

介绍：（to be cont.）

整个模型是从UN Economic and Social Council (ECOSOC)的文章中寻找SDG相关的词汇，来寻找SDG之间的关系。为了实现这样的模型，我们采用的词向量模型是Word2Vec模型。Word2Vec 是一种流行的自然语言处理技术，它可以将自然语言中的单词映射到低维向量空间中，使得单词在向量空间中的相对位置可以表示它们在语义上的相似性和相关性。除此之外Word2Vec使用了一种简单但高效的算法，称为Skip-gram或Continuous Bag-of-Words（CBOW）。这些算法不需要大量计算，因此Word2Vec的训练速度比其他模型更快。且与其他词向量模型相比，Word2Vec具有更好的性能。因此，Word2Vec 可以应用于我们的SDGs分析场景。但在实际使用中绝大部分项目会选择CBOW. 但由于我们训练用的语料库较小且关系更为复杂，所以我们选择skip-gram。因为在skip-gram中，每个单词都有一个独立的向量表示，因此每个向量可以专门捕捉该单词在语料库中的不同语义，包括它在不同上下文中的用法。

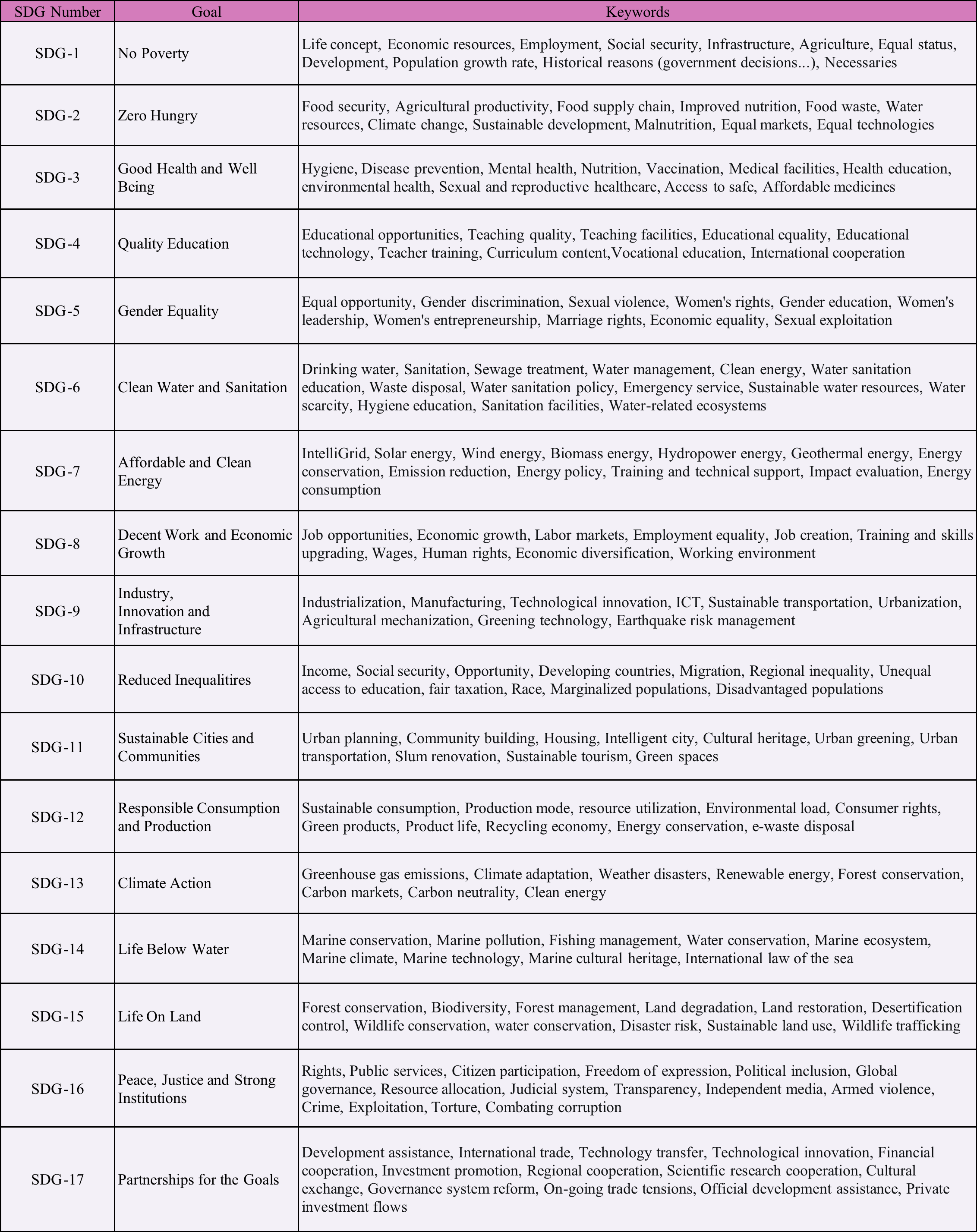
数据来源：

我们从ECOSOC发布的文章中选取了最近发布的57篇文章使用Natural Language Processing(NLP)技术进行分析。我们选取ECOSOC的文章的原因是其提供了广泛的数据，涵盖了各种可持续发展目标及其相关指标。这些数据不仅包括国际层面的数据，也包括国家层面的数据，因此可以对不同地区和国家的进展情况进行比较分析。并且，ECOSOC提供的数据基于标准化的指标和测量方法，可以确保数据在不同地区和国家之间具有一致性和可比性，从而更容易进行分析和评估。除此之外，ECOSOC提供的数据是公开可用的，数据源可以追溯到其来源，使得研究人员和政策制定者可以验证数据的准确性和可靠性。因此，我们选择根据ECOSOC近期发布的57篇文章的关键词来寻找17个SDG之间的network of the relationship.

