国际期刊《控制论与信息学》（IJCI）第12卷第2期，2023年4月

使用自然语言处理进行特征匹配

Muhammad Danial Khilji

数据科学系，Choreograph

摘要

特征匹配是匹配不同数据集的基本步骤。本文提出了一种新的混合模型：将基于预训练自然语言处理（NLP）模型 BERT 的模型与基于杰卡德相似度的统计模型并行使用，以衡量两个不同数据集中特征列表之间的相似度。这减少了搜索相关性或手动将一个数据集中的每个特征与另一个数据集中的特征进行匹配所需的时间。

关键词

BERT、余弦相似度、特征、匹配、语义相似度、相似度

1. 引言

特征匹配是许多数据处理流程中的第一步，在数据融合过程中发挥着至关重要的作用。数据融合是指利用一个数据集填充另一个数据集的过程。如果手动进行模式匹配，会非常耗时，尤其是在数据宽度很大的情况下。本文构建了一种基于语言的方法来进行特征匹配。当待匹配的数据集在性质上存在异质性时，这种方法尤为适用。

本文第 2 节介绍了模式匹配方面已发表的相关研究，并重点说明机器学习的进步如何改变了模式匹配的发展方向。文中还全面回顾了在相似度匹配中表现出色、也有望用于模式匹配的 BERT 相关模型。第 3 节深入解释了以 Python 语言设计并实现的架构。首先，将数据整理为适当的结构，并进行文本清理，使其更易于理解，以便在第 3.1 至 3.3 节中进行相似度度量。第 3.4 节详细介绍了用于衡量两个特征名称之间相似度的两种不同方法。随后，第 3.5 节说明了如何将结果整理成更易于理解和获取的形式。第 4 节讨论了结果，并说明结果为何会因所用数据类型而异。第 5 节总结了方法，并探讨了未来的改进方向。

2. 文献综述

鉴于特征匹配所需应用的广泛性，目前关于特征匹配的研究仍相当有限。《不透明列名的模式匹配》是最早研究两个不同数据集之间特征匹配的论文之一。该文还介绍了一种匹配列名和列值的方法。它着重探讨无解释结构匹配方法，即逐个元素匹配列值，与结构匹配形成对比 [1]。在此之前，E. Rahm 已开展了深入研究，指出

David C. Wyld 等（编）：DBML、CITE、IOTBC、EDUPAN、NLPAI -2023，
第251-260页，2023年。IJCI – 2023

数字对象标识符:
10.5121/ijci.2023.120218

模式匹配方法的全貌 [2]。早期方法的主要局限在于，它们专为数据库设计 [3]。目前流行的模式匹配模型是 COMA，它采用图网络方法，在多个数据集之间寻找相似特征 [4]。近期有一项研究探讨了采用图网络方法的模式匹配方法，并将 COMA 的性能与其他先进的机器学习模型进行了比较 [5]。

已有研究探讨模式匹配方法 [6]。大多数方法难以理解两个句子之间的语义相似性。本文简要分别介绍基于模式、基于实例和基于机器学习的方法。尽管基于机器学习的方法通常优于其他类型的方法，但效果也取决于具体应用场景。本文还提出了一种结合上述方法的新型端到端模式匹配方法，并将其与 Google 通用句子编码器（USE）模型进行比较 [7]。

随着机器学习领域的最新进展，Google AI 语言团队开发了双向变换器编码器表示（BERT）模型，该模型在多种自然语言处理任务中取得了最先进的结果 [8]。此后，BERT 模型已被用于从多个不同方面改进实体匹配 [9]。特征匹配的性质决定了它可被归类为实体匹配的一个分支。

