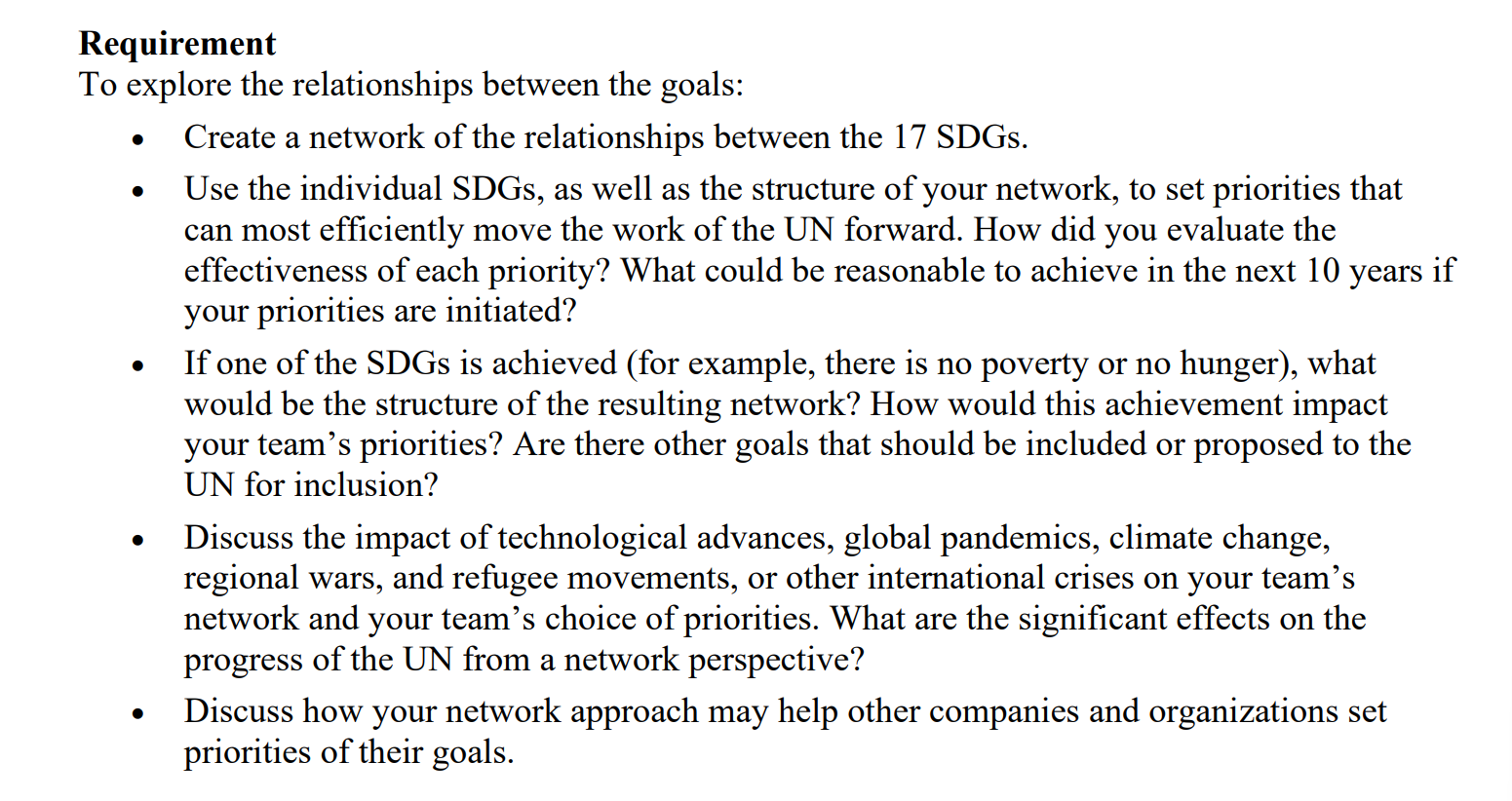
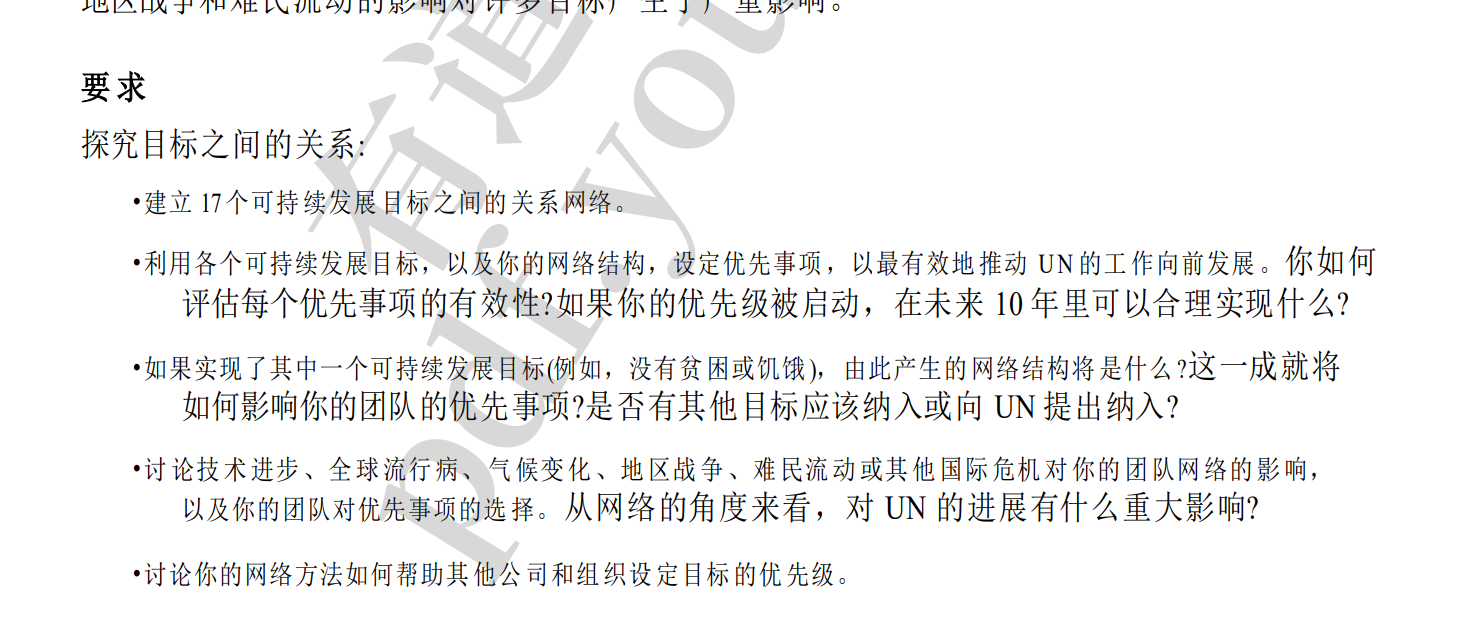
美赛草稿本





