Dataset "Commit Messages"

Qué	identificar patrones y tendencias en los mensajes de commit para entender mejor las prácticas de desarrollo en repositorios de código abierto.
Tipos de relaciones	Explorar correlaciones entre la frecuencia y naturaleza de los mensajes de commit, identificar patrones temporales, como momentos de alta actividad.
Predicciones o asociaciones	Inicialmente, me centraría en identificar asociaciones y patrones. Una vez establecida esta base, podría desarrollar modelos predictivos.
Tipo de datos y de información hay en cada columna	El dataset contiene hashes de commit, nombres de repositorios y sus respectivos mensajes de commit.
En caso de que haya varias tablas, ¿cómo se relacionan?	Cada hash de commit en commits.bin tiene un correspondiente nombre de repositorio en repos.txt.xz y un mensaje de commit en messages.txt.xz.
¿Necesita limpiar, agregar o procesar los datos antes de poder usarlos?	Sí, para iniciar, se incluirían acciones como la eliminación de duplicados y el manejo de valores nulos.

https://github.com/src-d/datasets/tree/master/CommitMessages