

AI Meets Archaeology: 新科技辅助文物分类与演化分析

魅力考古学课程论文

建议评阅老师：王晓琪

尹琪钦 211870080

南京大学物理学院

211870080@smail.nju.edu.cn

摘要

本文探讨了人工智能（AI）技术在考古学领域的应用，特别是其在分析历史文物属性和历史演化方面的潜力。介绍了科技考古学的概念，强调了其独特优势，并以古代钱币分类和云梦睡虎地秦简的文本分析为例，展示了 AI 技术的实际应用。提出了构建 AI 辅助考古分析框架的基本准则和步骤流程。介绍了原创性的考古分析 AI 模型工作，实现对不同历史时期“龍”字形态的分类和聚类，模型揭示了其演化路径，并能识别未知“龍”字图片的字体。文章展现了 AI 在文物分类、年代推断和属性推断等考古分析中的有效性，展现其辅助考古学的巨大应用前景，为科技考古提供了新的分析框架。

关键词：科技考古，人工智能，历史演化，文物分析

1 引言

在历史的长河中，考古学作为一门探索人类过往的科学，很大程度上依赖于实地挖掘、文献研究和物质分析等传统方法，始终致力于解读古代文明留下的物质文化遗产。随着科技的飞速发展，传统的考古学方法正逐渐与现代技术相结合，形成了一个新兴的交叉学科领域——科技考古学。这一领域利用先进的科技手段，如放射性核素测年、地球物理勘探、遥感技术、DNA 分析以及人工智能（AI），极大地提高了考古工作的效率和精确度，帮助解决传统考古学中的难题，为考古学研究开辟了新的视角和深度 [1]。

人工智能（AI）作为一门新兴的学科，已经在多个领域展现出其强大的数据处理和模式识别能力。在考古学领域，AI 技术正逐渐被应用于文物的分类、年代的推断、遗址的发现与分析等方面。AI 算法能够处理复杂的考古数据集，识别文物之间的细微差别，揭示文物背后的历史联系，并推测未知文物的属性。当前，AI 在考古学中的应用仍处于起步阶段，但其潜力已被广泛认可，特别是在分析历史演化趋势、分类未知文物以及推断其可能属性方面 [2][3]，并有望在未来的考古研究中发挥更加重要的作用。

在本次“魅力考古学”课程论文中，介绍了原创性开发的一个 AI 辅助的考古分析框架，该框架通过机器学习算法对已知汉字的历史演化进行初步分析，建立一个可靠的分类和属性推断模型。其中以汉字“龙”的五个历史时期演化分析为例，展示 AI 技术在解析具体考古问题中的应用。该模型容易推广并被应用于未知文物的分析，以期揭示其年代、用途和文化背景，为考古学研究提供启发。

通过对 AI 辅助考古分析框架的构建与测试，本课程项目不仅期望为考古学领域提供一种新的研究工具，更希望能够推动科技与考古学的深度融合，为发掘历史的真相和演化趋势提供新的途径。

2 科技考古学与 AI 的结合

科技考古学，作为考古学的一个分支，它利用现代科学技术对古代文明的物质文化遗产进行分析和研究。这一领域的研究不仅限于对文物的物理和化学属性的分析，还包括对古代社会结构、经济活动、文化交流和技术发展等方面的深入探讨。科技考古学的发展历史可以追溯到 20 世纪中叶，随着科学技术的进步，其研究方法和应用范围不断扩大，如今已成为考古学不可或缺的一部分。

科技考古学的重要性在于它能够提供传统考古学方法无法获取的信息。例如，通过放射性碳定年技术可以准确地确定文物的年代；通过同位素分析可以了解古代人类的饮食习惯；通过材料科学可以鉴定文物的制作工艺和原材料来源。这些信息对于重建古代社会的历史具有极其重要的价值。

近年来，人工智能（AI）技术在考古学领域的应用日益广泛，它为考古学研究带来了新的研究方法和启发。其中机器学习作为 AI 的一个分支，更是前景广阔，它通过算法从数据中学习并做出预测或决策，而无需进行显式的编程。在考古学中，机器学习可以应用于文物的分类、年代的推断、遗址的特征分析等任务。通过训练机器学习模型识别已知文物的特征，可以构建一个能够识别和分类未知文物的系统。此外，机器学习还可以帮助考古学家发现文物之间的潜在联系，为考古学研究提供新的视角。在此列举 AI 技术在考古学中的应用主要包括几个可能的方面：

