Lect 9: Dictionaries and Tolerant Retrieval

Dr. Subrat Kumar Nayak

Associate Professor

Dept. of CSE, ITER, SOADU

Recap

Type/token distinction

- Token – an instance of a word or term occurring in a document
- Type – an equivalence class of tokens
- *In June, the dog likes to chase the cat in the barn.*
- 12 word tokens, 9 word types

Recap…

Problems in tokenization

- What are the delimiters? Space? Apostrophe? Hyphen?
- For each of these: sometimes they delimit, sometimes they don't.
- No whitespace in many languages! (e.g., Chinese)
- No whitespace in Dutch, German, Swedish compounds (*Lebensversicherungsgesellschaftsangestellter*)

Recap…

Problems with equivalence classing

- A term is an equivalence class of tokens.
- How do we define equivalence classes?
- Numbers (3/20/91 vs. 20/3/91)
- Case folding
- Stemming, Porter stemmer
- Morphological analysis: inflectional vs. derivational

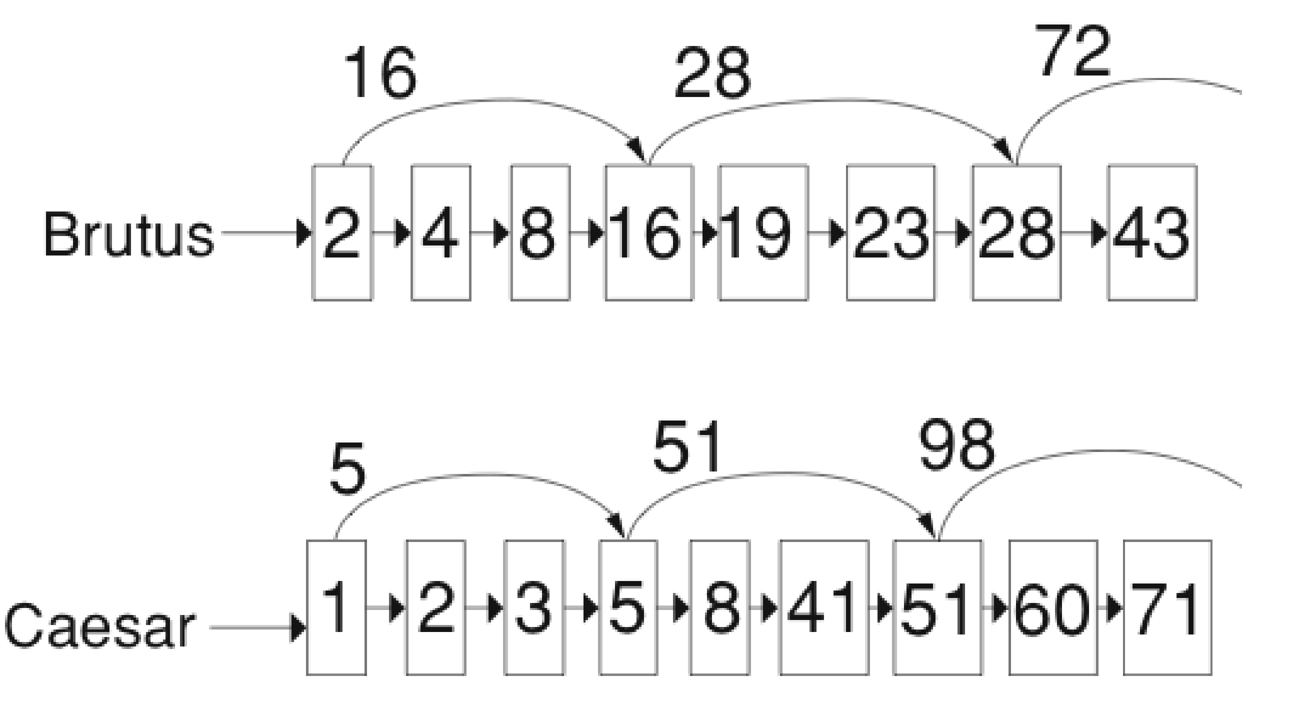
Eg: Happy to unhappy/ happyness **(Derivational)**

Determine to determines/ determining/ determined **(inflectional)**

- *Stemming* usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.
- *Lemmatization* usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma*.
- If confronted with the token *saw*, stemming might return just s, whereas lemmatization would attempt to return either see or *saw* depending on whether the use of the token was as a verb or a noun.
- Equivalence classing problems in other languages
  - More complex morphology than in English
  - Finnish: a single verb may have 12,000 different forms
  - Accents, umlauts

Recap…

Skip Pointers



Recap...