有一些关于实体匹配的实用研究论文，也可用于特征匹配。《基于变换器架构的实体匹配》是最早采用 BERT [9], XLNet [10], RoBERTa [11] 和 DistilBERT [12] ，并将它们与经典深度学习模型进行比较的论文。研究结果证明，变换器架构的表现胜过所有这些模型。另一项较早将 BERT 模型用于相似度匹配的研究，应用场景是匹配不同零售商的产品 [13]。该研究采用了更精简的先进 BERT 模型变体——BERT [8] 和 DistilBERT [12], ，利用来自不同电子商务网站的数据集进行产品匹配。研究人员在标准 BERT 模型上微调，并添加一个额外层，将其命名为 eComBERT。随后，研究比较了所有模型，结果显示，eComBERT 在不同数据集上均取得了显著提升的效果。

研究人员成功开发了一种专用于实体消歧的架构，名为 BERT-ER [14]。该架构通过增强对齐网络并引入“阻断解码器”模块，延后成对交互过程，同时剔除明显不相似的匹配项。结果表明，该架构以更短的运行时间取得了更高的 F1 分数。

紧接着，另一篇论文发表了，该论文展示了一种不同的方法，采用二元匹配和多类别分类 [9]。这种方法使训练过程中的最终决策由匹配／不匹配判断以及对每个实体描述的实体标识符进行预测共同构成，因此称为联合 BERT。除了使用神经网络模型之外，也存在其他方法，例如 SERIMI 提出的基于类别的匹配技术 [15, 16]。

3. 方法

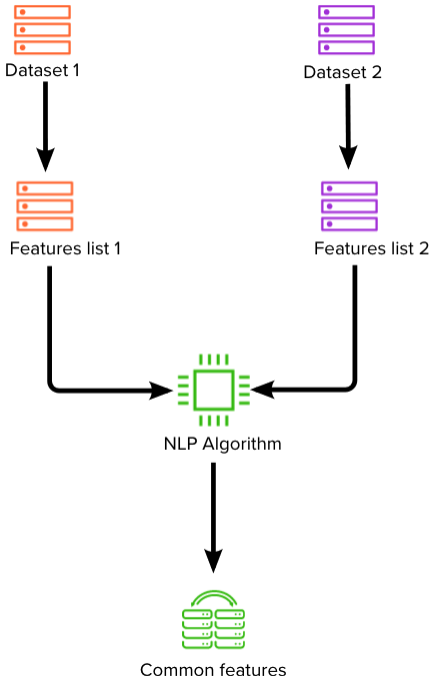


图1：主要工作流程

3.1. 预处理

首先，将两个不同数据集中的特征分别保存到两个独立的列表中。目前可以假设使用任意数据集，采用示例数据集得到的结果将在第 4 节讨论。为提供更多上下文并最终改善相似度结果，将额外说明特征的信息与两个特征列表拼接起来。这些额外信息是对每个特征的说明。或者，也可以将唯一的特征值与特征名称拼接起来。为移除不需要的数据，接下来对这两个特征列表进行预处理，清理数据，删除空记录以及可能不需要的列

3.2. 自然语言处理

自然语言处理旨在将文本转换为相似的结构。第一步是将文本转换为小写。然后移除可能导致结果质量较低的特殊字符，例如 & % $ 等。接着对每个句子进行分词，并将每个句子生成的词元列表分别保存到另一个列表中的不同位置。这样，最终列表中就包含多个已分词句子的列表。在此阶段还会移除停用词，例如“is、a、are、as”。最后一步是词形还原，将单词形式转换为现在时形式，例如将“pointing”转换为“point”，将“swimming”转换为“swim” [17]。