**问题重述**

2015年9月，193位世界领导人商定了17项全球可持续发展目标。如果这些目标得以实现，将意味着到2030年结束极端贫困、不平等和气候变化[1]。《2030年议程》是为实现这些目标而通过的框架和总体战略文件[7]。虽然2030议程是针对国家的，但它也可以帮助组织确定相关的可持续性方面。

因此，本文中的研究通过以下研究问题探讨了重要性分析在解决市政当局提出的挑战方面的有用性：当局如何将可持续性纳入其业务治理，哪些可持续性方面被列为优先事项？先前关于重要性分析的研究是否有助于确定可持续性方面的优先次序？

**SDG解释：**

<https://www.un.org/zh/70001/page/180631>

**备注:**

**Q1找个网络实现**

**Q2建立模型**

**Q3举例One SDG**

**Q4一个一个国际影响讨论过去**

**Q5模型应用到公司组织上面**

**模型假设：**

1. 假设我们收集到的数据属实。
2. 生存是文明的第一需要，零饥饿会被赋予很高的初始权重
3. 17个SDG之间是相互之间存在一定的关联性即依赖关系和相互影响，包括直接联系和间接联系。例如如果摆脱贫困就能实现优质教育，反过来如果实现优质教育可以摆脱贫困。这种关联性在某两者之间可能会很弱，其他两者之间可能会很强。
4. 不同需求之间遵循具有一定的次序关系，17个SDGs中有具有最大影响的目标，也有最有效实现的目标，还有优先考虑需要先实现的目标，以确保其他目标的可持续性发展。例如只有满足了生存需求才开始有尊严需求，但是这个次序关系在同级别的SDG面前可能并不是那么严格
5. 如果实现了其中一个SDG，那么整个网络结构将发生变化，与该SDG相关的指标将会受到改善，而其他SDGs相关的指标也可能会发生一些变化，则需要重新评估网络中各个SDG的得分和优先级。如果某个SDG的得分在实现另一个SDG后发生重大变化，可能需要重新评估其优先级。

