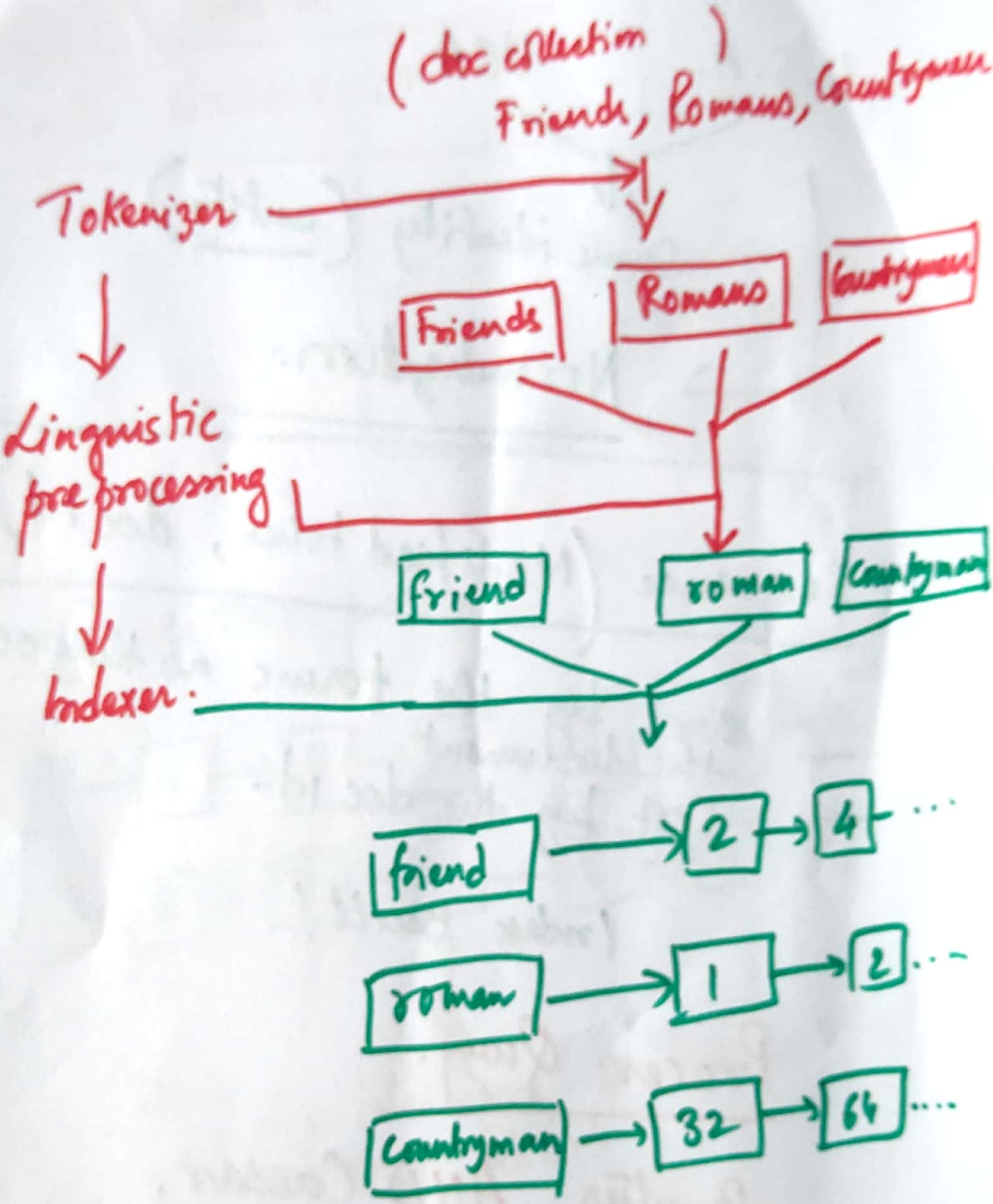
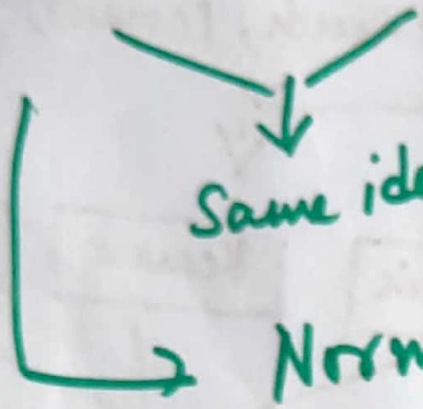


Tokenization: cutting/splitting chars into "word tokens" ← terms.

Language Model: a) Lower casing
b) stemming ←

U.S.A. USA



Same identity (entity)

Normalization:

Sequence (Modified token, doc id).