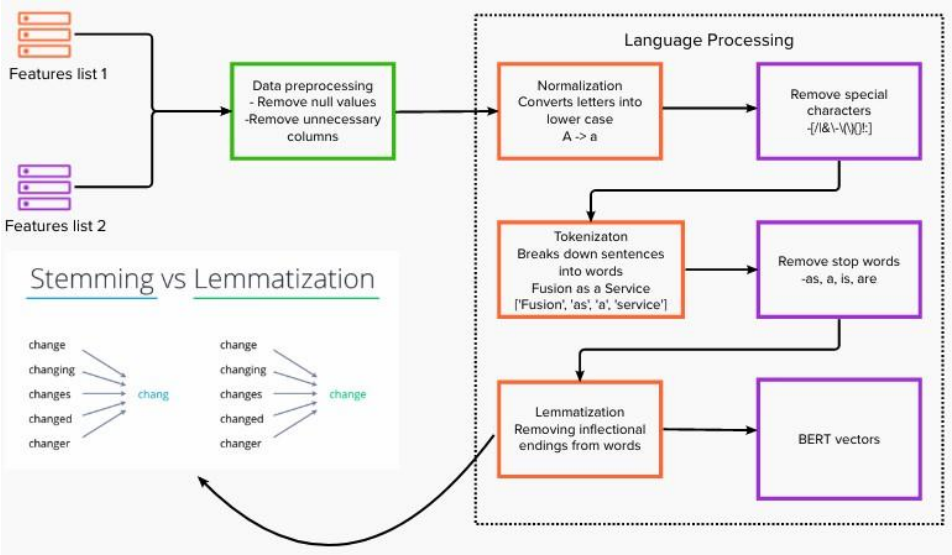


图2：数据输入和预处理阶段

3.3. 嵌入

单词完成词形还原后，使用预训练的标准 BERT 模型生成其嵌入 [8]。最初也使用过 JointBERT，但其表现不佳，因此被弃用 [9]。

3.4. 相似度度量

本文并行采用了两种方法：分别使用 BERT嵌入计算余弦相似度，以及将词形还原后的词元作为输入计算杰卡德相似度。采用这两种相似度方法的主要目的是：如果一种方法对某个匹配结果赋予的权重较低，另一种方法可能会赋予其较高的权重，从而产生相互矛盾的相似度结果。尽管 BERT 模型本身表现良好，但杰卡德相似度为整体模型提供了更大的灵活性，在使用特殊符号或行业术语的情况下也能发挥作用。杰卡德相似度的计算方式是：对于每个特征匹配，取两个不同特征列表中记录的词形还原后的词元集合，并计算它们之间的共同词语 [18]。例如：

特征列表 1 中的特征：AI 是我们的朋友，并且一直都是我们的
朋友。特征列表 2 中的特征：AI 和人类一直都是我们的朋友。

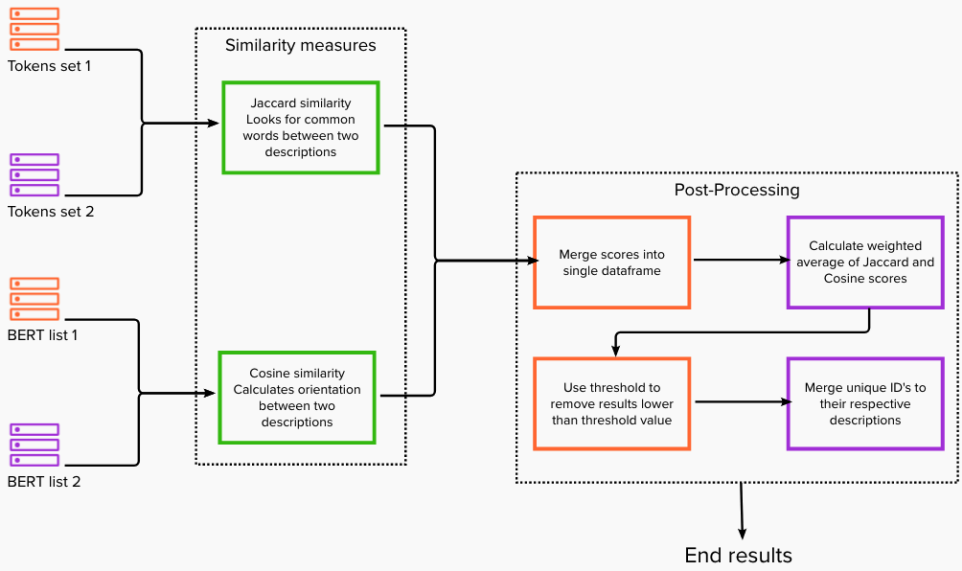


图3：预处理阶段输出至相似度度量和后处理

$$J(A,B) = \frac{|A \cap B|}{A \cup B}$$

其中，$|A \cap B|$ 是共同词元的总数，$|A \cup B|$ 是两个特征列表中所有词元的总数。A 是第一个特征列表中单个特征的词元集合，B 是第二个特征列表中单个特征的词元集合。结果如下图 4 所示。

