CMPE 493
INTRODUCTION TO
INFORMATION RETRIEVAL

Dictionaries

Department of Computer Engineering, Boğaziçi University
November 3, 2020

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				
⋮									

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

A naïve dictionary

- An array of struct:

term	document frequency	pointer to postings list
a	656,265	⟶
aachen	65	⟶
. . .	. . .	. . .
zulu	221	⟶
char[20]	int	Postings *
20 bytes	4/8 bytes	4/8 bytes

- How do we store a dictionary in memory efficiently?
- How do we quickly look up elements at query time?

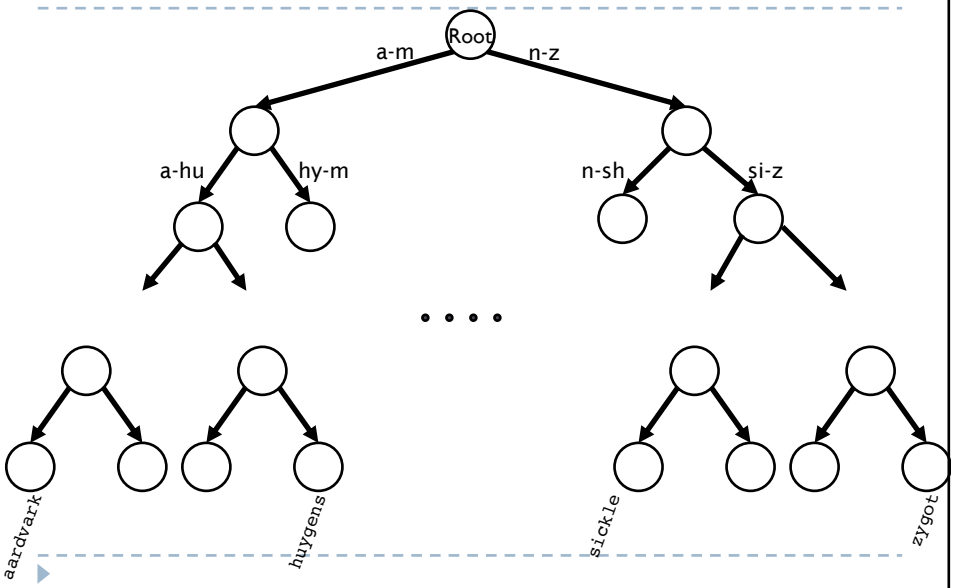
Data structures for looking up term

- Two main classes of data structures:
  - hashes
  - trees
- Some IR systems use hashes, some use trees.

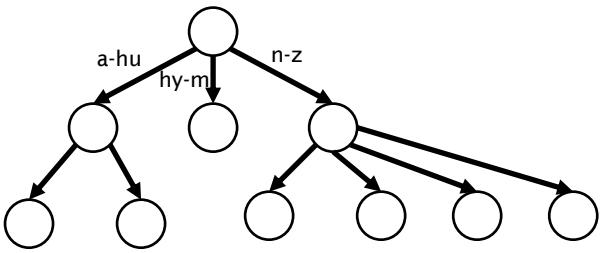
Hashes

- Each vocabulary term is hashed into an integer.
- Try to avoid collisions
- At query time, do the following: hash query term, resolve collisions, locate entry in fixed-width array
- Pros: Lookup in a hash is faster than lookup in a tree.
  - Lookup time is constant.
- Cons
  - no way to find minor variants (*resume* vs. *résumé*)
  - no prefix search (all terms starting with *automat*)
  - need to rehash everything periodically if vocabulary keeps growing

Tree: binary tree



Tree: B-tree



- Definition: Every internal node has a number of children in the interval [*a*,*b*] where *a, b* are appropriate natural numbers, e.g., [2,4].

