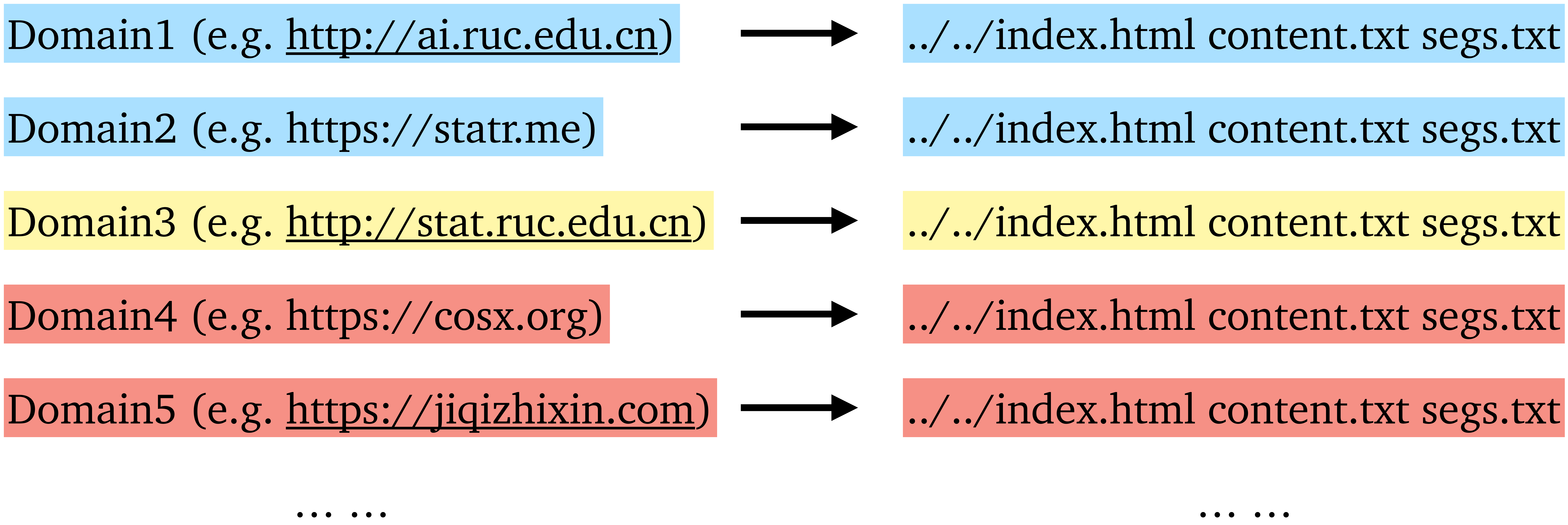
基于字典的 tf-idf & 字符串搜索系统

——《人工智能综合设计》课程报告

王胤博 2024.9.2

Walk-Through



Domain1 (e.g. http://ai.ruc.edu.cn) → ../../index.html content.txt segs.txt

Domain2 (e.g. https://statr.me) → ../../index.html content.txt segs.txt

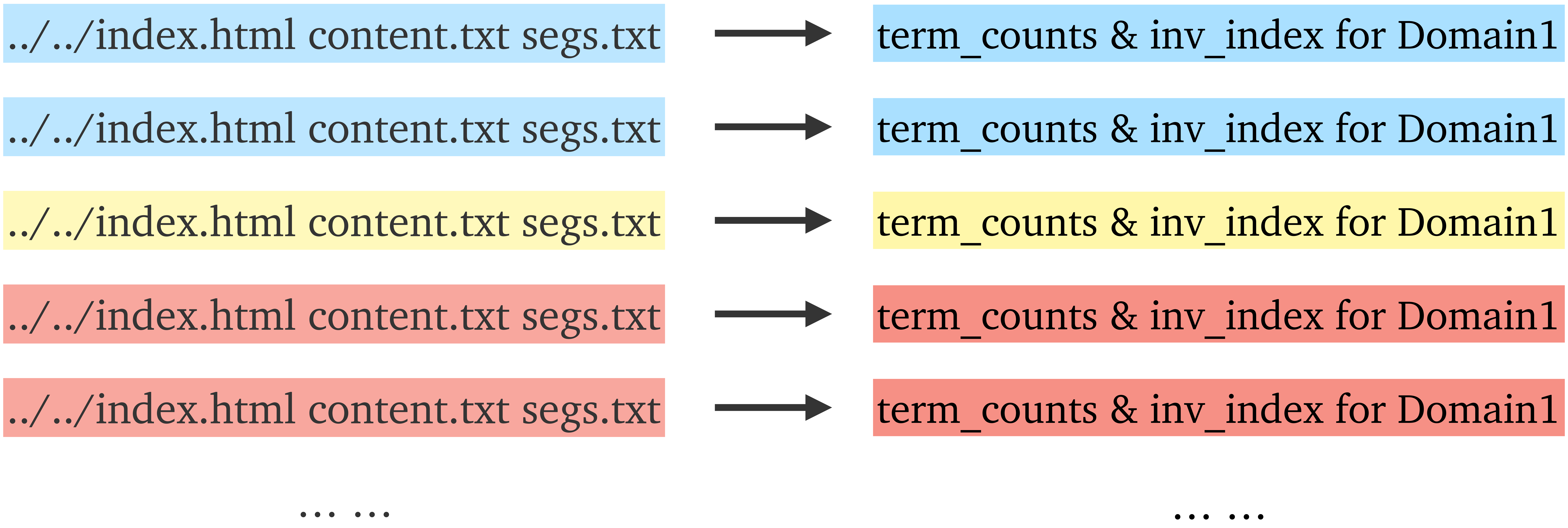
Domain3 (e.g. http://stat.ruc.edu.cn) → ../../index.html content.txt segs.txt

