library(textreuse)

minhash <- minhash\_generator(n = 240, seed = 3552)

head(minhash(c("turn tokens into", "tokens into hashes", "into hashes fast")))

dir <- system.file("extdata/ats", package = "textreuse")

corpus <- TextReuseCorpus(dir = dir, tokenizer = tokenize\_ngrams, n = 5,

minhash\_func = minhash, keep\_tokens = TRUE,

progress = FALSE)

head(minhashes(corpus[[1]]))

length(minhashes(corpus[[1]]))

lsh\_threshold(h = 200, b = 50)

lsh\_threshold(h = 240, b = 80)

lsh\_probability(h = 240, b = 80, s = 0.25)

lsh\_probability(h = 240, b = 80, s = 0.75)

buckets <- lsh(corpus, bands = 80, progress = FALSE)

buckets

baxter\_matches <- lsh\_query(buckets, "calltounconv00baxt")

baxter\_matches

candidates <- lsh\_candidates(buckets)

candidates

lsh\_compare(candidates, corpus, jaccard\_similarity, progress = FALSE)