图像识别与分类：图像识别与分类是人工智能领域的关键技术之一，它使得机器能够从图像中识别和区分不同的对象。在考古学中，这项技术尤其有价值，因为它可以用于自动化地识别和分类大量的考古发现，如陶器、工具、雕塑、钱币等。通过图像识别，考古学家能够节省大量的时间和资源，同时提高分类的准确性和效率。作为一个具体例子，借助科学计算软件 Mathematica，我们实现了基于机器学习对于“秦半两”、“汉五铢”、“开元通宝”这三种钱币的分类，结果如图 (1) 所示。

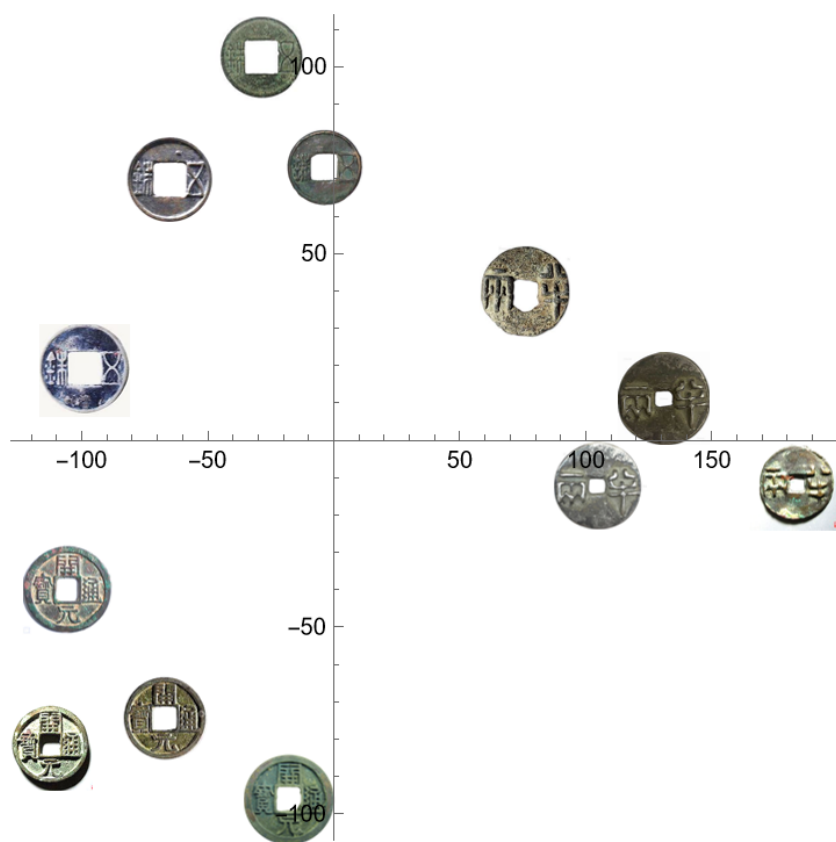


图 1: 三种古代钱币在特征空间中被良好区分

其中的具体步骤有:

1. 收集每种钱币的四张高质量图像，确保图像在不同的光照和角度下拍摄，以增加模型的泛化能力。
2. 对收集的图像进行背景淡化和对象锐化，突出钱币的特征，减少背景噪声对模型识别的干扰，以便机器学习模型能够更准确地识别和分类对象。
3. 使用图像处理技术提取钱币的关键特征，如形状（钱币的轮廓和整体形状）、文字和图案（钱币上的文字、符号和图案的布局和样式）、纹理（钱币表面的纹理特征，如磨损模式）等。这个过程涉及将图像转换为机器学习模型可以处理的数值形式，即特征向量。利用这些特征，机器学习模型能够在高维特征空间中对不同的钱币进行聚类。例如，通过无监督学习算法如 K-means 或层次聚类，模型可以将相似的钱币图像聚集在一起，形成不同的类别。
4. 模型训练：利用提取的特征训练机器学习模型，如支持向量机 (SVM) 或卷积神经网络 (CNN)，以识别和分类不同的钱币。
5. 使用训练好的模型对新的钱币图像进行分类，并与已知的分类结果进行比较，验证模型的准确性。