Trees

- Simplest: binary tree
- More usual: B-trees
- Trees require a standard ordering of characters and hence strings … but we standardly have one
- Pros:
  - Solves the prefix problem (terms starting with *hyp*)
- Cons:
  - Slower: O(log *N*) [and this requires *balanced* tree]
  - Rebalancing binary trees is expensive
    - But B-trees mitigate the rebalancing problem

Trie (Keyword Tree)

- A tree-based data structure, name comes from "retrieval"
- Each node stores a letter from the alphabet and each word corresponds to a path in this tree.
- Can be used to store the Dictionary.
  - Mark nodes corresponding to end of words and store link (pointer) to the associated postings.
  - Can be used to answer prefix searches such as in Auto-complete

Creating a Trie

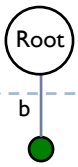
Root

Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana

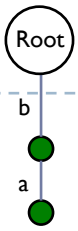
Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana



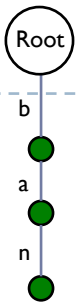
Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana



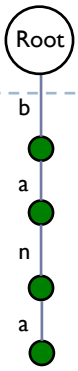
Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana



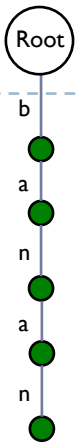
Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana



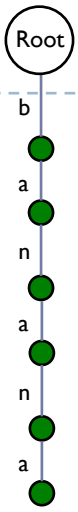
Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana



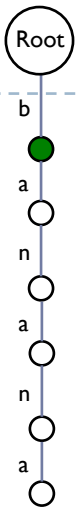
Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana



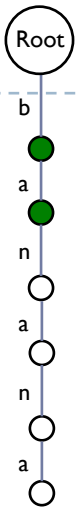
Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana



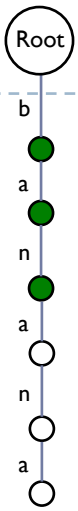
Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana



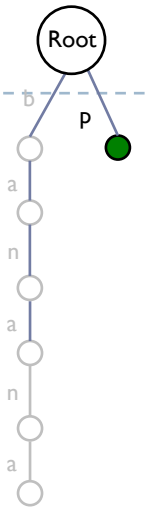
Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana



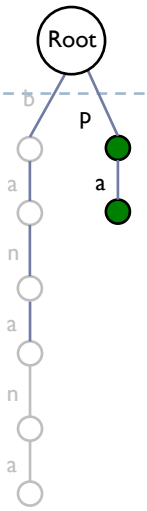
Dictionary

banana
ban
pan
nab
antenna
bandana
ananas
nana



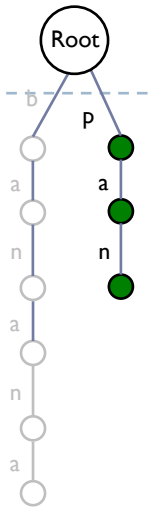
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



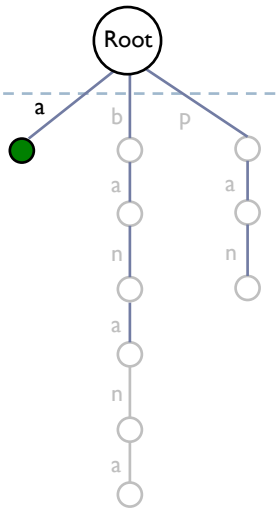
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



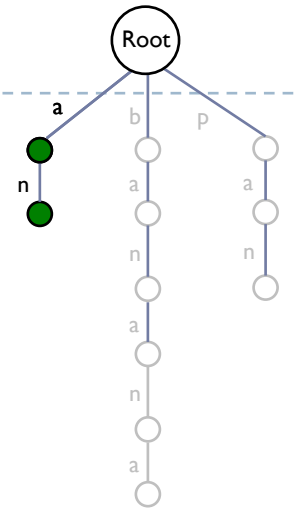
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



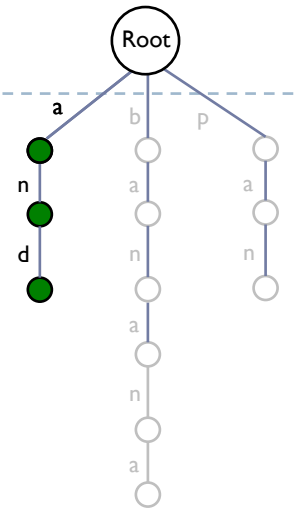
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