- sort by the terms as they occur in the document
- Sort by the doc id.

Index Built!

Process Query.

Brutus AND Caesar.

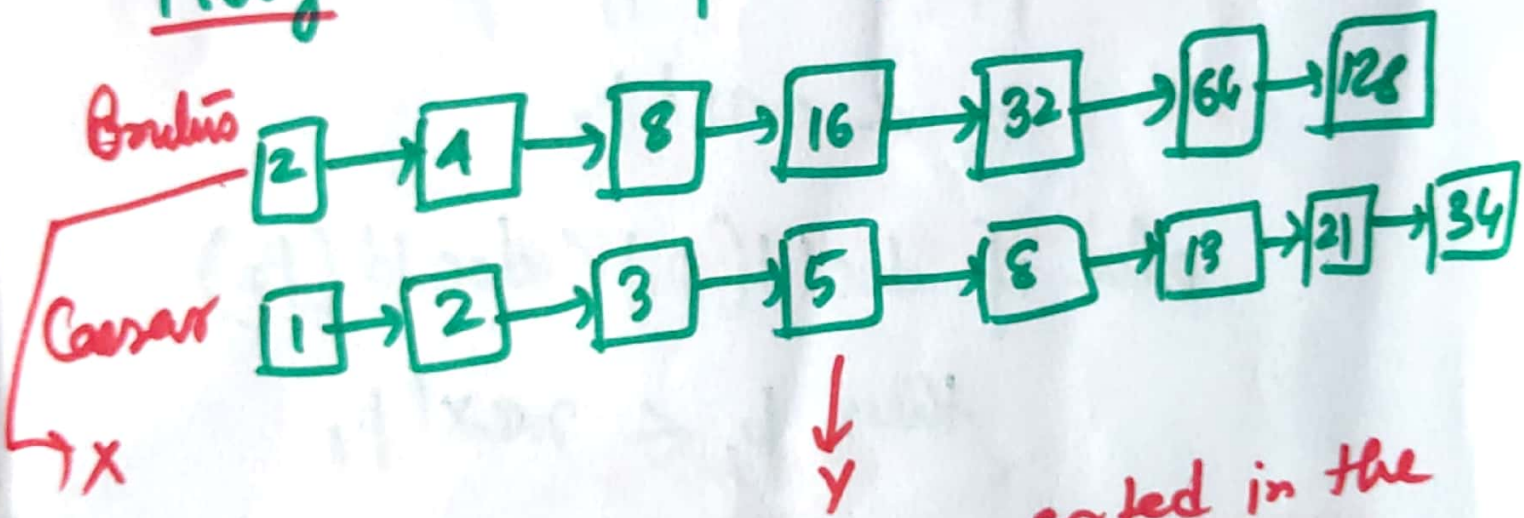
Locate Brutus. in dict

Locate Caesar in dict.

retrieve postings

retrieve postings.

"Merge" the two postings



doc ids are sorted in the posting lists.

$O(X+Y)$.

```
Intersect (p1, p2)
answer ← <>
while p1 ≠ NIL && p2 ≠ NIL
 do if docId(p1) = docId(p2)
      then ADD(answer, docId(p1))
           p1 ← next p1
           p2 ← next p2
      else if docId(p1) < docId(p2)
             then p1 ← next p1
             else p2 ← next p2
return answer.
```

Query Optimization

What is the best order in which we process our query.

query ← simple AND of some n query terms.

(document frequency of the term) # of the term across documents frequency → indicator of the size of the posting list.

two terms that have the smallest doc frequency.

OR queries?

(madding OR crowd) AND (ignoble AND strife)

Estimate the size of the OR?

worst case — sum of the doc frequencies.

PROCESS IN INCREASING ORDER OF OR SIZES

How popular do you think is Boolean retrieval?

- Not very popular.
- somewhat popular.
- used to be popular. ^ at the peak of popularity

Legal IR system:

WestLaw

www.westlaw.com.

700K users.

- What is the statute of limitations in cases involving the federal tort claims act?

LIMIT! /3 STATUTE ACTION /S

FEDERAL /2 TORT /3 CLAIM.

/3 → word windows (within 3 words)

/S → within the sentence

q → Google's default interpretation of the query $w_1 w_2 w_3 w_4 \ldots$

Default interpretation.

w_1 AND w_2 AND w_3 AND w_4 ....

When does this not hold?

(i) page contains a variant of w_i

(ii) presence of w_i in anchor text.

→ Order or ambiguity of the query.

→ sky AND blue ← color.

→ blue AND sky ← video

→ common AND wealth — games

→ wealth AND common → "common sense is your wealth"