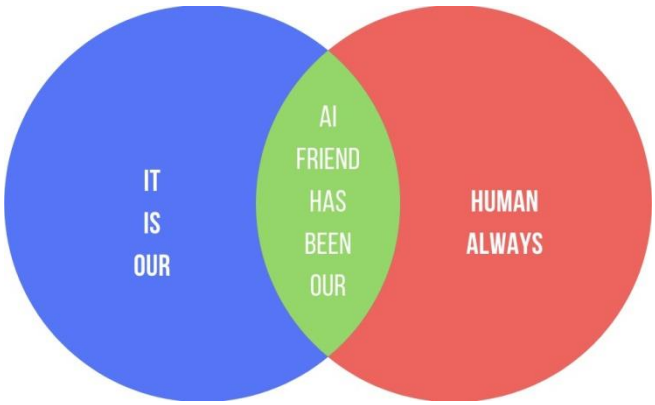


图4：维恩图，展示杰卡德相似度识别出的共同词语（绿色区域）

BERT 是一种专为自然语言处理相关任务设计并训练的神经网络架构。它能够理解含义相近的词语，并为意义非常相似的词语生成相似的数值嵌入 [8]。该模型的主要改进在于能够双向理解文本，而之前的模型只能从单一方向理解文本。余弦相似度以两个特征列表 [19] 中的 BERT 向量作为输入，然后将所有这些特征绘制在 n维空间中，使关系密切的向量彼此之间的夹角较小。它会计算每两个向量之间的方向关系，并按照以下公式计算相似度，结果如下面的图5所示：

$$Similarity = cos\theta = \frac{A.B}{||A||||B||}$$

Where A 和 B 分别是第一组和第二组特征中各个特征的 BERT 嵌入 ure 列表。

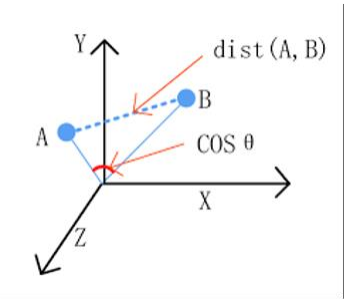


图5：余弦相似度。注意，dist(A,B) 在上面的公式中为 ‖A‖‖B‖ [20]

3.5. 后处理

测量相似度后，会将其转换为数据框，然后根据原始特征名称进行合并。接着计算杰卡德得分和余弦得分的加权平均值。权重由人工指定，但未来可以通过使用逻辑进行改进。在该算法中，提供了三种选择：