通过这种方法，考古学家可以快速地对大量文物进行自动化分类，甚至在没有详细历史记录的情况下，也能对未知的文物进行初步的年代和文化归属推断，一定程度上提高了工作效率。

文本分析：文本分析是 AI 在考古学中应用的另一重要领域，它涉及对古代文献和手稿的自动化解析，以提取有用信息和知识。通过文本分析，考古学家可以更好地理解古代语言、法律、社会结构和文化习俗。文本分析的关键技术包括自然语言处理 (NLP)、机器学习、模式识别和数据可视化。同样的，我们对云梦睡虎地秦简的文本内容进行了初步的分析，并介绍 AI 进一步潜在的应用。云梦睡虎地秦简作为出土的中国秦朝时期的实体文献资料，包含了丰富的法律、经济和社会信息 [4][5]。本次课程项目中实现了初步的工作，绘制了《秦律十八种》的词云图，词云图中词汇的大小与其在文本中出现的频率成正比，这有助于快速识别文本的关键主题和概念，可以直观地展示文本中出现频率较高的词汇，结果如图 (2)。词云图中直观展现了《秦律十八种》的主要内容，也能看出文本中有



图 2: 云梦睡虎地秦简-《秦律十八种》词云图

诸多缺失的内容和输入法尚未识别的文字，由于是初步的工作，这种局限性尚可接受。在词云图的基础上，AI 可以在文本分析中发挥更深入的作用，一个有可观应用价值的在于利用机器学习模型，

尤其是深度学习中的预测模型，可以推测文本中缺失或损坏部分的内容。例如，通过训练模型识别《秦律十八种》中已知的法律条文和语言表达模式，可以预测缺失文字的可能内容，揭示隐藏在文本背后的深层信息，为历史文献研究提供新的视角和工具。

此外还有若干应用方面如模式识别、三维重建，也可以借助机器学习识别和学习文物的特定模式，推测不同文化或时期的文物特征，对破损的文物进行虚拟修复，恢复其原始形态，为研究提供更完整的信息……

3 AI 辅助考古分析框架的构建

在构建 AI 辅助考古分析框架时，为了确保模型切实为考古工作带来便利，必须确立一系列基本的设计原则。首先，框架设计需融合多学科知识，因为考古学本身是一个跨学科领域，涉及历史、艺术、人类学和地质学等多个学科，这要求模型具备多领域的专业知识，以及对自然语言良好的处理能力，而此正好是大语言模型所擅长的。其次，框架应以数据为核心，利用大量考古数据来训练和优化 AI 模型，确保分析的准确性和可靠性。同时，框架的设计要具备可扩展性，以适应未来可能的数据和需求。此外，用户友好性也是设计时需要考虑的重要因素，框架应简单易用，使得非技术背景的考古学家也能轻松上手。

技术实现方面，AI 辅助考古分析框架的构建涉及多个关键步骤。首先是数据预处理，这一步骤至关重要，它包括数据清洗、规范化、缺失值处理等，以确保数据的质量。接下来是特征提取，从预处理后的数据中提取有助于模型训练的特征，如图像的纹理、形状、颜色特征，文本的词频、语法结构等。然后是模型选择，根据分析任务的类型（如分类、聚类、预测等）选择合适的机器学习模型。常见的模型包括决策树、支持向量机、神经网络等。模型训练与优化是随后的步骤，使用考古数据集对选定的模型进行训练，并采用交叉验证等方法优化模型参数。对于复杂的任务，可能需要使用集成学习方法，如随机森林或梯度提升机，以提高模型的性能。深度学习技术，尤其是卷积神经网络（CNN）或循环神经网络（RNN），在处理复杂的图像或文本分析任务时尤为关键。

框架的验证是确保其有效性的关键环节。可以通过使用已知文物数据集对模型进行测试，如年代、文化背景和类型都明确的文物，来评估模型的性能。性能评估可以通过准确率、召回率、F1 分数等指标来进行，这些都是计算机领域较为成熟的理论框架，适合交叉学科的应用。用户反馈是验证过程中不可或缺的一部分，通过收集使用框架的考古学家的反馈，可以了解框架在实际工作中的表现，并据此进行改进。持续迭代是框架验证的最后一步，根据验证结果和用户反馈，不断优化框架，以适应不断变化的研究需求。