Domain4 (e.g. https://cosx.org) → ../../index.html content.txt segs.txt

Domain5 (e.g. https://jiqizhixin.com) → ../../index.html content.txt segs.txt

… … → … …

Walk-Through



../../index.html content.txt segs.txt → term_counts & inv_index for Domain1

../../index.html content.txt segs.txt → term_counts & inv_index for Domain1

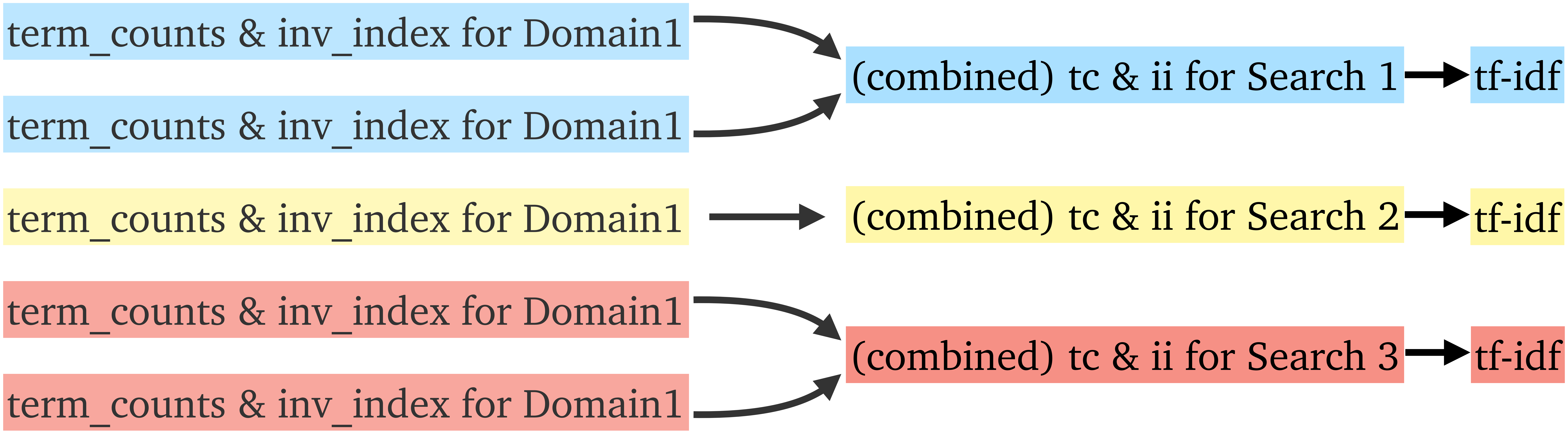
../../index.html content.txt segs.txt → term_counts & inv_index for Domain1

