Textové datové kolekce vytvářejí prostor pro nové úlohy využívající algoritmy z umělé inteligence. Tyto kolekce často obsahují různé osobní údaje a jiné citlivé informace, které komplikují jejich sdílení a další zpracování kvůli požadavkům na ochranu osobních údajů. Hledání osobních údajů je často řešeno pouze postupným procházením celého textu. Práce si proto klade za cíl vytvořit nástroj, který napomůže anotátorům snižovat riziko úniku osobních údajů z textových datových kolekcí. Nástroj pro snížení rizika využívá pseudonymizace (tj. nahrazování slov jinými slovy pomocí nějakého klíče). V průběhu anotačního procesu nástroj automaticky označuje slova jako „veřejná“, „soukromá“ a jako „podezřelá“. Úkolem anotátora je rozhodovat o „podezřelých slovech“ a dohledávat případné chybějící neoznačené citlivé informace. „Soukromá“ slova jsou poté předmětem procesu pseudonymizace. Nástroj k automatickému označovaní využívá rozpoznávač pojmenovaných entit a databázi pravidel. Databáze pravidel se sama průběžně vylepšuje při některých rozhodnutích anotátora. V rámci práce došlo k porovnání různých rozpoznávačů pojmenovaných entit pro účel vyhledávání osobních údajů na kolekci z projektu ELITR. Při porovnávání byla nalezena metoda, která zvýšila citlivost detekce pojmenovaných entit a tím i zvýšila citlivost detekce „podezřelých slov“. Součástí práce je úvod do právní problematiky ochrany osobních údajů při vytváření textových kolekcí.