以云梦秦简的 AI 分析为例，框架的验证和应用可以这样进行：首先，对秦简的数字化文本进行预处理，包括去除无关符号、分词、词性标注等，以准备数据用于后续分析。然后，提取文本特征，如词频、词汇的共现网络、语法结构等，这些特征对于训练文本分类模型至关重要。接着，使用提取的特征训练如支持向量机（SVM）或深度学习模型等分类器。在模型训练完成后，通过交叉验证等方法评估模型的性能，并与词云图等可视化结果进行对比，以验证模型的有效性。此外，利用训练好的模型预测秦简中缺失的文字，并通过考古学家的验证来评估预测的准确性。主题建模技术，如 LDA，可以用于分析秦简中的主题分布，为考古学家提供新的研究视角。最后，收集考古学家对 AI 分析结果的反馈，并据此对框架进行调整和优化。

可以看出，AI 辅助考古分析框架的构建是一个系统化的过程，它结合了多学科知识、数据科学、机器学习等多个领域的技术和方法。通过严格的设计原则、精心的技术实现和周到的框架验证，该框架有望成为考古学研究的有力工具，推动考古学领域的发展。随着 AI 技术的不断进步，该框架在考古学中的应用前景将更加广阔。

4 一个具体的例子：汉字“龙”的演化分析

文字演化研究一直是考古学的重要研究课题，它不仅为确定古代文物的年代提供了关键线索，而且揭示了文化交流、语言发展和技术进步的轨迹。文字变化反映了社会结构、艺术审美、宗教信仰及法律制度的演进，是理解古代文明不可或缺的一部分。此外，它还有助于我们探索不同社会群体的身份认同和知识传播的历史模式。在这次课程项目中，选择了汉字“龍”作为研究对象，借助 Mathematica 软件初步实现了对“龍”字的演变过程可视化，并实现了机器自动识别未知字体，提供了研究文字演化、辨别未知文字的新方法。

4.1 数据收集与预处理

数据收集是分析“龍”字演化的第一步。我们通过各种渠道搜集了“龍”字在不同历史时期的书写形式，初步分为甲骨文、周代金文、篆文、隶书和楷书五类，基本符合“龍”字历史演化规律，每一类各五张图片作为模型训练样本。这些样本不仅来自考古发掘的文物，也包括历史文献和书法作品的扫描件 [6][?]

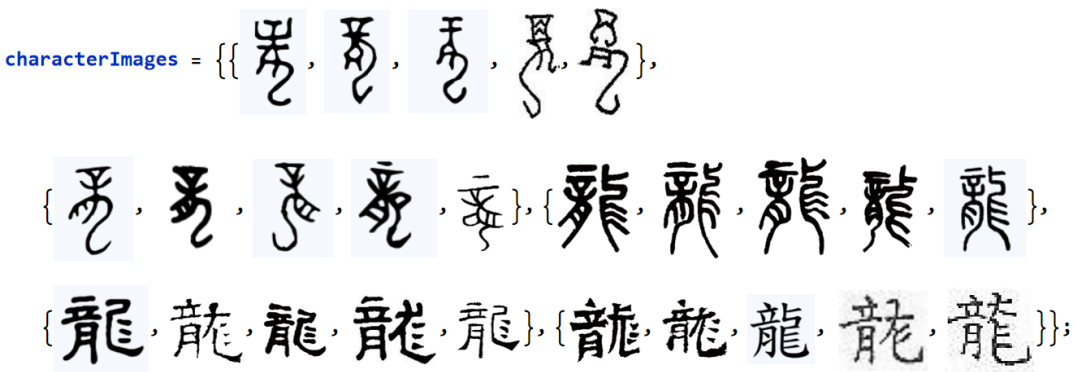


图 3: 五种字体“龍”字训练样本

为了使这些图像数据适用于机器学习模型，我们进行了一系列的数字化处理。包括图像的去噪、归一化、二值化以及形状特征的提取，以确保数据的质量和一致性。

4.2 机器学习模型训练

经过训练的机器学习模型能够准确地对“龍”字的不同历史形态进行分类和聚类。借助机器学习所擅长的特征提取功能，通过训练样本数据集，机器学习训练得到一个特征提取函数，每张“龍”字图片被特征提取函数映射到一个高维特征向量上，同时不在训练样本集中的文字也同样可以被特征提取函数映射为特征向量，如图 (4)，这允许我们寻找与某个未知文字（不在样本数据集中的称为“未知”，反之为“已知”）最接近的已知文字，如图 (5)，也是后续识别未知文字的基础。