../../index.html content.txt segs.txt → term_counts & inv_index for Domain1

../../index.html content.txt segs.txt → term_counts & inv_index for Domain1

… … … …

Walk-Through



term_counts & inv_index for Domain1

term_counts & inv_index for Domain1

(combined) tc & ii for Search 1

tf-idf

term_counts & inv_index for Domain1

(combined) tc & ii for Search 2

tf-idf

term_counts & inv_index for Domain1

term_counts & inv_index for Domain1

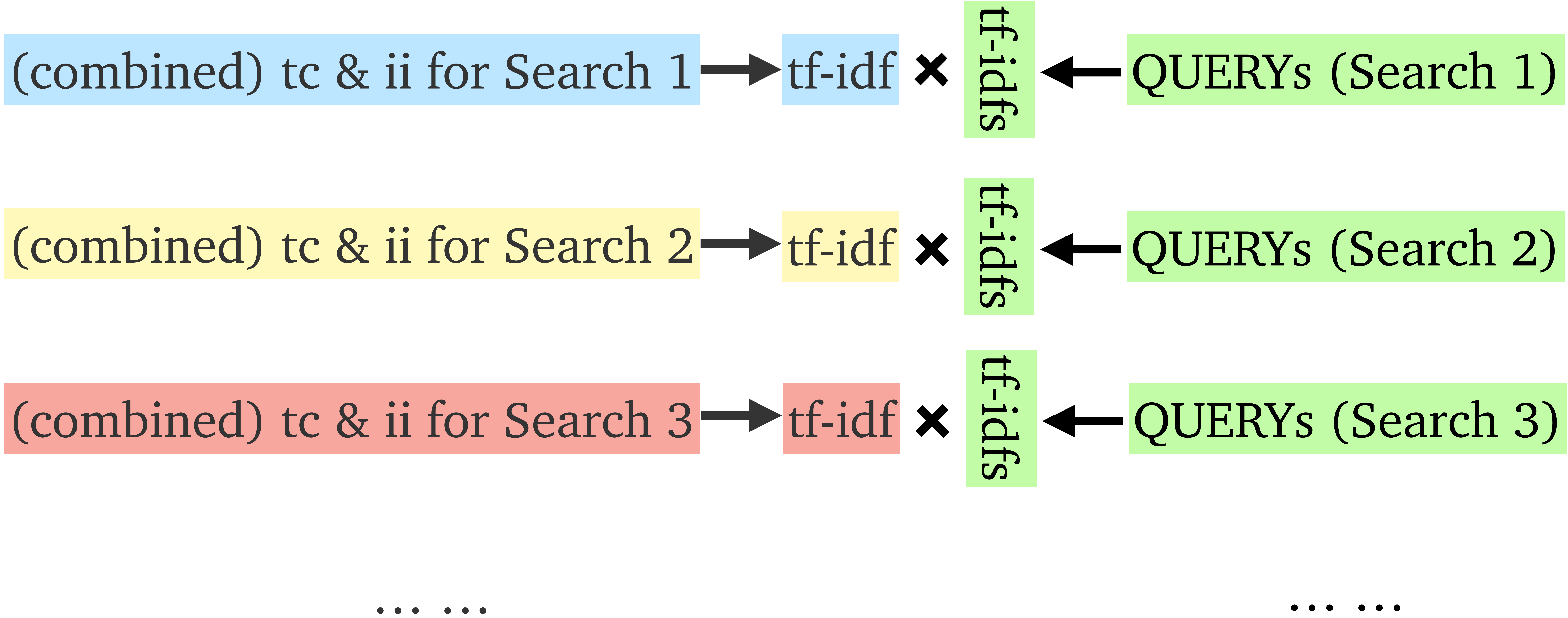
(combined) tc & ii for Search 3

tf-idf

… …

… …

Walk-Through



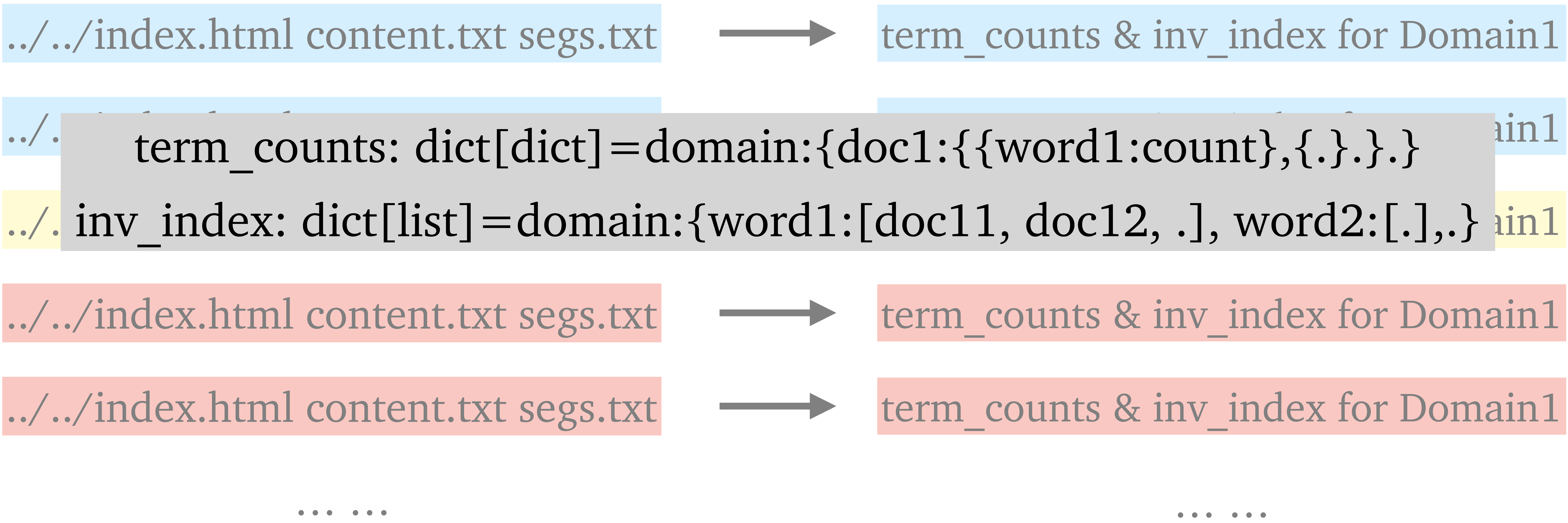
(combined) tc & ii for Search 1 → tf-idf × tf-idfs ← QUERYs (Search 1)

(combined) tc & ii for Search 2 → tf-idf × tf-idfs ← QUERYs (Search 2)

(combined) tc & ii for Search 3 → tf-idf × tf-idfs ← QUERYs (Search 3)