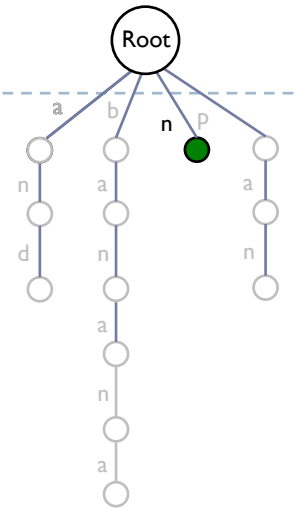
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



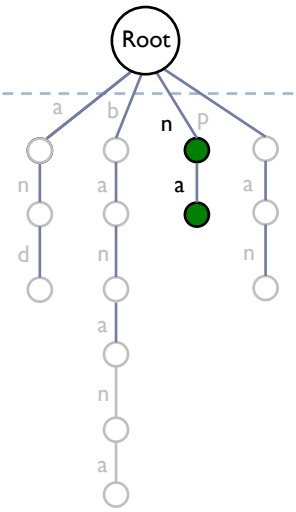
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



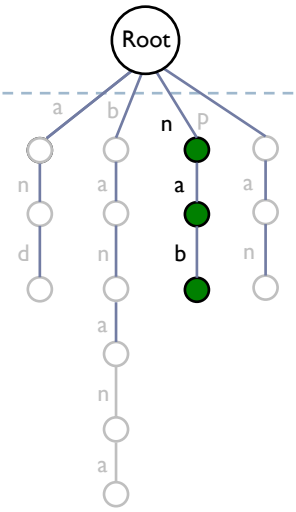
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



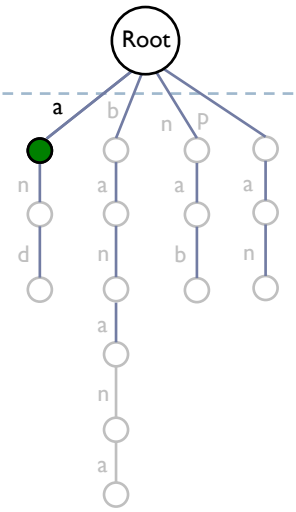
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



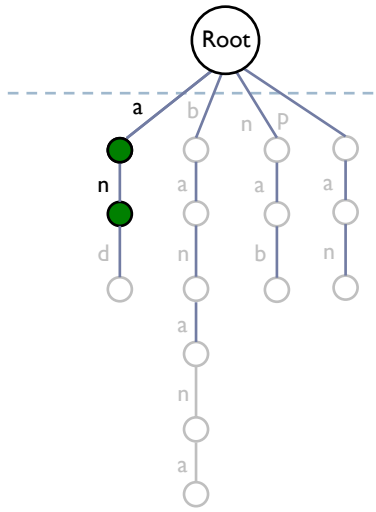
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



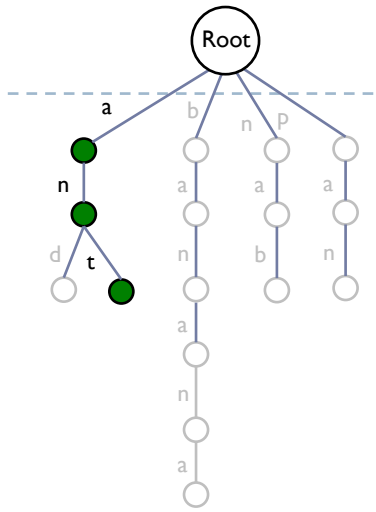
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



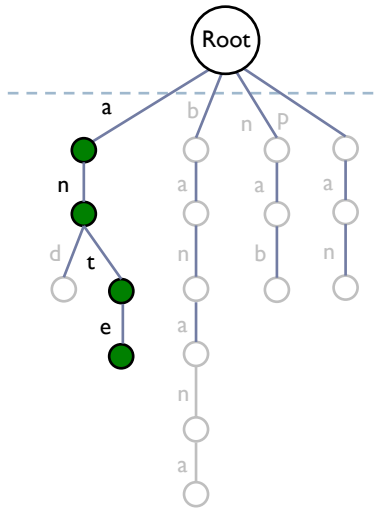
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