详细步骤：

1. 目标定位

我们根据联合国的17个SDG，将每个SDG细分为若干个与其相关的关键词，在我们选择57篇文章当中筛选相关关键词的出现频率。



1. 文章预处理

\*加公式 (laf

我们认为分词算法根据其核心思想主要分为两种，第一种是基于字典的分词，先把句子按照字典切分成词，再寻找词的最佳组合方式；第二种是基于字的分词，即由字构词，先把句子分成一个个字，再将字组合成词，寻找最优的切分策略，同时也可以转化成序列标注问题。归根结底，上述两种方法都可以归结为在图或者概率图上寻找最短路径的问题。这里，针对SDGs，因为我们都是从联合国官方文档里面获取相应的信息，所以，第一种方法可能更加适合我们的情况。最终，我们选择了这种情况。

文章预处理是指对原始文本进行处理，以便更好地应用自然语言处理（NLP）技术。文章预处理的目的是为了提高NLP任务的准确性和效率，减少错误和噪声，并为后续的数据分析和挖掘提供更加清晰的数据。将文章转换为机器可读的格式。对此，我们使用Python中的Natural Language Toolkit (NLTK)库对文章根据SDGs划分的关键词进行预处理，包括去除标点符号，分词，去除停用词等。此外，将文本分解成单词或标记，以便更好地理解文本的语义和结构。分词也可以帮助我们更好地理解句子和文档的含义。

同时，我们还需要进行词形还原和词干提取：在不同的形式下，同一个单词可能出现多次，如“running”和“runs”，它们的原形都是“run”。这样对数据进行处理因此，为了提高准确性，需要对单词进行词形还原或词干提取，使它们归一化为相同的形式。

此外，这些我们从联合国官网获取的文本数据通常包含一些不必要的信息，如标点符号、HTML标记、表情符号、网址等。这些噪声会干扰自然语言处理算法的准确性，因此需要去除。将句子分割成单词或子词，以便计算机可以理解和处理。在英文中，通常是以空格或标点符号为分隔符。这里，我们采用jieba词库进行分词，jieba采用了基于Trie树的分词算法，具有较高的分词速度和较小的内存占用。Trie树是基于以下搜索字典项目：

1. 从根节点开始第一次搜索
2. 取得要查找关键词的第一个字母，并根据该字母选择对应的子树并转到该子树继续进行检索
3. 在相应的子树上，取得要查找关键词的第二个字母,并进一步选择对应的子树进行检索。
4. 选代过程......
5. 在某个结点处，关键词的所有字母已被取出，则读取附在该结点上的信息，即完成查找其他操作类似处理

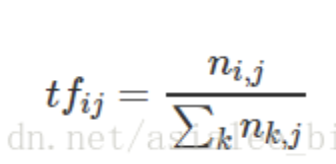
接着，我们还会采用停用词过滤的方法，停用词是指在文本中出现频率非常高，但并没有实际意义的词语，如“的”、“是”、“在”等。过滤掉这些停用词可以降低数据的噪声和冗余，提高后续处理的效率和质量。摆脱停用词对于任何自然语言处理任务都非常有意义。NLTK还允许用户根据自己的需要定制停止词列表。例如，他们可以从列表中添加或删除单词，或者根据自己的文本语料库创建一个新的列表。因此，我们会选择NLTK作为停用词表。

词干提取和词形还原：将单词的不同形态（如时态、复数形式）转换成其原始形式，以便提高文本的准确性和一致性。词干提取是指将单词缩减到其基本形式，如将“running”和“runs”都转换为“run”，而词形还原则是指将单词恢复到其原始形式，如将“am”, “is”, “are”都还原为“be”。

1. 筛选关键词

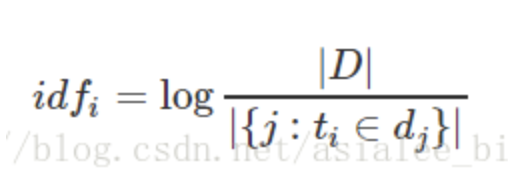
为了更好的寻找到文章的关键词，需要的文章进行处理。我们采用Term Frequency-Inverse Document Frequency(TF-IDF)算法，便于区分字词的重要性, 以此来筛选具有良好的类别区分能力词。TF-IDF算法能够准确地反映一个词汇在文档中的重要性。相比于仅使用Term Frequency，TF-IDF算法结合了Term Frequency和Inverse Document Frequency两个因素，可以更好地反映一个词汇在文档集合中的重要性。例如，对于一些高频词汇（如“the”、“be”等），它们在文档中出现的次数很多，但是它们在整个文档集合中出现的次数也很多，因此它们的TF-IDF值往往较低，可以被有效地过滤掉。下面是算法的具体步骤：

第一步，需要统计词汇出现的数量，也就是计算Term Frequency。在统计到文章词汇数量后，由于文章长短不齐，为了防止文章长短对数据造成一定影响，需要对该数据进行normalized processing。公式如下：



**N（i,j）**是该词在文章中出现的次数，分母则是文章中所有词汇出现的次数总和

第二步，我们需要计算Inverse Document Frequency（IDF），由**总文件数目除以包含该词语的文件的数目**，**再将得到的商取对数得到**。词汇的IDF越大，说明其类别区分能力越好。IDF计算公式如下：