… … … …

Dict-base



../../index.html content.txt segs.txt → term_counts & inv_index for Domain1

../. ain1

term_counts: dict[dict]=domain:{doc1:{{word1:count},{.}.}.}

inv_index: dict[list]=domain:{word1:[doc11, doc12, .], word2:[.],.}

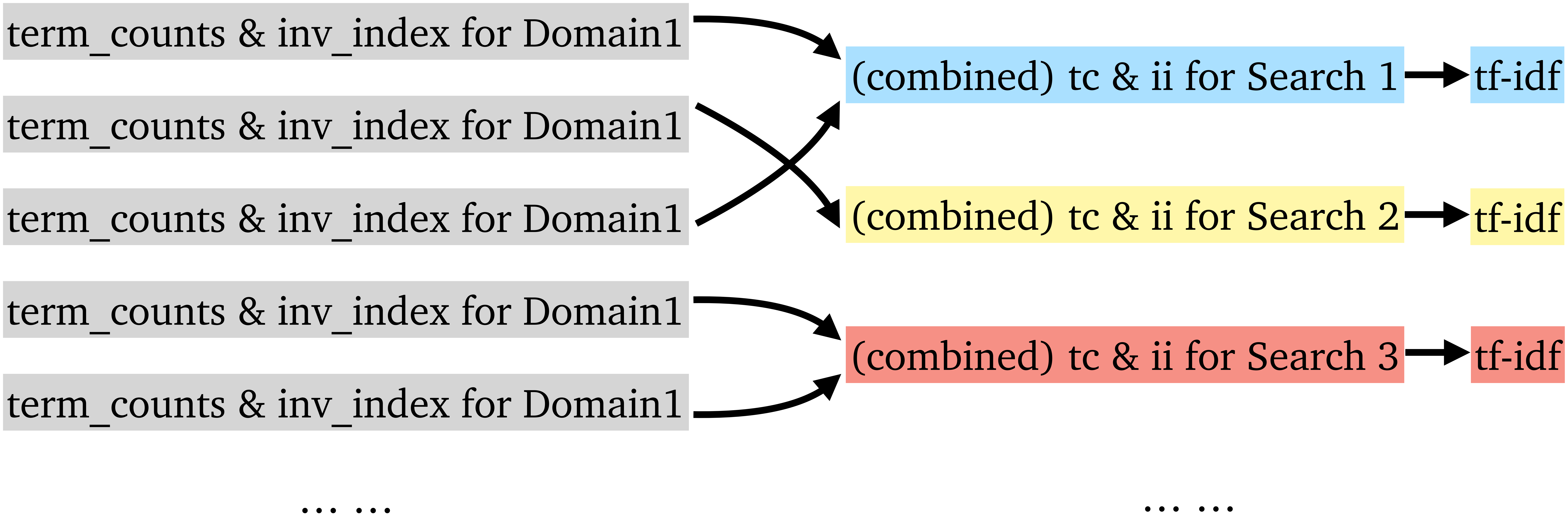
../. ain1

../../index.html content.txt segs.txt → term_counts & inv_index for Domain1

../../index.html content.txt segs.txt → term_counts & inv_index for Domain1

… … … …

Domain-base



