参数。

根据两种模型和两种算法自身的特点以及文献研究、多次实验，最终本研究将 sg 参数设为 1，即采用 skip-gram 模型，将 hs 参数设为 1，即采用 Hierarchical Softmax 算法，二者组合，采用 skip-gram + HS 框架，实验证明此种搭配下取得了更好的推荐效果。对于其它的参数，多次实验表明并不会对实验结果产生非常显著的影响，因此根据文献研究，均采取默认的参数。例如：alpha 表示初始的学习速率，在训练过程中会线性地递减到 min_alpha (学习率的最小值)；min_count 表示最低频率，可以对字典做截断，词频少于 min_count 次数的单词会被丢弃掉，默认值为 5；sample 表示高频词汇的随机降采样的配置阈值，默认为 1e-3，范围是(0, 1e-5)。

skip-gram + HS 框架是已知当前词，预测上下文(context(w),w)。输入层为当前词 w 的词向量，投影层为当前词 w 对应的词向量(恒等投影)，输出层为以词语在语料库中的词频作为权值构造的一颗二叉树[7]。叶子节点对应词汇表中的所有词语。假设叶子节点为 N 个，则非叶子节点为 N-1 个。叶子节点和非

Table 3. Main training parameters of the word2vec model
表 3. Word2vec 模型的主要训练参数

训练参数	参数含义
size	词向量的维数
window	训练窗口大小
alpha	学习速率
min_count	最低频率
workers	控制训练的并行数
sample	采样的阈值
sg	是否采用 skip-gram
cbow	是否采用 cbow，不可与 sg 同时使用
hs	是否采用 Hierarchical Softmax 算法
negative	是否采用 Negative Sampling 算法

叶子节点均对应一个向量。其中叶子节点对应的向量即为词向量，而非叶子节点对应的向量是一个辅助向量。

对于 skip-gram + HS 框架，给定一个训练样本(w,context(w))，词语 w 前后各 c 个词，context(w)包含 2c 个词。可以将通过 w 预测 context(w)的问题，即 $p\left(context\left(w\right)\middle|w\right)$ 转换为 2c 个通过 w 预测下一个词为 u 的问题 $p\left(u\middle|w\right)$，$u\in context\left(w\right)$，其目标函数为 $p\left(context\left(w\right)\middle|w\right)=\prod_{u\in context\left(w\right)}p\left(u\middle|w\right)$。其中，$p\left(u\middle|w\right)$ 可以利用将 u 视为叶子节点的思路来解决。

3.5. 兴趣推荐模块

兴趣推荐模块实现的功能是利用模型训练模块训练出的模型，针对每一首古诗词计算与其最相似的十首古诗，并将结果保存为 txt 文件，以备下次查询。在本模块中，调用了 Word2vec 模型的 similarity 方法，计算出古诗词之间的相似度，按相似度降序排序，截取前十首古诗词。

相似度计算公式采用的是余弦相似度。余弦相似度用向量空间中两个向量夹角的余弦值来衡量两个文本间的相似度，相比于距离度量，余弦相似度更加注重两个向量在方向上的差异，计算公式如下：

$$\cos\left(\theta\right)=\frac{\sum_{i=1}^{n}x_i,y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}}$$

其中，x_i,y_i 分别表示词语转化为的向量。在模型训练模块结束之后，通过构建词频向量，词频的数值为该词在句子中出现的次数，通过上述公式计算两个向量夹角的余弦值，结果越趋近于 1 表示相似度越高。进而，按照所得的相似度进行排序，截取前十名作为推荐的古诗词输出。

4. 系统实现和实验分析

本文利用 Python 语言开发了一个基于 word2vec 模型的古诗词个性化推荐系统，在系统完成之后，我们进行了实验。当给予关键词“送别”之后，系统首先通过该标题联系到相关的古诗信息，比如：作者、朝代、诗文、译文以及赏析等，然后，将这些内容转换为了相应的词向量。转换成的词向量和语料库中其它的词向量通过余弦相似度的计算得到彼此之间的距离关系并排序，最后截取前十名的古诗。系

统最终输出的十首古诗分别为：梅花、江南、南山、登高、相思、鸡鸣、春日、登楼、北风和风雨。因为送别诗中包含的情感大多是伤感离别的，因此推荐的大多是感伤惆怅的古诗。且送别诗的标志性事物有梅花、风雨、高楼等，因此推荐的古诗还包括和标志物相关的诗词。总体而言，推荐的十首古诗和用户给定的古诗从内容、语境、情感等方面具有很大的相似度，当用户想搜索与某首古诗相关联的诗词时，该系统能够提供较好的个性化智能推荐。

实验结果表明，系统基本实现了古诗词个性化推荐的目的，推荐的古诗大部分是与给定的古诗相关的，但在系统的运行效率和风格情感方面还有待改进。

5. 结论

Word2vec 是一款用于训练词向量的软件工具，提供了 CBOW 和 Skip-gram 两种模型。结合 hierarchy softmax 和 negative sampling 优化技术，Word2vec 可以快速高效地将词语表达成向量。Word2vec 本身的特点使其效率变得很高，主要包括：Word2vec 去掉了费时的非线性隐层；其次，Huffman 编码相当于做了一定聚类，不需要统计所有词对；而且 Word2vec 只需过一遍数据，不需要反复迭代。诸多优点相加使得 Word2vec 可以在百万数量级的词典和上亿的数据集上进行高效的训练。

诗歌是人类文学皇冠上璀璨的明珠。《诗经》而后，两千余年的时间里面，我国诗人辈出，其诗作若满天繁星，令人赞叹不已。随着计算机技术的不断发展，智能个性化推荐系统广泛地应用于生活中，为越来越多的用户提供了方便。然而，目前对于古诗词智能推荐的系统寥寥无几，古诗词网站大多只是对古诗词内容的简单展示，缺乏个性化推荐功能。本文利用深度学习的思想，基于 Word2vec 模型，实现了古诗词的智能个性化推荐，在促进古诗词的传播，弘扬中华优秀传统文化方面具有重要的意义。

基金项目

国家重点研发计划资助(2017YFB1400700)。

参考文献

[1] 习近平.《决胜全面建成小康社会 夺取新时代中国特色社会主义伟大胜利——在中国共产党第十九次全国代表大会上的报告(2017 年 10 月 18 日)》[J]. 美与时代(上), 2017(10): 4-22.

[2] 冰霜. BeautifulSoup 库基础及一般元素提取方法[EB/OL]. https://www.cnblogs.com/hanmk/p/8724162.html, 2018-04-05.

[3] 禾如月. 网络爬虫之规则之 requests 库入门[EB/OL]. https://blog.csdn.net/xiu_star/article/details/70156835, 2017-04-13.

[4] 武婷婷. 一种基于 WebMagic 和 Mahout 的信息搜集与推荐系统[J]. 软件导刊, 2016, 15(10): 1-3.

[5] 张祖平, 沈晓阳. 基于深度学习的用户行为推荐方法研究[J]. 计算机工程与应用.

[6] 唐明, 朱磊, 邹显春. 基于 Word2Vec 的一种文档向量表示[J]. 计算机科学, 2016, 43(6): 214-217.

[7] 周练. Word2vec 的工作原理及应用探究[J]. 图书情报导刊, 2015(2): 145-148.

[8] 王飞, 谭新. 一种基于 Word2Vec 的训练效果优化策略研究[J]. 计算机应用与软件, 2018(1): 97-102.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org