Positional indexes

- Postings lists in a nonpositional index: each posting is just a docID
- Postings lists in a positional index: each posting is a docID and a list of positions
- Example query: “to_1 be_2 or_3 not_4 to_5 be_6”
- TO, 993427:
  - ⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
  - 2: ⟨1, 17, 74, 222, 255⟩;
  - 4: ⟨8, 16, 190, 429, 433⟩;
  - 5: ⟨363, 367⟩;
  - 7: ⟨13, 23, 191⟩; . . . ⟩
- BE, 178239:
  - ⟨ 1: ⟨17, 25⟩;
  - 4: ⟨17, 191, 291, 430, 434⟩;
  - 5: ⟨14, 19, 101⟩; . . . ⟩ Document 4 is a match!

Recap…

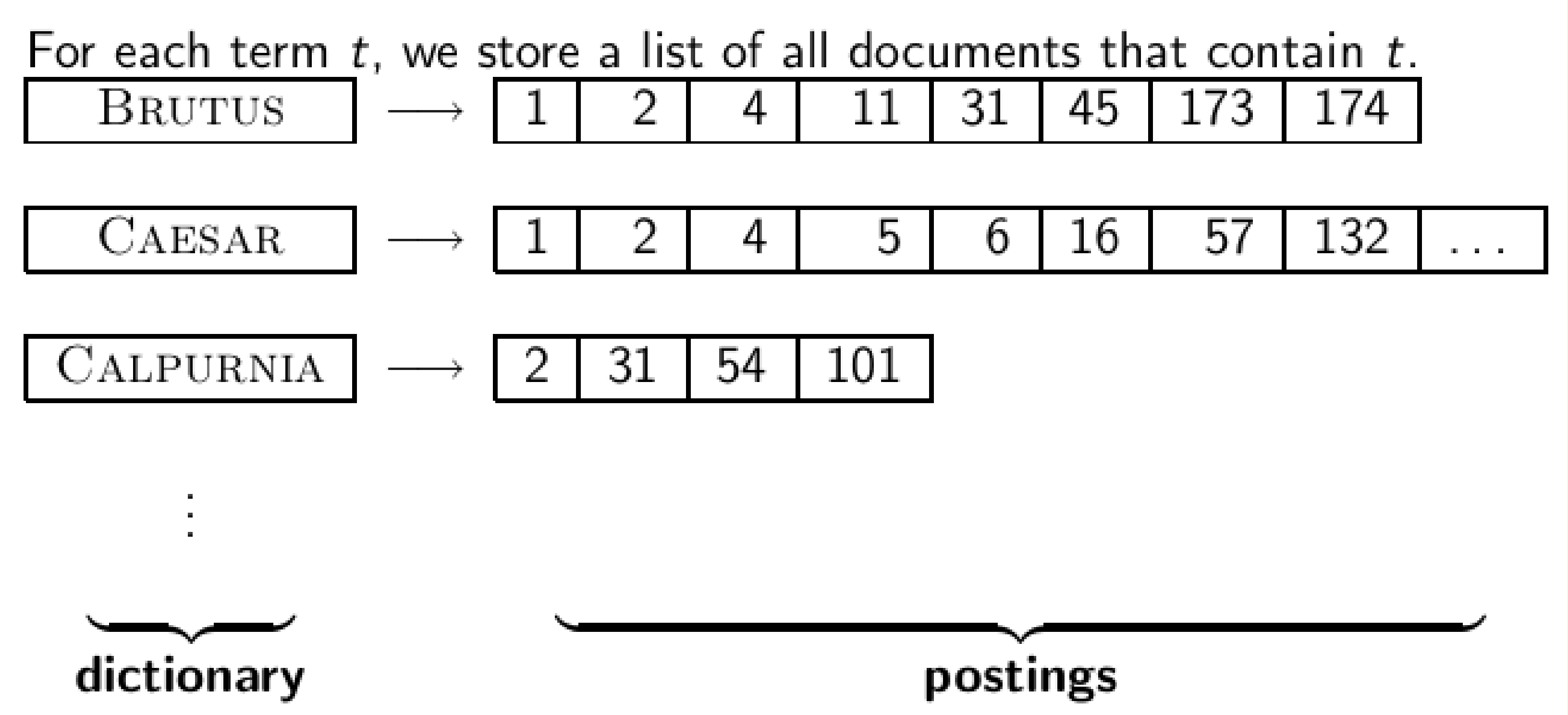
Positional indexes

- With a positional index, we can answer phrase queries.
- With a positional index, we can answer proximity queries.

Tolerant retrieval

- Tolerant retrieval: What to do if there is no exact match between query term and document term
- Wildcard queries
- Spelling correction

Inverted index



For each term t, we store a list of all documents that contain t.