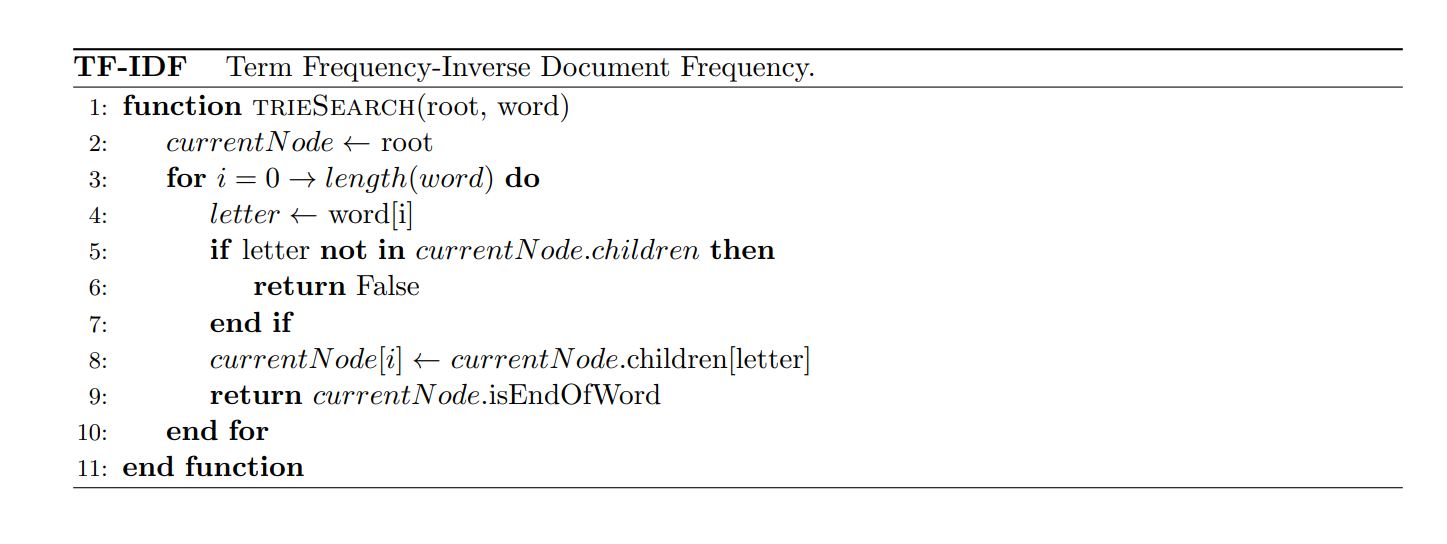


**ps:(这里分母应该是1+|{j:ti∈dj}|**

**|D|** **是文章总数**。**|{j:ti∈dj}| 表示包含该词汇的文章数目**

将TF与IDF相乘即可过滤掉常见的词汇。由此，我们可以提取每篇文章的关键词。

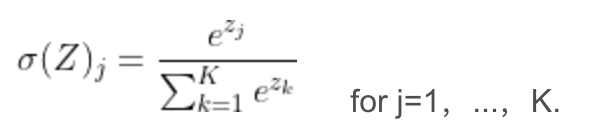
伪代码：



In this pseudocode, root is the root node of the Trie tree and word is the item that we want to search for in the dictionary. The function trieSearch iterates through each letter of the word and traverses the Trie tree accordingly. If a letter is not found in the currentNode's children, then the search terminates and the function returns False. If all letters in the word are found and the final node is marked as the end of a word, then the function returns True.

1. 训练Word2Vec模型

首先使用NLTK库构建词汇表，将每个单词映射到唯一的整数标识。接下来就是训练Word2Vec模型，设置skip-window=4，Hidden Layer=90。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Google Scholars及ECOSOC发布的其他文章，以此作为训练数据。我们使用TensorFlow作为工具来训练Word2Vec skip-gram模型。这里需要将上述所有文章作为输入。对外，我们采用one-hot进行编码将所有的字符转化成二进制的编码形式。再将one-hot编码后的词向量，通过神经网络的hidden layer，映射到一个低纬度的空间。再将hidden layer和输出层进行全连接。然后，我们需要将向量转换为概率分布的函数softmax用于计算当前单词与其他单词的相似度，以确定最可能的上下文单词。softmax函数可以表示为：