$$Cosine = 1 - Jaccard$$

选项 1：余弦相似度 = 0.7，杰卡德 = 0.3（默认）选项 2：余弦相似度= 杰卡德 = 0.5 选项 3：余弦相似度=0.3，杰卡德 = 0.7

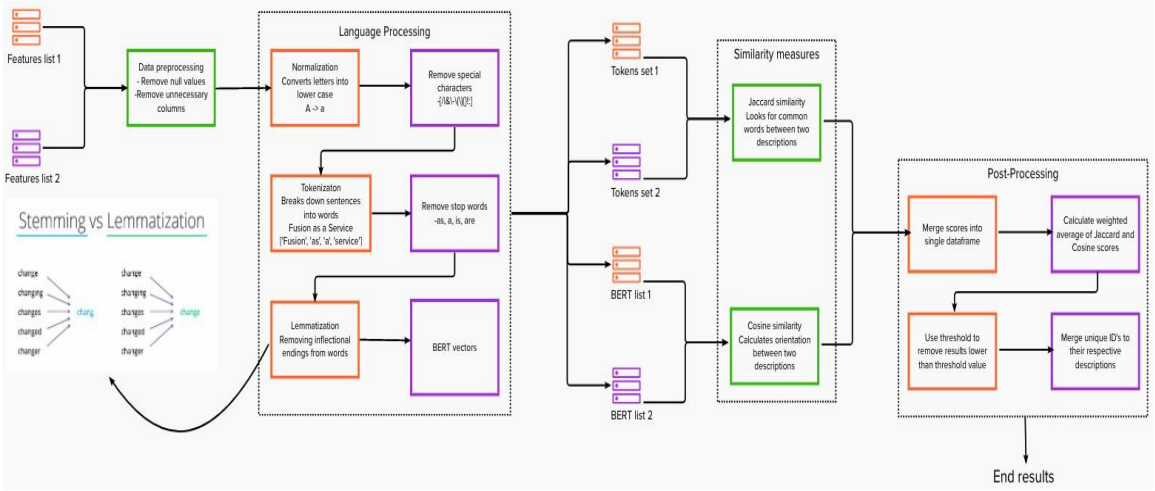
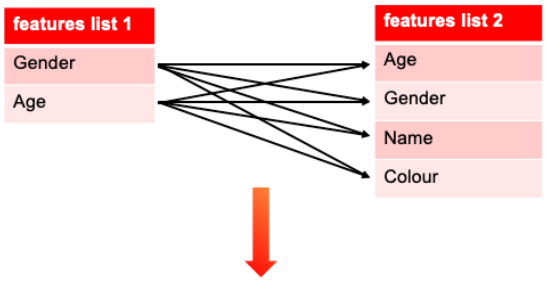


图6：端到端工作流详解

可能会出现许多得分很低的匹配项，甚至完全没有匹配项，因此会使用一个默认阈值。经观察，0.7 的阈值效果最佳，不过用户可以

在 0 到 1 的范围内自行输入阈值。得分高于 0.7 的匹配项通常最为接近，因此默认使用该值，但用户也可以使用较低的阈值来查看较弱的匹配项。这样，所有得分低于阈值的匹配项都会被删除。最后一步是可选的；如果特征具有唯一 ID，也可以在此阶段将其合并。

对于这两种方法，都会将一个列表中的所有特征与另一个列表中的所有特征进行比较，如下方图7所示。



features list 1	features list 2	score
Gender	Age	0
Gender	Gender	1
Gender	Name	0
Gender	Colour	0
Age	Age	1
Age	Gender	0
Age	Name	0
Age	Colour	0

图7：特征匹配示例

4. 结果

大多数结果是基于无法公开的数据生成的，因此，为了演示本文，使用了公开数据集。生成了两组测试结果：一组匹配 IMDB 与 Netflix 数据，另一组匹配珀斯与金县的房屋销售数据，这些数据下载自 Kaggle 网站 [2, 21–23]。

表1：珀斯与金县房屋销售数据之间的特征匹配（阈值 = 0.7）[2,23]

珀斯都会区 澳大利亚特征	美国金县特征	加权平均分
价格	价格	1
卧室	卧室	1
卧室	lat	0.699999988
价格	lōng	0.699999988
郊区	居住面积（平方英尺）15	0.699999928
土地面积	建造年份	0.643930674
郊区	条件	0.595897615

表 2：IMDB 与 Netflix 数据之间的特征匹配（阈值 = 0.7）[21, 22]

IMDB 特征	Netflix 特征	加权平均分
导演	导演	0.99999999
上映年份	上映年份	0.73863158
剧集名称	标题	0.712217
IMDb评分	评分	0.65647271
概述	描述	0.65626079
概述	列于	0.63430017
概述	type	0.62954128

在表1中，排名前两位的匹配项得分均为1，因为 BERT 将它们都理解为相同的词，而且它们都是拼写相同的常见词，因此杰卡德相似度也得出1。其余匹配项并非最佳匹配，因为两个数据集不包含相同类型的特征。在这种情况下，BERT 会尝试匹配含义可能相近的词，例如“郊区”与“条件”。虽然这两个词的含义完全不同，但 BERT 将它们理解为相似词。

