Near Duplicates

Near duplicates occur when there are only minor differences between multiple versions of a document, such as:

- Files that use the same template (e.g. bills or order forms)
- The same file changed over time (drafts and final versions)
- Original and a forwarded version of the same email

This is an issue because as humans we group near-duplicates together for investigation purposes, but simple machine deduplication techniques will not.