z就是我们的相似度

w\_c表示目标单词的one-hot向量

v\_c表示目标单词的词向量

u\_x表示除目标单词外第x个单词的词向量

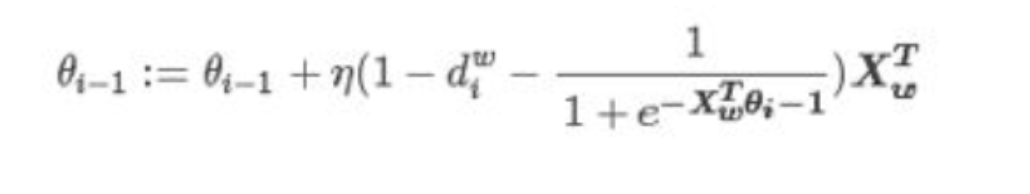
W表示目标单词矩阵

W'表示其他单词矩阵

词向量的维度是d

词汇表的维度是V

接下来，利用梯度上升法求似然函数的最大值。最后，我们需要更新Skip-gram模型的参数。更新参数的公式如下：

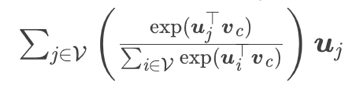


𝞋模型参数，一开始随机初始化

X(w)词向量

更新中心词的向量如下：（把这两个合并起来





V(c 中心词向量

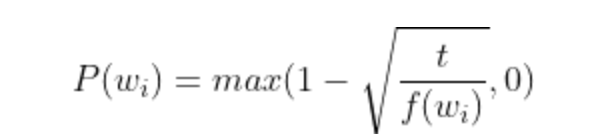
*uo*​背景词矩阵

背景词的索引为o

j的取值范围是− m ≤ j ≤ m , j不能为0

当梯度一旦收敛，停止更新，此时概率值为极大值，词向量为最优解。

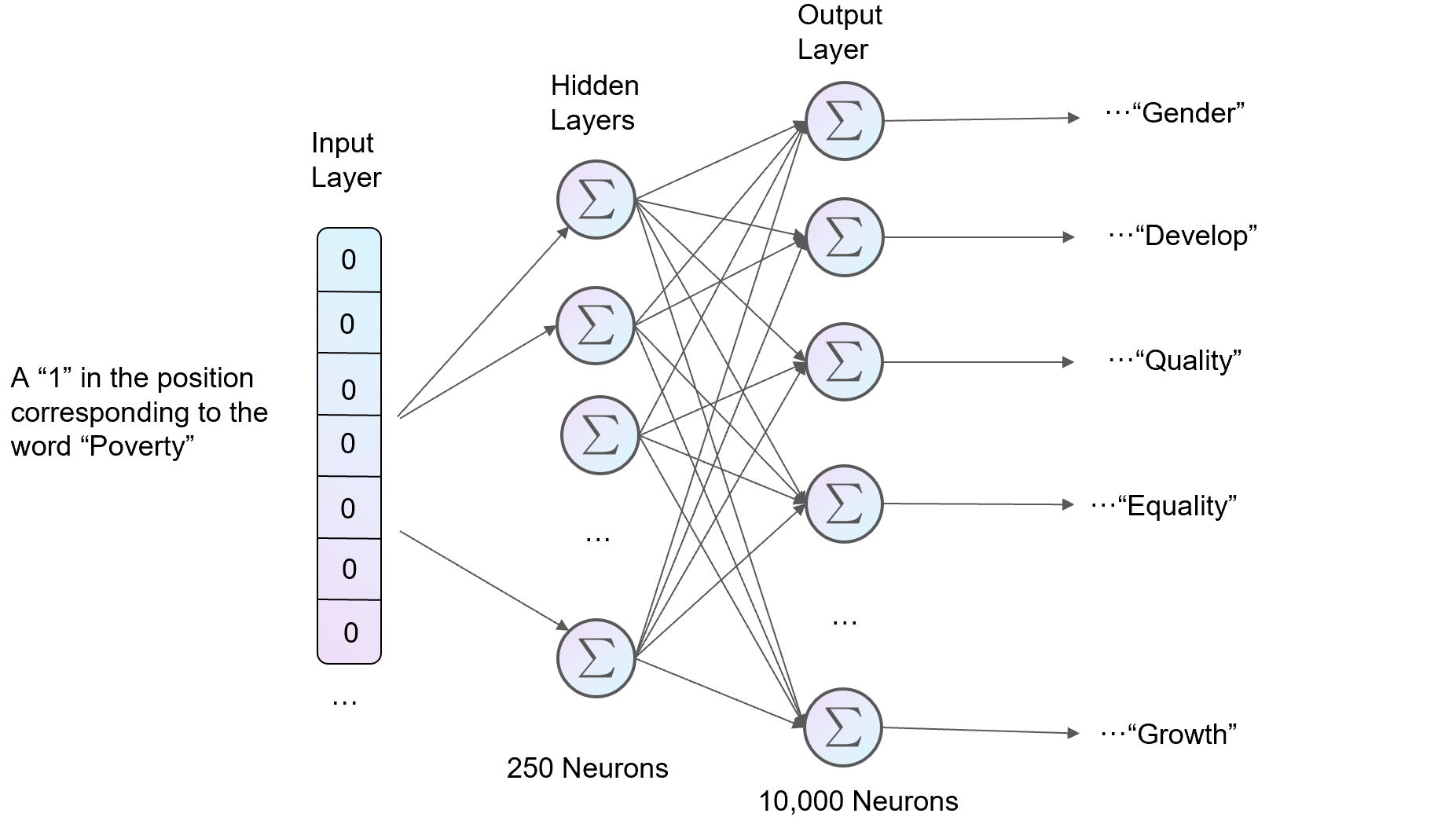
除此之外，我们还将使用二次采样法处理原始文本，这也是为了降低高频词在语料中出现的频次。方法是随机将高频的词抛弃，频率越高，被抛弃的概率就越大；频率越低，被抛弃的概率就越小。标点符号或冠词这样的高频词就会被抛弃，从而优化整个词表的词向量训练效果。具体来说，数据集中每个被索引词 w\_i将有一定概率被丢弃，该丢弃概率为：



其中f(w\_i)是w\_i 的个数与总词数之比

常数 t 是一个参数（此处设置为10^-4）。

So, the process actually looks like this:



模型实现：

We set up a Doc2vec model with 90 hidden layers. Considering the relatively small number of documents, we also performed a sensitivity analysis when training the doc2vec model to ensure the stability of the results.

在训练的过程中，我们以25位单位尝试了从50到300个隐藏层数量当中的最好的，迭代的增加了模型的隐藏层数量 which makes the model neither 过拟合 nor 欠拟合。The model will alter to some extent when there are fewer hidden layers, but when there are more hidden layers, the model is more likely to remain stable. This instability is expected given that our document is a collection of policy statements, and it means that additional iterations are needed to sample the limited number of unique terminology and settings that serve as symbols of the distinctions between the SDGs.