输出：17个SDG的实现顺序

目标： efficiently move the work of the UN forward

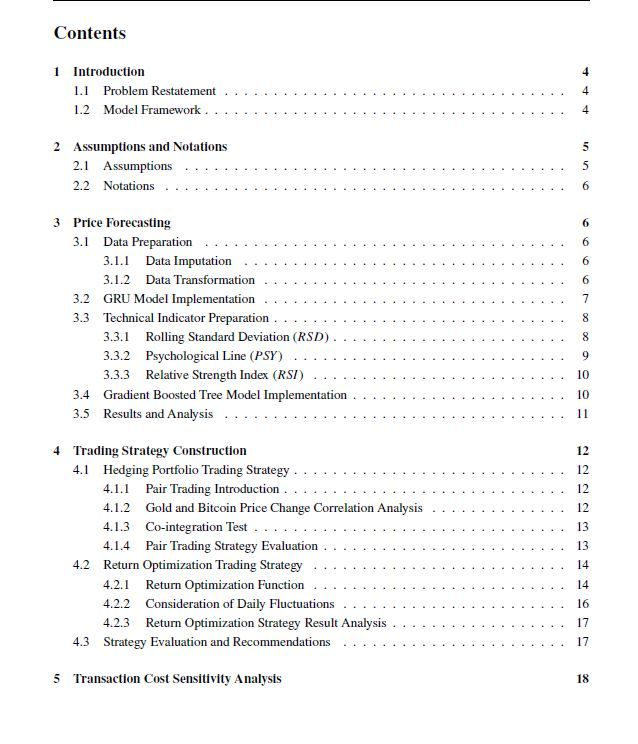
评价指标、衡量指标：促进世界发展的一个分数

**模型参数：**

(i)发生率（O），可根据相关利益相关者的百分比确定;（ii）被发现的可能性（D），指利益相关者的关注程度;（iii）严重程度（S），可根据问题对战略沟通目标的影响进行量化。

**论文框架**

原文链接：<https://www.zhihu.com/question/520508243/answer/2475682953>



#### Tqq用：

prob restatement(算是初稿，改写力度没有很大，需要改的话说一下）

The Sustainable Development Goals (SDGs) are a set of 17 goals established by the United Nations (UN) to create a sustainable future for all. These goals include no poverty, zero hunger, good health and well-being, quality education, gender equality, clean water and sanitation, affordable and clean energy, decent work and economic growth, industry innovation and infrastructure, reduced inequalities, sustainable cities and communities, responsible consumption and production, climate action, life below water, life on land, peace, justice and strong institutions, and partnerships for the goals. These 17 SDGs are interrelated, and one goal's achievement can affect others' achievement. Each block symbolizes one of the goals, and removing one block can destabilize the entire system. For example, the COVID-19 pandemic has impeded progress toward multiple objectives, including SDG 3 (Good Health and Well-being) and SDG 4(Quality Education). Also, the epidemic has exacerbated existing inequality and poverty (SDG 1). Reaching the SDGs is a dynamic process that needs balancing national and international goals, limited resources, and other elements such as technology developments, pandemics, climate change, regional conflicts, and refugee movements. Many stakeholders need to collaborate to overcome these hurdles.

The following content will examine the connections between the 17 SDGs and develop a network illustrating their relationships. Analyze the efficiency of each priority and provide a realistic estimate of what can be accomplished in the next decade with their implementation. Identify any other goals that should be included or proposed to the United Nations for inclusion, taking into account the potential changes in the network structure resulting from achieving an SDG. Global challenges affect the team's network and priorities, including technology, climate change, local wars, refugee flows, and pandemics. We will investigate these crises' network effects on UN development. Explain how our network model may assist other organizations in establishing priorities for their goals. Prioritize the UN's efforts by assessing the SDGs and network structure. Create a system to evaluate each priority. Setting these priorities will help the UN accomplish its goals in 10 years.

paper notes:

for identifying SDG interconnections by applying Natural Language

Processing (NLP) tools and network science techniques

我们寻求解决两个主要问题:(i)联合国关于可持续发展的官方政策话语中产生了哪些相互关联的 可持续发展目标的相互关联和集群?(ii)在关于可持续发展目标的科学论述中是否也发现了类似的相互联系?

在文章的第一部分，我们使用这种组合方法来

(i) 描述和总结围绕可持续发展目标的国际政策讨论，以联合国关于每年实现每个目标取得的进展的官方报告为代表；

(ii) 描绘联合国产生的可持续发展目标相互联系网络

政策话语； 和

(iii) 确定联合国倾向于类似讨论的可持续发展目标集群，暗示它们可以。 由决策者或科学家同时解决。

可持续发展目标相互依存关系网络是第二波研究的主题。 该领域的一项开创性研究通过检查官方描述、具体目标和指标中共享或交叉引用的具体目标，量化了可持续发展目标之间的协同作用程度6。

然而，这项努力因未能考虑到目标之间的权衡和潜在冲突领域而受到批评 5, 19。

我们分析了联合国经济及社会理事会 (ECOSOC) 每年向联合国秘书长提交的关于每个可持续发展目标的所有“进展和信息”(P/I) 报告的语料库，从 2016 年开始（此类报告的第一年） 到 2020 年（最后一份可用报告的年份）36。 这些报告描述了每个可持续发展目标每年取得的全球进展，并提供有关每个目标的其他描述性信息。 所有 85 份报告（17 个 SDG，5 年）均使用标准文本预处理程序进行了清理和清理，并由 SDG 进行了拼接。 生成的 17 个 P/I 文档的长度范围为 1662 到 6743 个标记（平均值 = 3092，标准偏差 = 1363 个标记）。 为了提取显着信息，我们应用了标准的 NLP 文本预处理步骤，包括标记化、词形还原和词性 (POS) 标记。 为了描述语料库内容，我们将 TextRank 算法应用于文档 13、37 的词形化内容。 有关文本预处理和 TextRank 分析结果的更多详细信息，请参见 SI 附录。