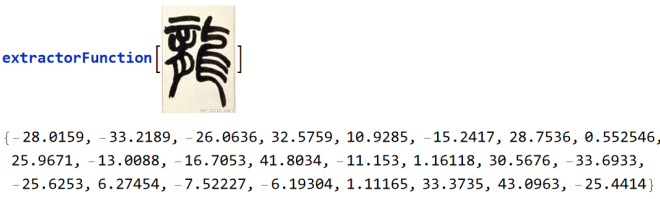


图 4: 将未知文字映射为特征向量

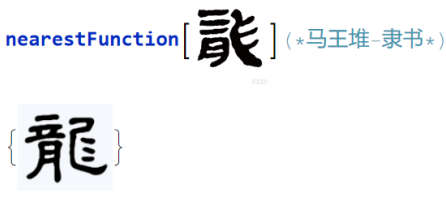


图 5: 寻找与未知文字接近的已知文字

在被压缩后的特征空间二维平面上，模型通过聚类算法对“龍”字的不同历史形态进行聚类，以识别和区分不同的书写风格，将相似的相似的“龍”字形态聚集在一起，如图 (6)，揭示了“龍”字演化的内在规律，这种方法很好反映了历史的连续性。通过可视化技术，我们可以直观地观察到“龍”字从甲骨文到楷书的演化路径，图中显示为从下到上的演化方向，以及不同历史时期“龍”字形态的分布情况。