While during the changing of the number of hidden layers, 我们发现了4-教育问题的变化往往会连带着影响1-贫困和2-平衡，甚至还会影响5-Gender和16-Peace。这表明这三者之间存在着一定的关联。此外，9-工业，8-经济，7-能源这三项经常聚集在一起，这可能是因为他们都与国家的宏观层面有关。

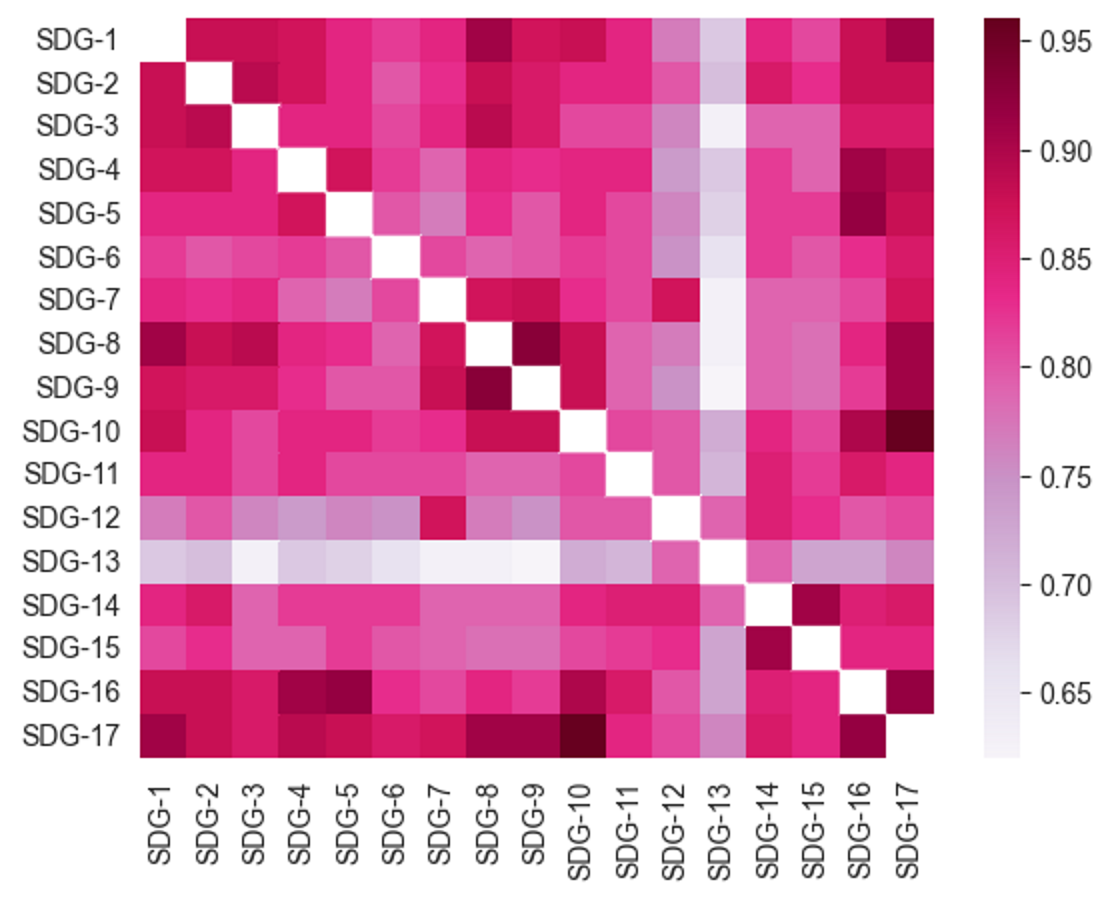
此外，The observation of three closely connected SDG pairs across all models—5-Gender and 16-Peace, 10-Inequality and 17-Partnership, 14-Aquatic and 15-Terrestrial—indicates the possibility of both micro- and macro-scalability. The Variations and Relationships Among Sustainable Development Objectives

归一化欧几里得距离 (laf

The normalized Euclidian distance extracted from the doc2vec model presented in the main text is also used to obtain largely accurate results when applying complete-link clustering to documents. It can be seen in this figure.

The analysis has identified three pairs of SDGs that are closely related to each other: poverty and hunger, aquatic and terrestrial life, and inequality and partnership. It also indicates that the development goals related to poverty, hunger, health, education, gender equity, peace, sanitation, and sustainable infrastructure tend to be similar to each other based on the findings of the first analysis.

However, the analysis also found that the relationships between poverty, economy, and industry are more significant compared to those between education, gender, and peace when clusters of goals are assigned. This implies that the SDGs should not be viewed as isolated clusters, but rather as complex and interconnected networks, as there are many interdependencies and relationships between them.



Generally speaking, anther method cosine similarity is more likely to be used during academic research. Cosine similarity is invariant to the magnitude of the vectors and only considers the direction, which can be useful when the magnitude of the vectors is not meaningful or when the focus is on the similarities of the direction between the vectors. On the other hand, normalized Euclidean distance is commonly used when comparing the similarity between points in a low-dimensional space. In this case, the focus is on both the direction and magnitude of the vectors. Normalized Euclidean distance is the Euclidean distance between the vectors when they are normalized to have unit length, which can make it easier to interpret distances between vectors.



Task3:

Climate change has been mentioned so many times in recent years because it is one of the most pressing and urgent global challenges that we face today.

Climate change is a major global threat that affects many aspects of our lives, from food and water security to human health and the economy. SDG 13 specifically aims to take urgent action to combat climate change and its impacts. The targets under this SDG include reducing greenhouse gas emissions, promoting climate resilience and adaptation, increasing awareness and capacity on climate change, and mobilizing climate financing.

As climate change continues to intensify, it is essential to address its impacts and implement measures to mitigate and adapt to them. By achieving the targets under SDG 13, countries can help limit the worst effects of climate change, reduce vulnerability to its impacts, and transition to a more sustainable, low-carbon future. Therefore, SDG 13 is critical to achieving a more sustainable and equitable world for all.