为了映射可持续发展目标之间的相互依赖关系，我们需要分布式数字表示， 我们可以计算相似性度量。这是使用文档嵌入模型(doc2vec)实现的， 一种在NLP中使用的技术，用于在继承单词语义 后生成文档的数字表示，这些语义基于训练数据语料库中相似的词汇上下文中的使用。然后使用连续词袋(CBOW) doc2vec模型(5字窗口)300维和250次训练迭代来量化P/I文档之间的 语义重叠38. doc2vec 模型将文档 ID 参数引入 CBOW word2vec 模型（一种浅层神经网络，旨在根据单词搭配生成单词嵌入），以根据文档标识符和更广泛的词汇上下文预测单词。 此信息用于为每个 SDG 的 P/I 文档生成文档嵌入（即数字向量表示）15。 文档嵌入之间的余弦相似度捕捉联合国报告中可持续发展目标之间的话语接近度或重叠度。 这些相似性表示为可持续发展目标的加权网络，每个网络边缘的权重表示联合国关于两个可持续发展目标的报告之间的归一化余弦相似性。 为了识别相似 SDG 的聚类，对 SDG 之间的余弦相似度得分矩阵进行了完整链接聚类分析，并将 Louvain 社区检测算法应用于 SDG 相似度网络 39,40。 当从在 UN P/I 语料库（图 1）上训练的模型切换到在 11.8 GB 的美国新闻报道上训练的模型（图 S4）时，平均余弦相似度增加到 0.94（sd = 0.02），但没有 可持续发展目标之间的话语相似性模式存在实质性差异。 话语相似性的结果对于 doc2vec 模型的不同规范和其他 NLP 方法（例如潜在语义分析（请参阅 SI 附录））也是稳健的。

为了评估可持续发展目标联合国报告中的重叠与科学文献中可持续发展目标之间的整合之间的一致性，我们收集了所有经过同行评审的科学文章（2000 年至 2020 年之间）的元数据，这些文章涉及三对中具有最强余弦的六个可持续发展目标中的至少一个。 联合国报告（可持续发展目标 5 和 16、10 和 17、14 和 15）中的相似性分数。 这些出版物是使用 dimensions.ai SDG 分类器41 识别的，这是一种受监督的机器学习算法，在基于手动策划的扩展搜索字符串（每个目标一个）的基础上收集的 SDG 相关出版物上进行训练。 生成的出版物数据集包含 21 年来的 779,901 篇文章（N2000 = 12,727 篇文章，N2020 = 67,916，年平均值 = 37,138，年 SD = 21,119），每篇都被分类器标记为与感兴趣的六个可持续发展目标中的一个或多个相关。 自 2012 年以来，我们首次使用该数据集来计算科学文献中六个 SDG 之间的重叠。对于每对 SDG，它们的文章之间的重叠是使用 Jaccard 指数来衡量的，Jaccard 指数是两组之间对齐的常用衡量标准42：文章被归类为 与两个 SDGs（集交集）相关，相对于所有分类为与任一 SDGs（集并集）相关的文章。 从出版物数据集中，我们还提取了 2000 年至 2020 年三对中每对 SDG 的 21 个科学家年度合着网络。 这产生了总共 63 个网络（21 年 × 3 个 SDG 对），拥有 2210 至 51,270 名作者和 1887 至 316,769 名协作关系。 作者被分配了三个标签之一，表明他们的出版物在当年对每对 SDG 做出了贡献（SDG A、SDG B、SDG A + B）。 提出了 SDG 学术整合的三个衡量标准：(1) Jaccard 指数表示对两个 SDG（SDG A + B）做出贡献的作者相对于对 SDG 中任一 SDG 做出贡献的所有作者（SDG A、SDG B 或 可持续发展目标 A + B）； (2) 共同作者网络 40,43 Louvain 社区的平均 SDG 熵反映了作者在同一合作科学家社区中对不同 SDG 集群做出贡献的程度； (3) SDG 标签共同作者网络的分类系数44 衡量具有相同 SDG 标签的作者之间发生合作的程度，0 (1) 表示具有不同 SDG 标签的作者之间的最大（最小）合作组合。

\*基于 doc2vec 嵌入中余弦相似性的 SDG 之间的话语重叠。 (A) 余弦相似度矩阵的热图与矩阵层次聚类的树状图。 (B) 余弦相似性加权网络（节点颜色代表通过 Louvain 网络社区检测算法识别的子组 40）。

Supplementry

文本预处理和词性标注

UDPipe 的树库依赖解析器应用于 UN P/I 报告1-5，以生成用于所有后续分析的词性 (POS) 注释。 UDPipe 是一个多语言自然语言处理 (NLP) 套件，可在 R 包存储库 CRAN 和 GitHub 上找到，包括用于标记化、词性标记、词形还原和依赖项解析的管道6。 提取的词形还原和词性标记应用于分析的不同阶段。

