Principio

Para identificar la similitud entre los proyectos basándonos en su descripción, entendiéndose como tal su fichero README, utilizamos el "análisis semántico latente" (LSA, usando la abreviatura en inglés), que es una técnica ampliamente utilizada en el procesamiento del lenguaje natural. LSA transforma cada texto en un vector, en un espacio de características. En nuestro caso, las características son palabras que ocurren en las descripciones. A continuación, se crea una matriz que contiene todos los vectores: las columnas representan las descripciones de los proyectos y las filas representan palabras únicas. Por consiguiente, el número de filas puede ascender a decenas de miles de palabras.

Con el fin de identificar las características relevantes de esta matriz, la "descomposición de valores singulares" (SVD, usando la abreviatura en inglés), que es una técnica de reducción de dimensión, se utiliza para reducir el número de líneas -palabras-, manteniendo y resaltando la similitud entre columnas-descripción -. La dimensión de esta matriz de aproximación se establece mediante un hiperparámetro que es el número de temas, comúnmente llamado como tópicos. En este marco, un tópico consiste en un conjunto de palabras con pesos asociados que definen la contribución de cada palabra a la dirección de este tópico. Basándose en esta matriz de aproximación de baja dimensión, la similitud entre dos columnas -descripciones- se calcula utilizando el coseno del ángulo entre estos dos vectores.