Therefore, the SDG we choose is 13 - climate action. If we take apart SDG 13 - Climate Action, it would have negative effects on other SDGs as well. Climate change is a global issue that affects several aspects of our lives, and its impacts are interconnected with other global challenges, including poverty, hunger, health, and biodiversity. Here are some of the effects that taking apart SDG 13 could have on other SDGs:

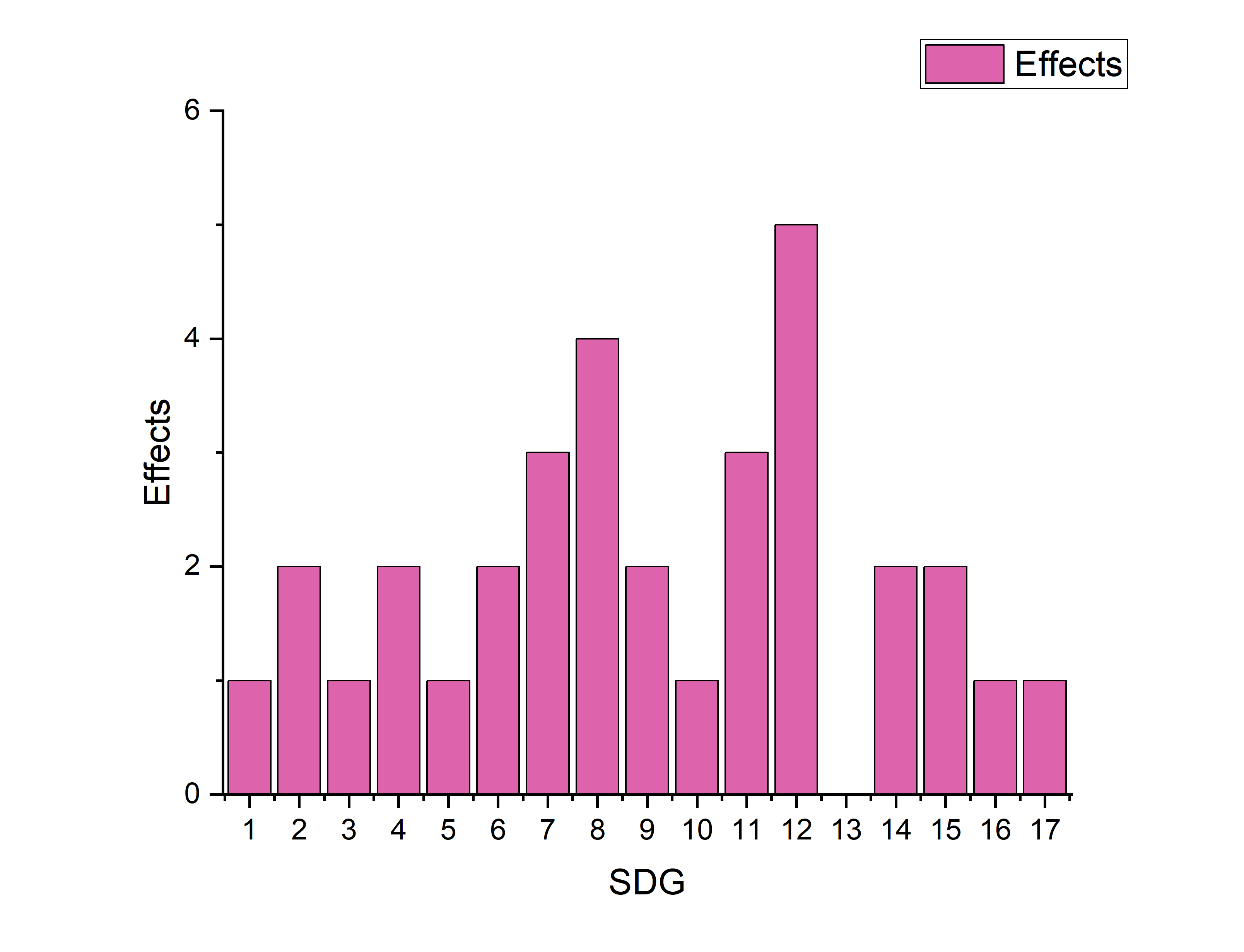
SDG 1 - No Poverty: Climate change can exacerbate poverty, particularly in developing countries, by affecting agricultural productivity, increasing the frequency and intensity of extreme weather events, and increasing health risks. Without SDG 13, there may be less investment in climate adaptation measures, making it harder to achieve SDG 1 - No Poverty.

SDG 2 - Zero Hunger: Climate change affects food security and agriculture by reducing crop yields and increasing food prices. SDG 13 includes targets to promote climate-resilient agriculture and increase the availability of climate financing for agriculture. Without SDG 13, progress towards SDG 2 - Zero Hunger could be hindered.

SDG 3 - Good Health and Well-being: Climate change can exacerbate health risks, particularly in vulnerable populations. SDG 13 includes targets to promote awareness of climate change impacts on health and to increase investments in health systems to improve resilience to climate change. Without SDG 13, it may be more challenging to achieve SDG 3 - Good Health and Well-being.

