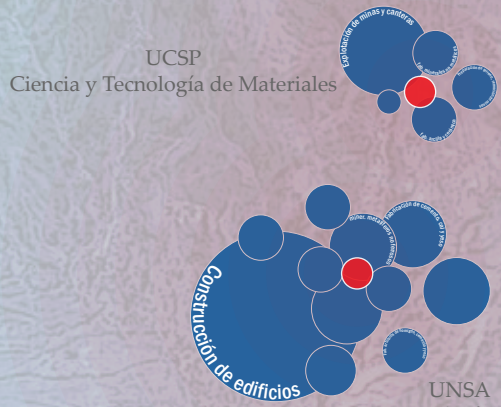
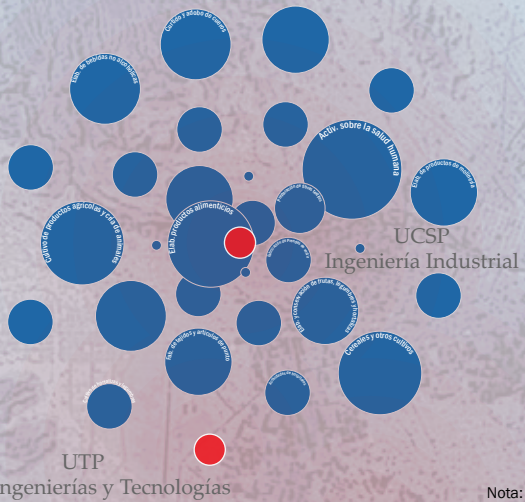
Potencial de vinculación entre los Centros de Investigación y las industrias de la región de Arequipa, Perú





Nota: La cercanía entre industrias y los centros de investigación se calculó utilizando herramientas de inteligencia artificial y análisis semántico (Embedding de OpenAI). Se mide la distancia semántica entre los títulos de los proyectos de investigación de los centros y la descripción completa de las actividades económicas de las empresas. La representación visual se realiza con el algoritmo t-SNE.
Fuentes: Se utiliza información actualizada a 2023 de los centros de investigación autorizados bajo la Ley 30309 y los datos de 2015 de la estructura empresarial en Arequipa bajo la clasificación CIIU rev. 3.

iAXTA

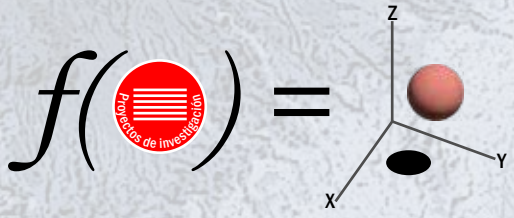
¿Cómo se calcularon las distancias?

Se extrae la informacion de los proyectos de investigación aplicada con presupuesto mayor a 100 000 soles de cada centro de investigación. Se han considerado solo a los centros autorizados por CONCYTEC bajo la ley que permite la exoneración de impuestos por actividades de investigación (Ley 30309).

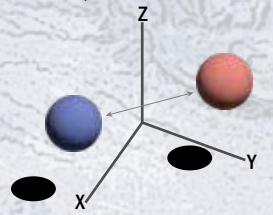
Se obtiene la información de las empresas medianas y grandes en Arequipa del censo de empresas de 2015 del Ministerio de la Producción. Este censo contiene la actividad realizada por cada empresa bajo la clasificación CIIU rev 3. Se ha utilizado la descripción detallada de las actividades. Además, el censo registra el rango de ventas anuales en el que se ubica la empresa. La información en rangos representa un problema para la agregación a nivel industrial. Se ha utilizado la marca de clase en cada rango de ventas.

La información de los proyectos de cada centro y la descripción de las actividades de cada industria corresponden a objetos de texto.

Para medir la cercanía entre objetos, se procede a representar a cada objeto de texto como un punto en un espacio vectorial semántico (Embeddings). Se utiliza el modelo text-embedding-ada-002 de OpenAI cuya salida son vectores de 1 536 dimensiones. A pesar de que el espacio semántico utilizado no ha sido diseñado para una tarea específica como la de agrupación de disciplinas científicas y campos industriales, los resultados son aceptables. Una tarea pendiente es la generación de este tipo de espacios semánticos específicos.

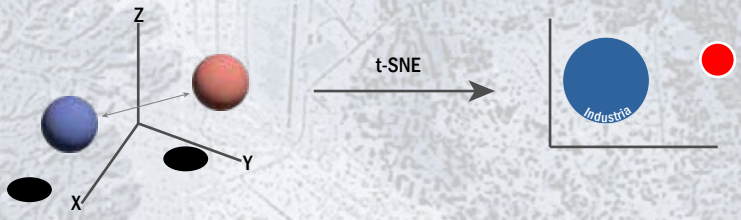


Se miden las distancias entre objetos en el espacio vectorial mediante la distancia del coseno (ampliamente utilizada en análisis semántico).



$$Distance_{cos}(a,b) = \frac{\langle a,b \rangle}{\|a\|\|b\|}$$

Finalmente, el espacio vectorial de 1 536 dimensiones es simplificado con el algoritmo t-SNE en 2 dimensiones con fines de visualización.



Fuentes de información:
Centros de investigación: https://www.datosabiertos.gob.pe/dataset/centros-de-investigaci%C3%B3n-autorizados-vigentes-por-concytec-2023-1 y portales de transparencia de cada universidad.
Censo de empresas: https://www.produce.gob.pe/index.php/datosabiertos
Clasificación CIIU rev. 3: https://unstats.un.org/unsd/publication/seriesm/seriesm_4rev3_1s.pdf

iAxta