Labor Space: A Unifying Representation of the Labor Market via Large Language Models

[1]

Abstract

The labor market is a complex ecosystem comprising diverse, inter- connected entities, such as industries, occupations, skills, and firms. Due to the lack of a systematic method to map these heterogeneous entities together, each entity has been analyzed in isolation or only through pairwise relationships, inhibiting comprehensive under- standing of the whole ecosystem. Here, we introduce Labor Space, a vector-space embedding of heterogeneous labor market entities, derived through applying a large language model with fine-tuning. Labor Space exposes the complex relational fabric of various labor market constituents, facilitating coherent integrative analysis of industries, occupations, skills, and firms, while retaining type-specific clustering. We demonstrate its unprecedented analytical capacities, including positioning heterogeneous entities on an economic axes, such as "Manufacturing-Healthcare". Furthermore, by allowing vec- tor arithmetic of these entities, Labor Space enables the exploration of complex inter-unit relations, and subsequently the estimation of the ramifications of economic shocks on individual units and their ripple effect across the labor market. We posit that Labor Space provides policymakers and business leaders with a comprehensive unifying framework for labor market analysis and simulation, fostering more nuanced and effective strategic decision-making.

Summary

Performed the following procedures to determine their "Labor Space" is the space the vectors of jobs/professions occupy:

1. fine-tune the original BERT model by:
2. use HuggingFace's “fill mask” pipeline for context learning. Here, the context learning aims to adjust the pre-trained model to the context of the labor market, through additional training with a domain-specific corpus for each entity (see data section for the domain-specific data)
3. set the maximum token length to 512 and configured the hyperparameters for three epochs, using a batch size of 8 and a learning rate of 2e-5
4. process the textual descriptions of the entities, using the BERT's tokenizer function. The BERT tokenizer, known as Wordpiece, encodes raw text data into token sequences and maps these tokens to their respective token IDs.
5. BERT then maps these token sequences to a matrix, where each row comprises 768-dimensional vectors representing each token ID
6. we compute a linear combination of individual word vectors. This is achieved by summing the embeddings of all the words in the sequence and dividing by the word count, thus capturing the overall semantic essence of the description.

Other potential insights are the use of the vector's to see how teams compare to each other or find their difference when transitioning over time through simple vector subtraction. Might be able to look at summing vectors overtime.

Data

Took data from NAICS, ONET, ESCO, Crunchbase

Methods

See summary

References

[1] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models, 2023.