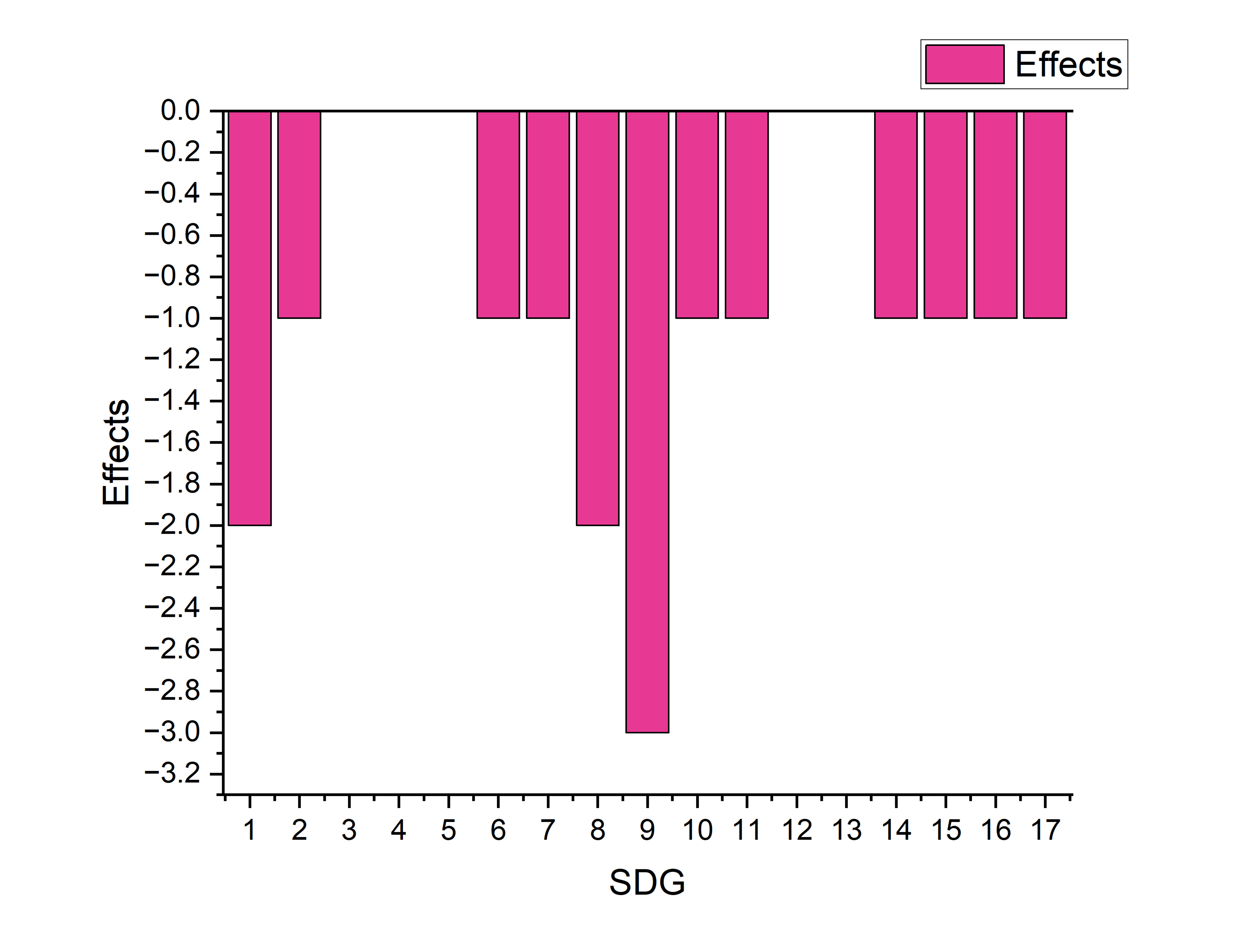
SDG 14 - Life Below Water and SDG 15 - Life on Land: Climate change affects ocean and terrestrial ecosystems, leading to biodiversity loss and species extinction. SDG 13 includes targets to reduce greenhouse gas emissions and promote sustainable land use and fisheries management. Without SDG 13, it may be more challenging to achieve SDG 14 - Life Below Water and SDG 15 - Life on Land.

负面图



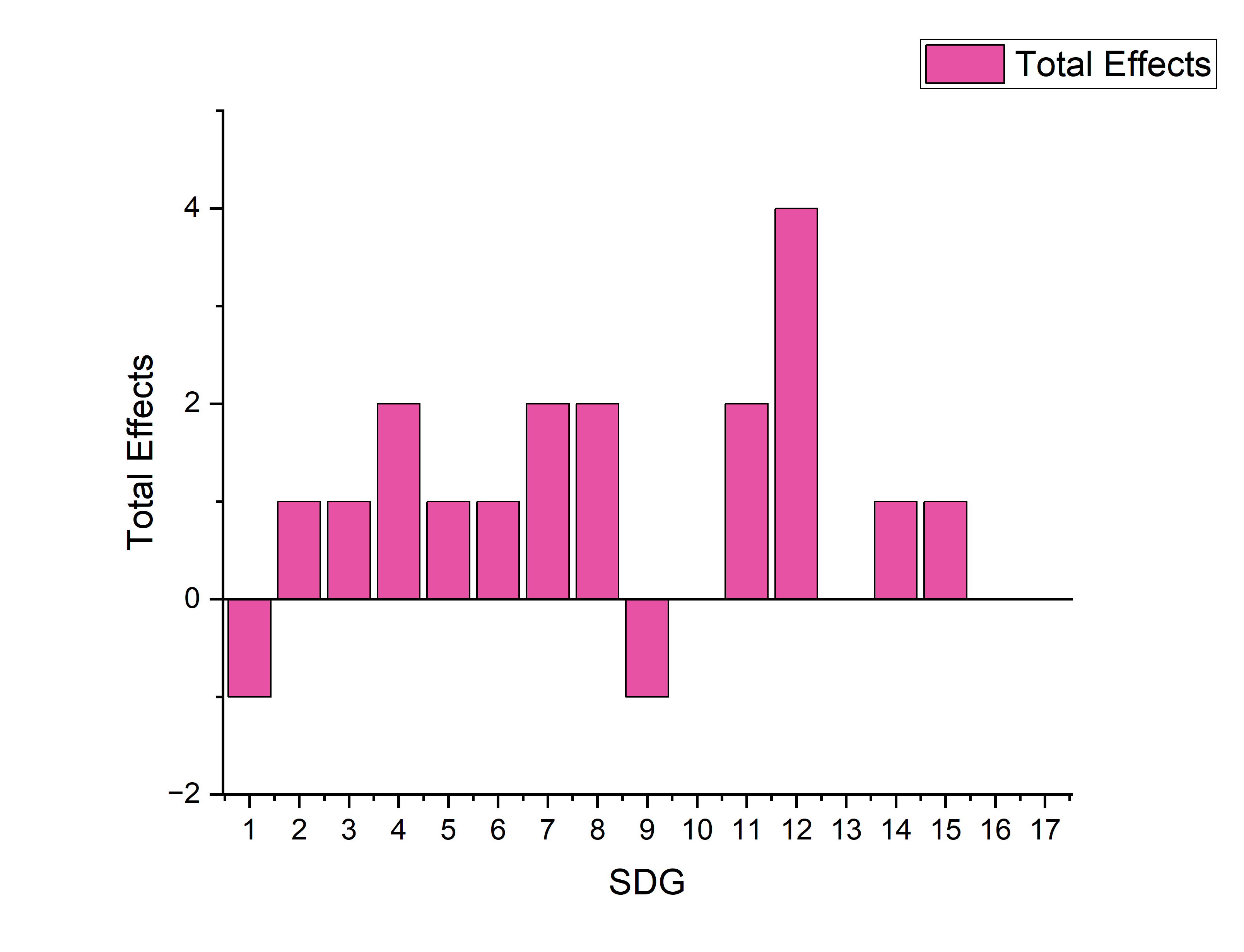
There is some evidence to suggest that some areas of the world may see some good effects from climate change, at least temporarily. Any benefits, however, are likely to be somewhat limited. For instance, higher average temperatures in temperate regions could lend assistance to initiatives made to boost agricultural production (2.3). According to the research, however, the countries that already have a high standard of living are the ones most likely to benefit from these favourable benefits, which will only serve to exacerbate the imbalance that exists between countries with high and low standards of living.