term_counts & inv_index for Domain1

term_counts & inv_index for Domain1

term_counts & inv_index for Domain1

term_counts & inv_index for Domain1

term_counts & inv_index for Domain1

… …

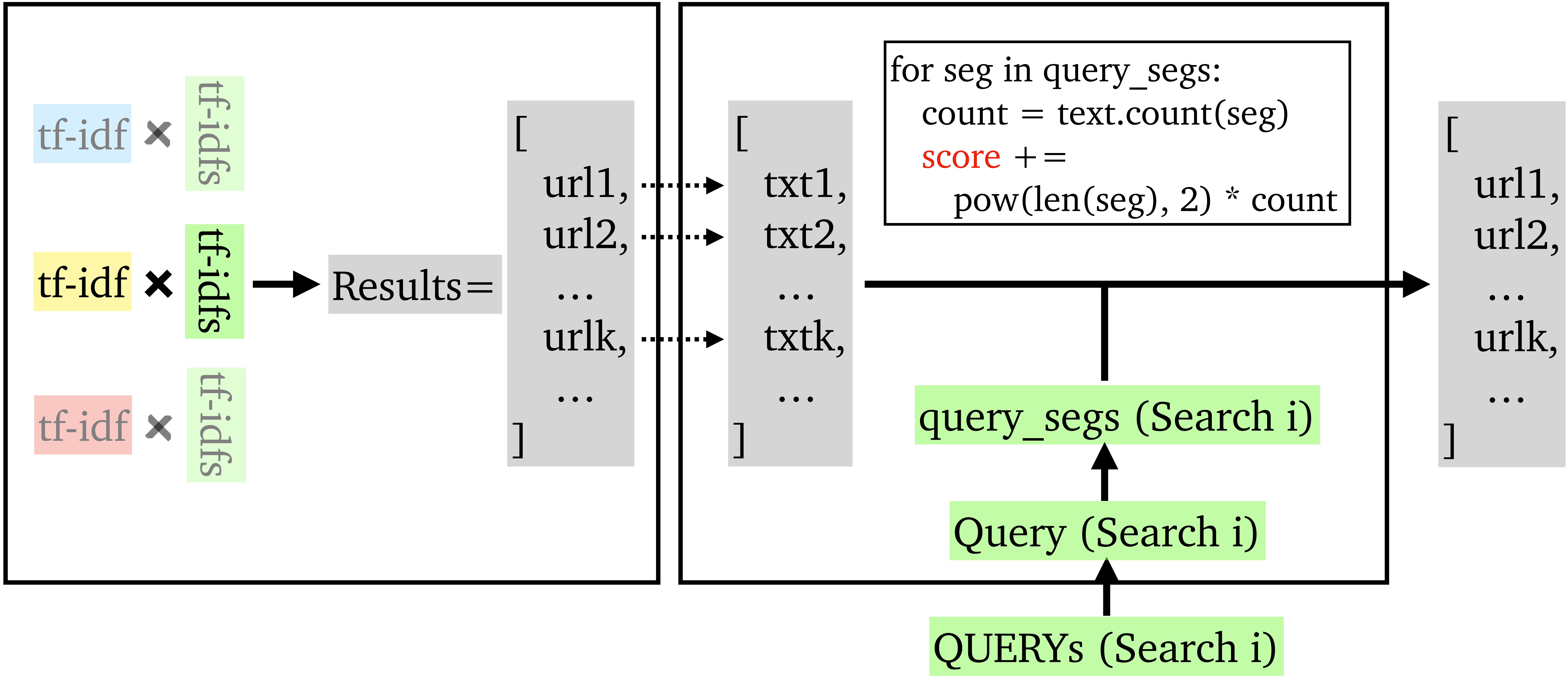
(combined) tc & ii for Search 1 → tf-idf

(combined) tc & ii for Search 2 → tf-idf

(combined) tc & ii for Search 3 → tf-idf

… …

Search



```
for seg in query_segs:
   count = text.count(seg)
   score +=
      pow(len(seg), 2) * count
```