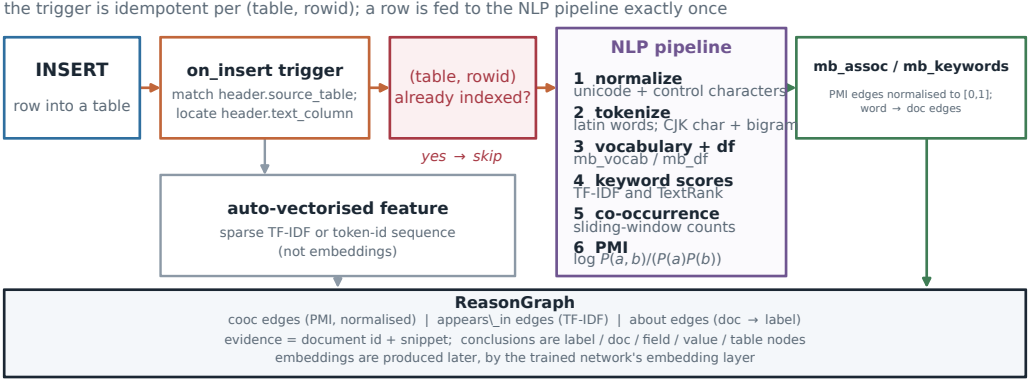

the trigger is idempotent per (table, rowid); a row is fed to the NLP pipeline exactly once
INSERT
row into a table
on_insert trigger
match header.source_table;
locate header.text_column
(table, rowid)
already indexed?
yes → skip
NLP pipeline
1 normalize
unicode + control characters
2 tokenize
latin words; CJK char + bigram
3 vocabulary + df
mb_vocab / mb_df
4 keyword scores
TF-IDF and TextRank
5 co-occurrence
sliding-window counts
6 PMI
log P(a, b)/(P(a)P(b))
mb_assoc / mb_keywords
PMI edges normalised to [0,1];
word → doc edges
auto-vectorised feature
sparse TF-IDF or token-id sequence
(not embeddings)
ReasonGraph
cooc edges (PMI, normalised) | appears_in edges (TF-IDF) | about edges (doc → label)
evidence = document id + snippet; conclusions are label / doc / field / value / table nodes
embeddings are produced later, by the trained network's embedding layer