表 2 展示了截然不同的表现。尽管排名前两位的匹配结果完全一致，但 BERT 却将它们识别为不同项，因此平均得分较低。相比之下，杰卡德相似度对这两项匹配的评分为 1，因此最终加权平均分有所提高。这类情况正是使用简单相似度方法而非复杂神经网络模型的主要原因。我们最初也使用了 JointBERT，但它在我们的应用场景中表现不佳。可以考虑训练自己的模型，但与标准 BERT 模型相比，JointBERT 需要相对更多的数据才能完成学习 [17]。

本研究使用了预训练模型，但也可以针对特定数据训练模型。不过，这类模型更容易出现偏差等问题，可通过选择偏差探索和去偏方法加以改善 [24]。

5. 结论

本文展示了如何将 BERT 与简单的统计方法结合，用于查找含义相似的特征名称。这一过程可用于数据融合等场景。也可以并行使用两种以上的相似度度量，例如 Monge-Elkan 和软 TF-IDF 等 [25–27]。尽管采用更多方法会得到更好的结果，但算法的计算成本也会随之增加。因此，我们仅使用了两种最符合需求的方法。一般建议将余弦相似度与其他能取得最佳结果的方法结合使用。影响结果的主要因素是最适合问题的相似度类型；例如，本问题属于基于词元和基于序列的相似度类别 [28]。

致谢

我谨向 Choreograph 的数据科学部门表示感谢，感谢他们给予我研究和完善这一方法的机会。当时的经理 Victor Moron Tejero 确保我拥有开展研究所需的一切资源。融合即服务（FaaS）团队让我更深入地了解了这个问题，并通过持续反馈帮助我改进设计。Giuseppe De Vitis 为我提供了不同的视角，并说服我进一步完善架构。作为我目前的经理，Nicholas Iles 在需要时帮助我联系到合适的人。Georgios Giasemidis 在将这项研究发表出来的过程中给予了我很大帮助，指导我寻找