用于 TextRank 分析的文本预处理。 正如在应用 NLP7,8 中变得普遍的那样，我们的分析在通过 TextRank 提取 SDG 关键字时专门关注词形还原的名词和形容词——形成语料库的主题词或短语。 具体来说，TextRank 的 5 词移动窗口以这样一种方式应用于语料库，即在结果图中只有名词和形容词被分配一个节点，只有当这些名词和形容词出现在彼此 5 个词以内时，它们之间才会出现联系，并且 通过 POS 标签相关（例如，“优先考虑水的可用性，而不是短期宏观经济增长”会将“可用性”与“水”和“优先级”联系起来，但不会将“优先级”与“宏观经济”或“ 生长”）。 在使用 R TextMining (TM) 包 9 的停用词功能检测到停用词后，此分析中省略了停用词。 在对联合国的 SDG 进展和信息 (P/I) 文本进行定性检查后，我们将以下词添加到 TM 包提供的标准停用词列表中：平均、真实、许多、水平、数据、总计。

用于 doc2vec 分析的文本预处理。 这部分分析采用了稍微不同的预处理管道。 自监督连续词袋 (CBOW) Word2Vec 模型是 doc2vec 模型的基础，它超越了简单地将单词放置在网络中，并学习在给定重复上下文的情况下预测单词（例如，“发展”可以从 上下文“可持续……目标”）10,11。 在此分析中，只关注名词和形容词会从 Word2Vec 学习的单词上下文中删除关键信息。 因此，我们应用了更细致的预处理管道。 对于每个文本，我们使用 Penn Treebank 项目 12 中的 POS 标签识别、提取和连接所有词形还原词，该项目区分副词（RB 标签）、限定词（DT 标签）、不同模式的动词（VB 和相关标签），以及 除了名词和形容词之外，还有许多其他词类。 POS 标签被附加到每个单词（例如“goal”→“goal//NN”），确保我们的 Word2Vec 模型只将占据相同 POS 的单词识别为同义词（例如“chosen”作为 choose 的过去分词将是 “选择//VBN”，而“选择”作为形容词将是“选择//JJ”）。 这种更精确的 POS 程序使我们的模型能够生成更准确的文档嵌入，通过识别相同的词根据其词性位置具有不同的含义，更好地导航语言的细微差别。

#### Supporting contents

SDG P/I 报告中的重要内容：TextRank 分析

我们应用 TextRank 算法来简洁地描述与每个 SDG 有关的进度和信息 (P/I) 文档中的内容。 TextRank 是一种基于网络的模型，用于识别文本文档中的关键词和短语 13。 预训练的 UDPipe 树库解析器使我们能够识别所有名词和形容词，即最有可能包含主题词的词性。 TextRank 算法生成一个 8 词搭配网络（图形），其中每个节点都是一个词，每个链接表示 8 词窗口内两个词之间的共现。 然后根据 Google 在该搭配网络中的 PageRank 中心性分数对单词进行排名 14。

TextRank 识别单个关键词，但不直接检测关键短语。 基于文本的图形表示，它不直接处理 n-gram（即连续的单词序列）：与所有中心性度量一样，PageRank 中心性是针对每个图形节点（即一个单词）计算的，而不是针对序列计算的 的节点。 因此，我们通过 (1) 识别频繁出现的 2 到 8 个单词的 n-gram，以及 (2) 获得短语级别的 PageRank 中心性作为每个 n-gram 的组成单词的平均 PageRank 分数来获得关键短语。

此描述性分析的结果显示在图 S1 中。 为了确保跨 SDG 的可比性，PageRank 分数由 SDG 标准化。 17 个 SDG P/I 文件在语言使用上有很大的重叠。 事实上，在所有 17 个 P/I 文本中确定的 1,599 个独特关键词和短语中，有 306 个 (20%) 由一个或多个可持续发展目标共享。 与 SDG 特定关键字的中低 PageRank 分数相比，大多数人在 PageRank 上的得分较低或极高（图 S1 B）。 图 S2 描述了基于完全匹配的 TextRank 关键字的分析映射 SDG 重叠的结果。 Jaccard 指数用于表示一对 SDGs 共享的单词和短语在其完整的关键词和短语集中的比例 15。 在生成的 Jaccard 相似性矩阵上使用完整链接聚类，我们提取了以下三个可持续发展目标集群：(1) 10-不平等、2-饥饿、6-卫生、16-和平、5-性别； (2) 8-经济，1-贫困，4-教育，17-伙伴关系，3-健康； (3) 9-工业、7-能源、11-定居点、12-消费、14-水生、13-气候、15-陆地。 与下面介绍的基于主题模型的复制不同，这些集群部分偏离正文中报告的 doc2vec 结果。 然而，它们确实证实了正文分析中报告的一些强烈的相互联系，例如 5-性别与 16-和平、8-经济与-健康，或 14-水生与 15-陆地之间的联系。 一方面，TextRank 关键字重叠分析与主题模型和 doc2vec 分析之间的结果差异是预期的，因为主题模型和 doc2vec 旨在识别除了精确单词匹配之外的同义词，如所讨论的 在正文中。 5-Gender 和 16-Peace 的分组表明它们在关键词/短语的使用上非常匹配，表明其中大量重复使用的语言。 相比之下，与主要文本分析相比，17-Partnerships 和 10-Inequality 之间的距离增加表明这对 SDGs 不使用相同的语言，而是在相似的词上下文中使用相似的语言（或同义词），这决定了它们更强的关系 在我们的主题模型中。