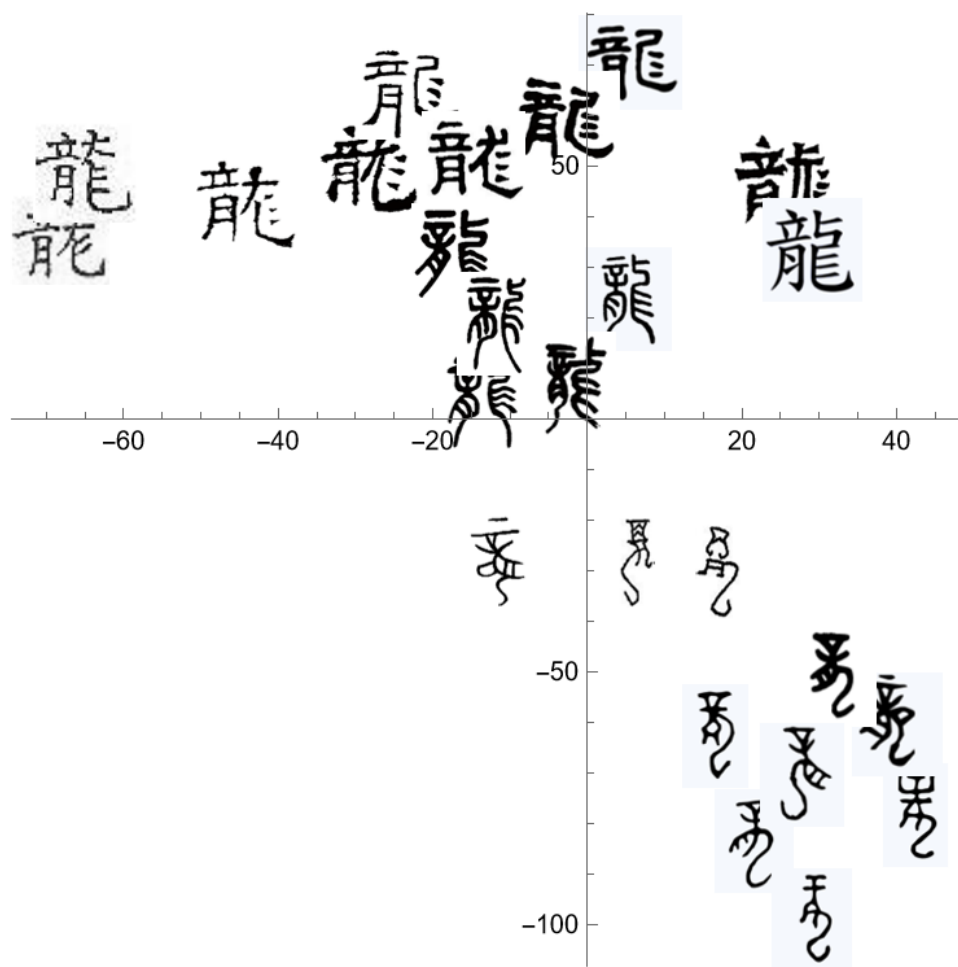



图 6: “龍”字聚类显示其演化趋势

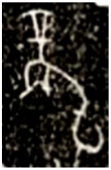
此外，该模型还能够识别和分类未在样本中的“龍”字所对应的字体，将其与训练样本中的图文字对比寻找其可能的分类，如图 (7)，选择了两个训练样本集之外的“龍”字，分别属于战国文字 (介于周代金文和篆文) 和甲骨文，AI 模型都成功识别了其对应字体，并在甲骨文的例子中，给出了未知文字可能分属于各类字体的概率。我们开发的这个原创性分析框架为考古工作提供了一种新的工具，以辨别未知文物上“龍”字的字体，并一定程度上反映了对应文物年代。

4.3 分析结果

通过 AI 框架得到的“龍”字演化分析结果，我们成功地将不同历史时期的“龍”字样本进行了分类和聚类。二维平面聚类可视化的结果清晰地展示了“龍”字从甲骨文到楷书的演化路径，揭示了不同历史时期书写风格的变迁。此外，该框架还能够有效辨别未在样本中的“龍”字图片的字体，为未知文物的年代鉴定提供了新的技术手段，即使不是专业的考古工作者也能应用这个框架做出初步的历史判断。

`characterClassifier` [] (*战国文字*)

篆文

`ReverseSort@characterClassifier` [ , "Probabilities"]
|反规范排序

(*《甲骨文合集》27021 (《国博》203)*)

<| 甲骨文 → 0.996989, 隶书 → 0.00301079,

篆文 → 1.14565×10^{-8} , 周代金文 → 1.33252×10^{-17} , 楷书 → 1.43602×10^{-31} |>

图 7: AI 识别未知“龍”字

然而,我们必须认识到,尽管这一模型在特定任务上表现出色,但它仍然存在局限性。例如,对于形态高度相似或受损严重的“龍”字样本,模型的识别准确率可能会下降。此外,模型的泛化能力也有待进一步提高,以便更好地适应不同的考古场景和文物。

总的来说,通过对“龍”字演化的 AI 辅助分析的具体创新案例,展示了机器学习技术在考古学研究中的应用潜力。这一研究不仅直观展现了汉字演化的趋势,也为其他具有历史演化特性的对象提供了分析框架和方法。随着技术的不断发展,我们期待 AI 技术能够在考古学领域发挥更大的作用,为文化遗产的保护和研究做出更多的贡献。

5 总结与展望

本次课程项目研究的成果主要体现在以下几个方面:首先,展示了 AI 在图像和文本的识别、分析领域的初步应用,指出了 AI 辅助考古研究的可能道路。其次,建立了一个 AI 辅助的汉字演化分析框架,为考古学研究提供了新的视角和方法。本次研究展示了 AI 技术在考古学领域的应用潜力,为未来的研究提供了宝贵的经验和启示。

展望未来, AI 在考古学领域的应用前景广阔。随着技术的进步,可以预见, AI 将被更广泛地应用于文物的识别、分类和年代测定,极大地提高考古工作的效率和准确性。同时, AI 技术也将推动考古学研究方法的创新,为解决复杂的考古问题提供新的途径。此外, AI 技术在文化遗产保护、虚拟修复和数字化展示等方面的应用也将得到进一步的发展。然而,也必须注意到, AI 技术在考古学领域的应用还面临着数据质量、模型泛化能力、伦理和法律等挑战。因此,未来的研究需要在提高 AI 技术的性能的同时,也要考虑这些因素,以确保 AI 技术在考古学领域的健康发展。