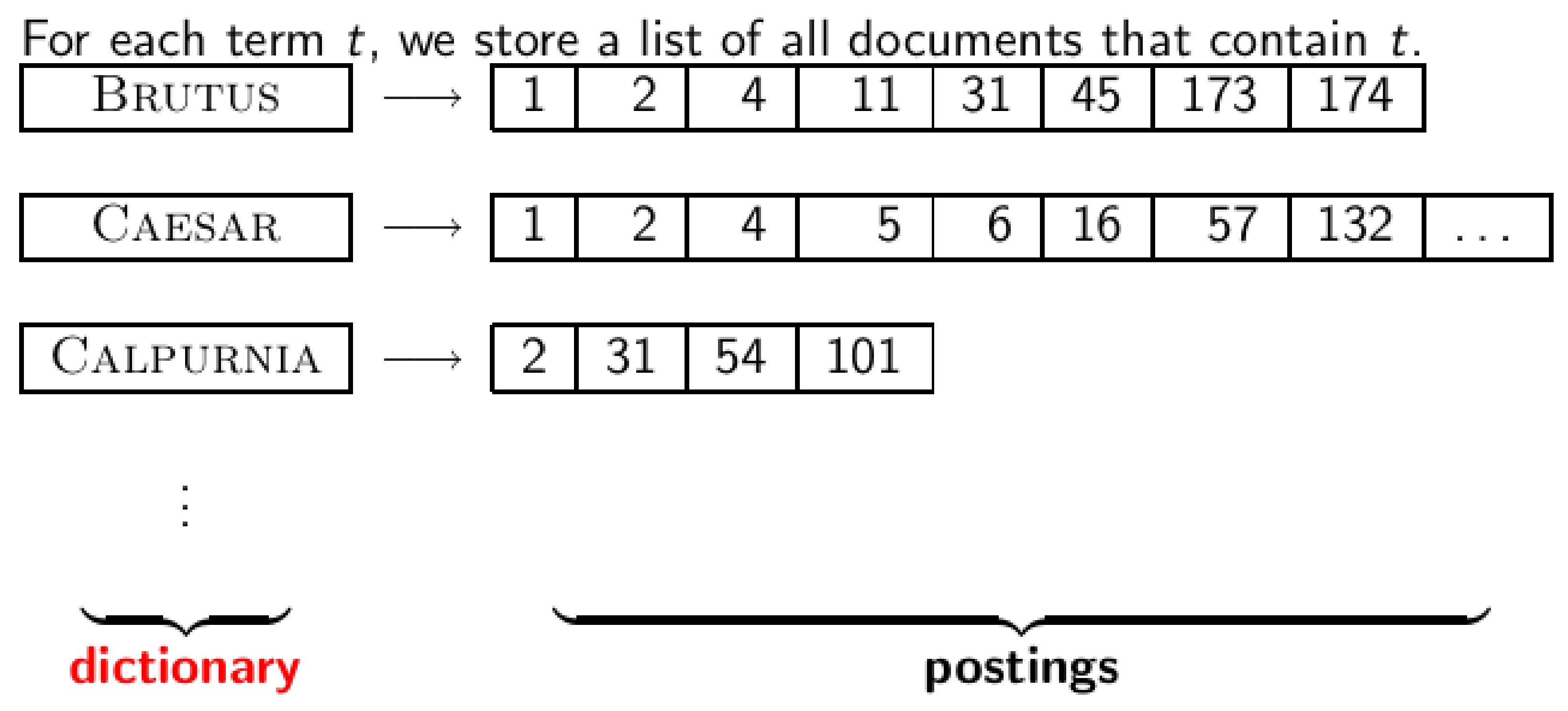
BRUTUS → 1 | 2 | 4 | 11 | 31 | 45 | 173 | 174

CAESAR → 1 | 2 | 4 | 5 | 6 | 16 | 57 | 132 | ...

CALPURNIA → 2 | 31 | 54 | 101

⋮

Inverted index



For each term t, we store a list of all documents that contain t.

Brutus $\longrightarrow$ 1 | 2 | 4 | 11 | 31 | 45 | 173 | 174

Caesar $\longrightarrow$ 1 | 2 | 4 | 5 | 6 | 16 | 57 | 132 | ...

Calpurnia $\longrightarrow$ 2 | 31 | 54 | 101

⋮

dictionary

postings

Dictionaries

- The dictionary is the data structure for storing the term vocabulary.
- Term vocabulary: the data
- Dictionary: the data structure for storing the term vocabulary

Dictionary as array of fixed-width entries

- For each term, we need to store a couple of items:
  - document frequency
  - pointer to postings list
  - . . .
- Assume for the time being that we can store this information in a fixed-length entry.
- Assume that we store these entries in an array.

Data structures for looking up term

- Two main classes of data structures: hashes and trees
- Some IR systems use hashes, some use trees.
- Criteria for when to use hashes vs. trees:
  - Is there a fixed number of terms or will it keep growing?
  - What are the relative frequencies with which various keys will be accessed?
  - How many terms are we likely to have?

Recap...