**联合国话语中 SDG 之间的相似之处：使用替代模型规范和 NLP 方法的稳健性检查**

具有不同隐藏层数的 Doc2vec 模型。 考虑到我们的文档相对较小，我们在训练我们的 doc2vec 模型时进行了敏感性分析，以确保报告结果的稳定性。 按照惯例，正文中报告的模型被指定为包括 300 个隐藏层（维度）、250 次迭代、一个 5 字移动窗口和负采样 16、17。 在训练期间，我们迭代地增加模型中隐藏层的数量（10、50、100、150 和 300），直到我们达到嵌入导致稳定集群的点。 从图 S3 中可以看出，无论模型规格如何，以人类为中心的可持续发展目标和环境可持续发展目标之间出现了明显的分歧。 然而，以人类为中心的可持续发展目标的层次聚类存在差异。 在 300 个隐藏层处，我们观察到 4-Education 在集群之间的放置存在微小变化。 我们怀疑该可持续发展目标倾向于在 5-性别和 16-和平的配对以及 1-贫困、2-饥饿和 3-健康的三元组之间浮动，表明 4-教育与这两个问题的广泛相关性 性别不平等和贫困（及其相关因素）。 尽管如此，4-教育一直被置于个人层面的以人类为中心的可持续发展目标中，而 9-工业、8-经济、17-伙伴关系、10-不平等和 7-能源都被置于宏观的以人类为中心的可持续发展目标集群中——所有这些都与 正文中报告的模型。 正文中检查的三个密切相关的 SDG 对（5-性别和 16-和平、10-不平等和 17-伙伴关系、14-水生和 15-陆地）在所有模型规范中一致出现，更广泛的区别也是如此 在宏观和个人层面的 SDGs 之间（10 隐藏层模型除外）。 然而，随着我们减少隐藏层的数量，以人类为中心的集群变得越来越不稳定，特别是在隐藏层数量较少的 6-Sanitation、7-Energy 和 11-Settlements 的位置。 这种不稳定性是意料之中的，因为我们的文件是一组政策文本，其中包含大量与可持续性相关的词汇字典，这意味着需要更多的迭代来抽取代表可持续发展目标之间差异的稀缺独特词汇和上下文（见 图 S1)。

在美国新闻报道语料库上预训练的 Doc2vec 模型。 虽然正文中介绍的 doc2vec 模型依赖于在 UN P/I 报告语料库上训练的词嵌入，但我们还测试了结果对预训练 doc2vec 模型的敏感性。 该模型在来自众多来源（路透社、纽约时报、洛杉矶时报等）的 11.8GB 在线美国新闻报道上进行了训练（具有 500 个隐藏层、10 字移动窗口和负采样的 CBOW）。 新闻报道总计 23 亿字，可从 word2vec++ GitHub 存储库下载18。 如图 S4 所示，该分析的结果在很大程度上与我们的主要模型中确定的集群和相互依赖性一致。 例如，11-定居点和 6-卫生设施分别与其余可持续发展目标配对。 环境可持续发展目标（12-消费、13-气候、14-水生、15-陆地）被分配了自己的集群。 复制了最强的配对（10-不平等和 17-伙伴关系；14-水生和 15-陆生；5-性别和 16-和平）。

然而，与正文中呈现的结果有一个显着差异：1-贫困和 2-饥饿与 10-不平等和 17-伙伴关系形成了一个独特的集群，分别与个人层面的以人类为中心的可持续发展目标（3-健康， 4-教育、5-性别、16-和平）和宏观经济可持续发展目标（7-能源、8-经济、9-工业）。 这种改变的原因很复杂，取决于 Word2Vec 模型从美国新闻语料库中学到了什么。 回想一下，在训练过程中，Word2Vec 根据相似上下文中的单词使用来学习单词关联 11。 鉴于替代模型是在美国新闻语料库上训练的，它已经从美国新闻文章中学习了单词关联。 因此，告知 doc2vec 对可持续发展目标之间关系的表示的同义词现在是从联合国以外的来源中先验地学习的。 事实上，我们选择在主要文章中报告在联合国 P/I 语料库上训练的模型（尽管联合国 P/I 语料库比美国新闻语料库小得多）的原因是联合国文本反映了联合国的观点、优先事项、 以及对可持续发展目标及其相互依存关系的观察——而不是美国新闻媒体中出现的对可持续发展目标相关主题和语言的更笼统的表述。 两种模型之间的差异可能是由于 UN P/I 文本为 doc2vec 模型提供了对组成语言更好的“理解”。 也可能是，与联合国文本相比，美国新闻媒体倾向于更频繁地讨论贫困和饥饿方面的全球不平等，而不是获得教育、公共卫生、性别和制度稳定方面的问题。