合适的期刊，并教我如何构建研究论文。我衷心感谢你们每一位。

参考文献

[1] J. Kang 和 J. F. Naughton，《关于不透明列名和数据值的模式匹配》，载于 2003 年 ACM SIGMOD 数据管理国际会议论文集，第 205–216 页，2003 年。[2] Harlfoxem，《美国金县房屋销售》。[3] Z. Bellahsene、A. Bonifati、F. Duchateau 和 Y. Velegrakis，《关于模式匹配与映射的评估》，载于《模式匹配与映射》，第 253–291 页，施普林格，2011 年。[4] H.-H. Do 和 E. Rahm，《COMA——一种灵活组合模式匹配方法的系统》，载于*VLDB'02*：第 28 届超大型数据库国际会议论文集，第 61–621 页，Elsevier，2002 年。[5] F. Rahutomo、T. Kitasuka 和 M. Aritsugi，《语义余弦相似度》，载于第七届国际科学与技术学生会议 ICAST，卷 4，第 1 页，2012 年。[6] B. H¨attasch、M. Truong-Ngoc、A. Schmidt 和 C. Binnig，《这就是匹配：一种使用嵌入进行模式匹配的两步方法》，arXiv 预印本 arXiv:2203.04366，2022 年。[7] D. Cer、Y. Yang、S.-y. Kong、N. Hua、N. Limtiaco、R. S. John、N. Constant、[8] M. Guajardo-Cespedes、S. Yuan、C. Tar 等，《通用句子编码器》，arXiv 预印本 arXiv:1803.11175，2018 年。[9] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，arXiv 预印本 arXiv:1810.04805，2018 年。[10] M. Paganelli、F. D. Buono、A. Baraldi 和 F. Guerra，《分析 BERT 在实体匹配中的表现》，《VLDB基金会会刊》，第 15 卷，第 8 期，第 1726–1738 页，2022 年。[11] L. Wang、Z. Chen 和 J. Wu，《车载自组网中基于车辆移动关联的数据转发机会路由》，《信息》，第 8 卷，第 4 期，第 140 页，2017 年。[12] Y. Liu 、 M. Ott 、 N. Goyal 、 J. Du 、 M. Joshi 、 D. Chen 、 O. Levy 、 M. Lewis、L. Zettlemoyer 和 V. Stoyanov，《RoBERTa：一种稳健优化的 BERT 预训练方法》，arXiv 预印本 arXiv:19 07 11692，2019 年。[13] S. Araujo、D. T. Tran、A. P. de Vries 和 D. Schwabe，《SERIMI：用于异构数据集间实例匹配的基于类别的匹配》，《IEEE 知识与数据工程汇刊》，第 27 卷，第 5 期，第 1397–1440 页，2014 年。[14] V. Sanh、L. Debut、J. Chaumond 和 T. Wolf，《DistilBERT：BERT 的蒸馏版本，更小、更快、更便宜、更轻量》，arXiv 预印本 arXiv:1910.01108，2019 年。[15] B. Li、Y. Miao、Y. Wang、Y. Sun 和 W. Wang，《提升基于 BERT 的实体消歧的效率与效果》，载于 AAAI 人工智能会议论文集，第 35 卷，第 13226–13233 页，2021 年。[16] D. Rodrigues 和 A. d. Silva，《机器学习技术在模式匹配网络问题中的研究》，《巴西计算机学会期刊》，第 27 卷，第 1 期，第 1–29 页，2021 年。[17] G. Papadakis 和 W. Nejdl，《异构信息空间中的高效实体消歧方法》，载于 2011 年 IEEE 第 27 届数据工程国际会议研讨会论文集，第 304–307 页，IEEE，2011 年。[18] R. Peeters 和 C. Bizer，《面向实体匹配的 BERT 双目标微调》，[19] 《VLDB基金会会刊》，第 14 卷，第 10 期，第 1913–1921 页，2021 年。[20] S. Niwattanakul、J. Singthongchai、E. Naenudorn 和 S. Wanapu，《使用杰卡德系数进行关键词相似度计算》，载于国际工程师与计算机科学家多学科会议论文集，第 1 卷，第 380–384 页，2013 年。[21] E. Rahm 和 P. A. Bernstein，《自动模式匹配方法综述》，《VLDB期刊》，第 10 卷，第 4 期，第 334–350 页，2001 年。[22] J. Tracz、P. I. W´ojcik、K. Jasinska-Kobus、R. Belluzzo、R. Mroczkowski 和 I. Gawlik，《基于 BERT 的产品匹配相似度学习》，载于电子商务自然语言处理研讨会论文集，第 66–75 页，2020 年。[23] H. Shankhdhar，《IMDB 电影数据集》。[24] S. Bhansal，《Netflix 电影和电视节目》。[25] M. S. Zainal，《珀斯房价》。

[26] G. Zhang、B. Bai、J. Liang、K. Bai、S. Chang、M. Yu、C. Zhu 和 T. Zhao，《自然语言句子匹配数据集中的选择偏差探索与去偏方法》，arXiv 预印本 arXiv:1905.06221，2019 年。[27] S. Jimenez、C. Becerra、A. Gelbukh 和 F. Gonzalez，《用于近似文本字符串比较的广义 Mongue-Elkan 方法》，载于国际智能文本处理与计算语言学会议论文集，第 559–570 页，施普林格，2009 年。[28] D. Eder 和 F. Gregg，《soft-TFIDF》。[29] M. del Pilar Angeles 和 A. E. Gamez，《汉明距离、Jaro 和 Monge–Elkan 方法比较》，DBKDA 2015，第 73 页，2015 年。[30] 《文本距离度量比较》。

作者

Muhammad Danial Khilji 是一名数据科学家，拥有不到一年的相关领域工作经验。他获得了应用数据科学硕士学位和电气工程学士学位，主修人工智能和机器学习。