随着 AI 技术的不断发展和完善,我们有理由相信,它将成为考古学领域的重要工具,为文化遗产的保护和研究做出更大的贡献。通过跨学科的合作和创新, AI 技术将开启考古学研究的新篇章。

参考文献

- [1] Argyro Argyrou and Athos Agapiou. A review of artificial intelligence and remote sensing for archaeological research. *Remote Sensing*, 14(23):6000, 2022.

- [2] Archith Iyer and Manoj Franklin. Ai-powered archaeology: Determining the origin culture of various ancient artifacts using machine learning. *Journal of Student Research*, 11(1), 2022.
- [3] Athos Agapiou and Vasiliki Lysandrou. Interacting with the artificial intelligence (ai) language model chatgpt: a synopsis of earth observation and remote sensing in archaeology. *Heritage*, 6(5):4072–4085, 2023.
- [4] 黄盛璋. 云梦秦简辨正. 考古学报, 1:1–26, 1979.
- [5] 季勋. 云梦睡虎地秦简概述. 文物, 5:1–6, 1976.
- [6] 徐中舒. 甲骨文字典. 四川辞书出版社, 2021.

课程作业检测系统
文本复制检测报告单(简洁)

No: BC2024060715104715269442266

检测时间: 2024-06-07 15:10:47

篇名: AI Meets Archaeology: 辅助分析未知文物属性与历史演化. pdf

作者: 尹琪钦 (211870080)

授课教师: 王晓琪, 张良仁, 张学锋, 赵星宇, 殷洁

检测机构: 南京大学

文件名: AI Meets Archaeology: 辅助分析未知文物属性与历史演化. pdf

检测系统: 课程作业检测系统 (课程学习全过程综合培养平台)

检测类型: 课程作业

检测范围: 中国学术期刊网络出版总库
中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
中国重要会议论文全文数据库
中国重要报纸全文数据库
中国专利全文数据库
图书资源
优先出版文献库
互联网资源(包含贴吧等论坛资源)
英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
港澳台学术文献库
互联网文档资源
源代码库
CNKI大成编客-原创作品库
大学生论文联合比库
课程作业联合比库
机构自建比库

时间范围: 1915-01-01至2024-06-07

检测结果

总文字复制比: 0%

跨语言检测结果: -

去除引用文献复制比: 0%

去除本人文献复制比: 0%

单篇最大文字复制比: 0%

重复字数:	[0]	总段落数:	[1]
总字数:	[6730]	疑似段落数:	[0]
单篇最大重复字数:	[0]	前部重合字数:	[0]
疑似段落最大重合字数:	[0]	后部重合字数:	[0]
疑似段落最小重合字数:	[0]		

指标: ☐ 疑似剽窃观点 ☐ 疑似剽窃文字表述 ☐ 疑似整体剽窃 ☐ 过度引用

相似表格: 0 相似公式: 没有公式 疑似文字的图片: 0

1. AI Meets Archaeology: 辅助分析未知文物属性与历史演化. pdf

总字数: 6730

相似文献列表

说明: 1. 总文字复制比: 被检测论文总重合字数在总字数中所占的比例

2. 去除引用文献复制比: 去除系统识别为引用的文献后, 计算出来的重合字数在总字数中所占的比例

3. 去除本人文献复制比: 去除作者本人文献后, 计算出来的重合字数在总字数中所占的比例

4. 单篇最大文字复制比: 被检测文献与所有相似文献比对后, 重合字数占总字数的比例最大的那一篇文献的文字复制比

5. 复制比: 按照“四舍五入”规则, 保留1位小数

6. 指标是由系统根据《学术论文不端行为的界定标准》自动生成的

7. 红色文字表示文字复制部分; 绿色文字表示引用部分(包括系统自动识别为引用的部分); 棕灰色文字表示系统依据作者姓名识别的本人其他文献部分

8. 本报告单仅对您所选择的比对时间范围、资源范围内的检测结果负责



 amlc@cnki.net

 <https://check.cnki.net/>