Doc2vec 模型分别应用于不同年份的 P/I 报告。 作为 doc2vec 的最终验证步骤，我们将最终模型重新应用于按年份分类的联合国 SDG P/I 文件。 这种分解产生了 85 个文本：17 个 SDG x 2016 年到 2020 年之间的 5 年。此验证步骤的目标是测试模型准确地将每个 SDG 的所有年度迭代组合在一起的能力。 从图 S5 中可以看出，当将完整链接聚类应用于从正文中提供的 doc2vec 模型中提取的文档余弦相似度分数时，我们取得了基本准确的结果。 图 S5 突出显示了 5 个组，与正文中讨论的 SDG 集群数量相同； 和 17 个组，对应于 17 个可持续发展目标。 在聚合和分解分析中确定的 5 个集群之间可以看到一些主要相似之处。 特别是，环境可持续发展目标与以人类为中心的可持续发展目标之间存在明显的分歧。 分类分析中最值得注意的偏离是 (a) 1-贫困与 10-不平等放在同一组中（尽管该组的重点与 6-卫生、7-能源、11-定居点、17-伙伴关系相关）， (b) 2-饥饿与 14-水生和 15-陆生一起被纳入环境集群。

将 85 份文本分成 17 组（每个 SDG 一组）时，我们的模型成功地将 SDG 1、3、4、6、11、13、14 和 15 的所有 5 份年度 P/I 文件分组。此外，12-Consumption 被正确地分为两个不同的组（2016-2018 年和 2019-2020 年），表明 2018 年后讨论该可持续发展目标的方式发生了转变。 与主要分析的结果一致，我们的模型努力将正文中讨论的两个最强对的 P/I 文件分开：5-性别和 16-和平，10-不平等和 17-伙伴关系。 它创建了两个组合集群，收集了每对的 10 份年度文件。 然而，该模型能够在两个独立的集群中正确区分 14-水生和 15-陆生（主要分析中第三强的对）。 最后，8-Economy 和 9-Industry 的 10 份文件也聚集在同一个集群中，这与主要分析中检测到的这两个 SDGs 之间的强联系一致。 最终，我们在 P/I 文本上训练的模型能够很好地识别讨论同一 SDG 的文本组。 一个额外的 doc2vec 模型在美国新闻语料库（如上所述）上训练并应用于 85 个分类的年度 P/I 文件，识别出大致相同的 5 和 17 个集群（图 S6），进一步证明我们的主要结果对 要么是用于模型训练的文本语料库，要么是 SDG 对年度 P/I 文档的聚合。

在词频-逆文档频率上训练的潜在语义分析。 我们还测试了我们的结果对一种完全不同的 NLP 方法的稳健性，即基于术语频率-逆文档频率 (TF-IDF) 的加权文档-术语矩阵的潜在语义分析 (LSA)（图 S7）19， 20. 按照 Deerwester 等人描述的程序，我们将 17 个 P/I 文档转换为术语文档矩阵 (TDM) 21。每个矩阵行代表 17 个文档中的一个或多个中的一个术语，每一列是其中一个 17 个 SDG P/I 文件。 停用词（包括 SDG P/I 文本特有的一组手动精选词，例如“per”、“cent”、“average”、“total”）和出现次数少于 4 次的词已从 TDM 中删除。 我们进一步对 TDM 进行子集化，使其仅包含具有特定 Penn Treebank POS 标签（单数名词、复数名词、单数专有名词、复数专有名词、形容词、比较形容词、最高级形容词）的词。 然后将术语频率转换为 TF-IDF，根据每个术语在每个文档中的频率和包含该术语的文档的倒数频率对每个术语进行加权，以便在随后将奇异值分解应用于矩阵时流行的术语具有较小的权重。

正文中呈现的 doc2vec 结果与 LSA 结果之间有很强的相似性。 环境可持续发展目标和以人类为中心的可持续发展目标之间再次存在明显分歧。 就像在 doc2vec 分析中一样，连接最紧密的三个对是 5-性别和 16-和平、10-不平等和 17-伙伴关系、14-水生和 15-陆地。 与主要分析的结果一样，1-贫困、2-饥饿、3-健康、4-教育、5-性别和 16-和平都相对接近，6-卫生和 11-定居点的基础设施可持续发展目标也是如此 . LSA 结果的层次聚类仅在将 8-Economy 与 1-Poverty 而非 9-Industry 配对时偏离主要结果。 然而，图 1a 中显示的结果清楚地表明 1-贫困、8-经济 (cos = .91) 和 9-工业 (cos = .87) 之间存在很强的余弦相似性。 这只是一个例子，其中 1-贫困与 2-饥饿、3-健康、4-教育、5-性别和 16-和平（.84<cos<.88，平均值 = . 87, sd = .02) 在分配集群时超过 1-贫困、8-经济和 9-工业之间的关系。 这说明了将可持续发展目标概念化为明确定义的、有界的集群而不是复杂的相互依存网络的固有局限性。

**隶属于三对高度相关的可持续发展目标的研究人员的共同作者网络**

图 S8 到 S10 是 3 个网络的可视化图，描绘了与正文中检查的三对密切相关的可持续发展目标（5-性别和 16-和平，10-不平等和 17）中的每一个相关的论文的同行评审出版物共同作者 -伙伴关系，14-水生和 15-陆生）。 每个网络的所有年份 (2000-2020) 都汇集在一起。 在主要文章的材料和方法部分中更详细地描述了构成的年度网络。