正面图



In summary, taking apart SDG 13 would not only have negative effects on climate change but also hinder progress towards other SDGs, which are interconnected with the challenges of climate change.

总的图



Task4：

模型评价（优势/劣势

优势：

1. 预测准确性

该模型可以通过大量数据的学习和训练，提高预测准确性。虽然在使用该模型时，仅选取了ECOSO最近发布的57篇文章，但理论上，可以选取ECOSO发布的所有文章作为原始数据，甚至可以选取其他质量有保证的文章作为原始数据（如，各个国家官方报告等）。随着原始数据的增加，该模型可以更加有效的从数据中发现潜在规律与network of relationship，使得预测准确性更高。

1. 可扩展性

这个模型可以拓展到有关SDG其他方面的分析，如SDG的进展情况，SDG的政策和法规，SDG的实践案例，SDG的监测和评估等（这里应该是有后续task用到的东西，后续看到记得删掉）。联合国文献中记录了许多关于实现SDGs的实践案例，使用该模型选取然后分析这些案例可以帮助其他国家和组织了解其他地区在实现SDGs方面的经验和做法。提取有关在一些条件下，有效现实SDGs的方法论及一些思路等有效建议。如果对模型进行稍微修改加入大量相关文章，甚至可以根据其他情况相似国家的发展方针，对一些较为落后的国家发展提供一些有效的建议，指导等。当该模型扩展到大规模数据加上不断的复杂化，甚至可以制作国家政策打分系统等。

1. 可以使用更为官方且全面的原始数据

使用NLP技术可以更大可能的获取更为全面的数据。如果选用非NLP技术去建模，需要选取基于数字的原始数据。目前，很难找到可以涉及17个SDGs还有各个大洲及各种发展程度的国家/地区的原始数据。相比而言，该模型可以通过收集的文章多样性来达到分析更为全面的目的。

劣势：

1. 语义鸿沟

语义鸿沟是指不同文化、语境或背景之间的语言差异和误解，这可能会对Word2Vec在SDGs之间关系分析的准确性产生一定影响。由于Word2Vec是一种基于统计的自然语言处理方法，因此它需要大量的训练数据来捕捉词语之间的语义关系。但是，不同的语言会有不同的词汇和语法规则，这可能会导致训练数据的不同，进而影响Word2Vec模型的准确性。并且，SDGs的内容涉及到许多不同领域的知识，如经济、环境、社会等。因此，在进行SDGs之间关系分析时，必须对各个领域的专业术语和背景有深入的了解，以避免因语境误解而影响Word2Vec模型的准确性。除此之外，由于SDGs的复杂性和多样性，可能存在某些SDGs之间关系的数据稀疏性，即某些关系没有足够的数据支持。这可能会导致Word2Vec模型无法捕捉到这些关系的准确性，从而影响分析结果的准确性。

为了避免过大的语义鸿沟问题，我们选择使用大量的、多样化的Google Scholars及ECOSOC发布的其他文章来作为训练数据来训练Word2Vec模型，以尽可能减少语言差异对模型准确性的影响。并且我们有去了解SDGs各领域的专业术语和背景，并对文本进行了一点人为干预，以确保模型可以正确地捕捉语境。

1. 语料库的限制

Word2Vec需要大量的语料库来训练模型，但是在SDGs的领域，相关的语料库可能相对较少，因此，Word2Vec在SDGs的应用可能会受到语料库的限制。并且语料库的限制可能会影响Word2Vec在SDGs之间关系分析的准确性。

1. 模型的泛化能力

Word2Vec的模型训练是基于具体的语料库进行的，它的泛化能力可能受到训练数据的局限。因此，Word2Vec在分析SDGs之间关系时，可能会存在泛化能力不足的问题。

敏感性分析 （不太懂，慢慢查

附录:

