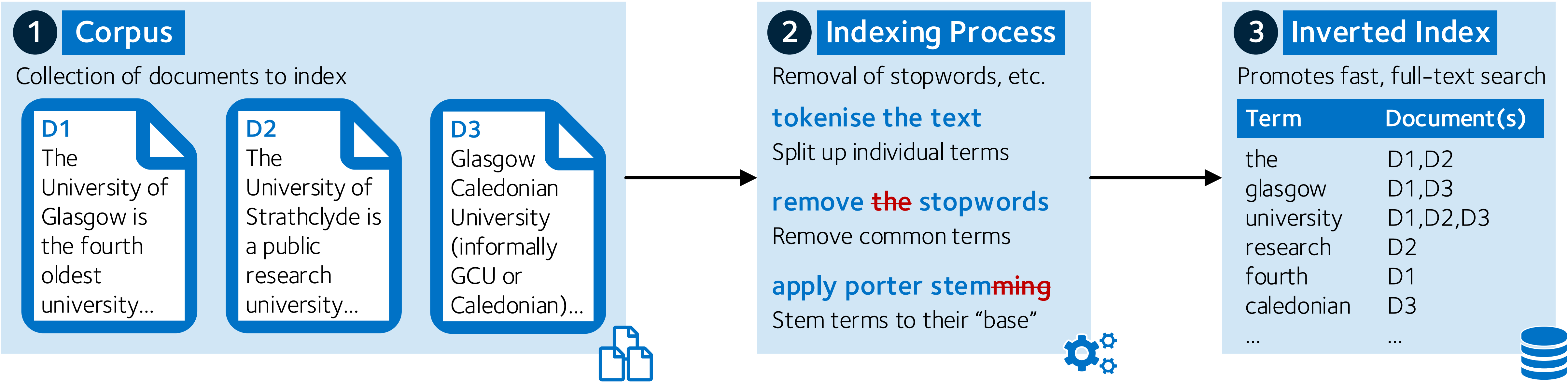

1 Corpus

Collection of documents to index

D1 The University of Glasgow is the fourth oldest university…

D2 The University of Strathclyde is a public research university…

D3 Glasgow Caledonian University (informally GCU or Caledonian)…

2 Indexing Process

Removal of stopwords, etc.

tokenise the text
Split up individual terms

remove ~~the~~ stopwords
Remove common terms

apply porter stem~~ming~~
Stem terms to their “base”

3 Inverted Index

Promotes fast, full-text search

Term	Document(s)
the	D1,D2
glasgow	D1,D3
university	D1,D2,D3
research	D2
fourth	D1
caledonian	D3
…	…