id **content**

[web page 1][text1, text2, …]

[web page 2][text1, text2, …]

⋮

Train doc2vec model on Content Token Sequence

id **content+tags**

[web page 1] [tag1, text1, …]

[web page 2] [tag1, text2, …]

⋮

Train doc2vec model on Content+Tags Token Sequence

id **tags**

[web page 1] [tag1, tag2, …]

[web page 2] [tag1, tag2, …]

⋮

Train doc2vec model on Tags Token Sequence

Extract Content

Content Token Sequence

Collect Content + Tags

Content+Tags Token Sequence

training set

Extract Tags

Tags Token Sequence

PHASE 1 - TRAINING DOC2VEC

training set (labeled)

Compute Embeddings → **Compute Similarities** → **Train Classifier** → SAF

PHASE 2 - TRAINING STATE ABSTRACTION FUNCTION

Crawler

web page A   embedding A

web page B   embedding B

similarity SAF

✓ Distinct

! Duplicate

PHASE 3 - CRAWLING

Web App Crawl Model → **Crawl Segmentation** → Test Cases

PHASE 4 - TEST CREATION