**Process:**

TextRank 预处理(好像可以用 BM25/ TF-IDF算法 来替换），Doc2vec + PV-DM(替代CBOW)-》Word2Vec + Skip-gram

Doc2vec 公式

<https://zhuanlan.zhihu.com/p/336921474>

TextRank 公式

<https://www.jianshu.com/p/8b27f729e30f>

BM25

<https://www.cnblogs.com/geeks-reign/p/Okapi_BM25.html#:~:text=BM25,%E7%AE%97%E6%B3%95%E5%AE%9E%E8%B4%A8%E4%B8%8A%E6%98%AF%E4%B8%80%E4%B8%AA%E7%94%A8%E4%BA%8E%E4%BF%A1%E6%81%AF%E6%A3%80%E7%B4%A2%E4%B8%AD%EF%BC%8C%E5%AF%B9%E7%BB%99%E5%AE%9A%E6%9F%A5%E8%AF%A2%EF%BC%88query%EF%BC%89%E5%92%8C%E8%8B%A5%E5%B9%B2%E2%80%9C%E7%9B%B8%E5%85%B3%E2%80%9D%E6%96%87%E6%A1%A3%EF%BC%88document%EF%BC%89%E8%BF%9B%E8%A1%8C%E7%9B%B8%E5%85%B3%E6%80%A7%E6%8E%92%E5%BA%8F%E6%89%93%E5%88%86%E7%9A%84%E6%8E%92%E5%BA%8F%E5%87%BD%E6%95%B0%E3%80%82>

TF-IDF

<https://blog.csdn.net/weixin_43283397/article/details/107409456>

选取联合国官网最新发表的42篇文章，使用TF-IDF算法进行预处理

TF-IDF 预处理

Doc2Vec + PVDM

热力图 (Normalized Eucleanian Distance)

协方差表示

Python 热力图

<https://blog.quantinsti.com/creating-heatmap-using-python-seaborn/#:~:text=A%20heatmap%20is%20a%20two,as%20per%20the%20creator's%20requirement.>

Python Correlation Matrix

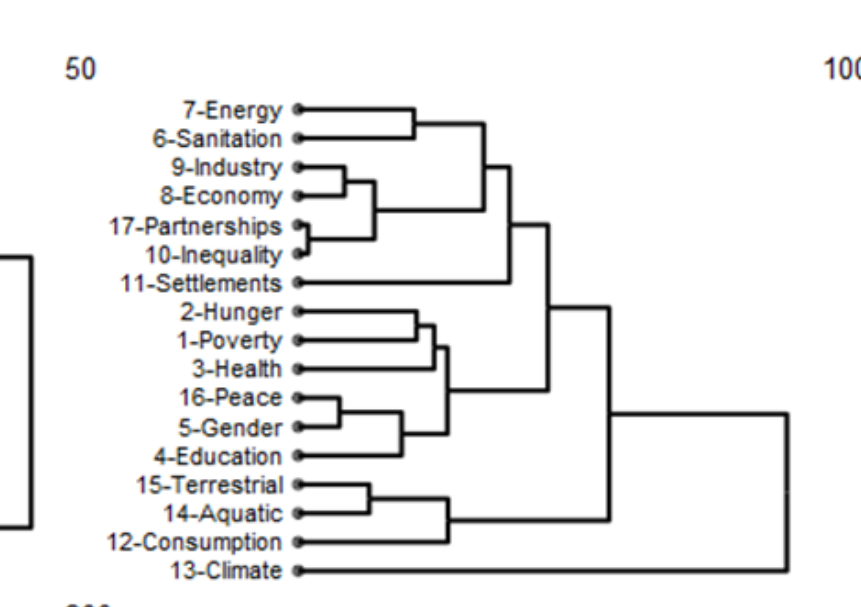
<https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec>

Normalized Eucleanian DistanceCorrelation Matrix

归一化欧几里得距离

90个隐藏层

90 Layers



### To do list

周日上午：

欧姐：(按顺序来做)

1. 模型假设

2. 剩下几问有什么思路

3. 找reference

tqq： 深度学习(pca有机会就加上) 跟一遍，把图花了，跟模型假设保持一致（一二 问），ppt画图

- 知道概念，知道在做什么，然后paraphrase就可以了

- 改数据：

\* 越奇怪越好

\* 图不能太像

画图素材：

<https://docs.google.com/presentation/d/1ZuAg-cqMxDREOXDL3mFwg6HPOSiB8izBv9eO0dI8Exg/edit?usp=sharing>

Xmind

周日需要完成的：至少前4问, 最好都写完

**Ziou 2.18**

**References:**

1. United Nations. Sustainable Development Goals: 17 Goals to Transform our World.[Sustainable Development Goals: 17 Goals to Transform our World | United Nations](https://www.un.org/en/exhibits/page/sdgs-17-goals-transform-world)(2020)
2. National Bureau of Stastistics of China. International Stastical YearBook (2020)