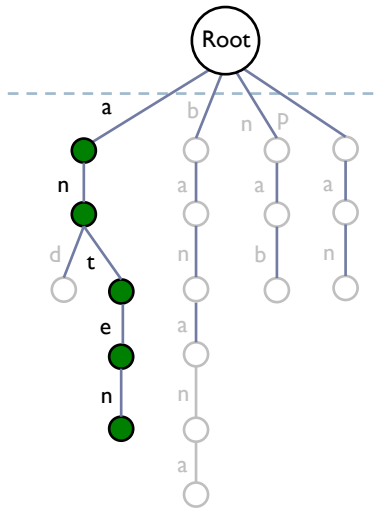
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



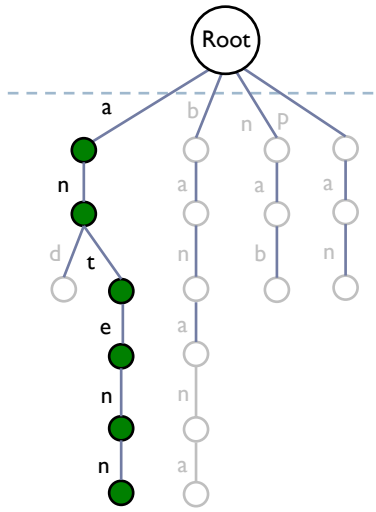
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



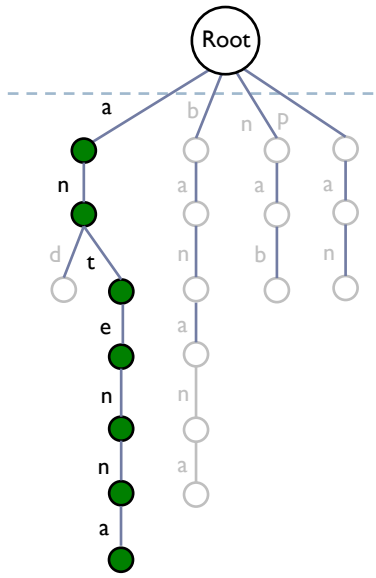
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



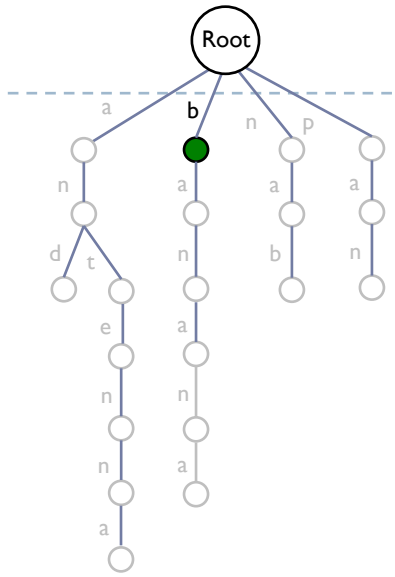
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



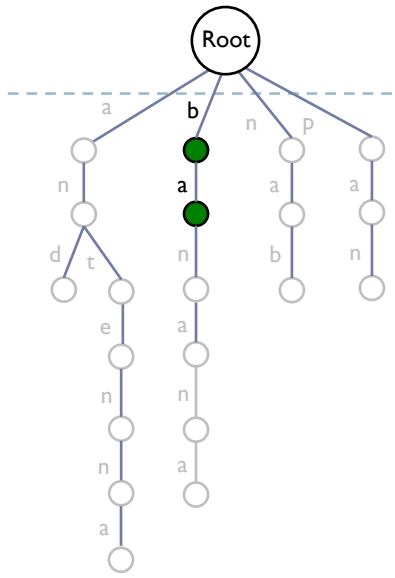
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



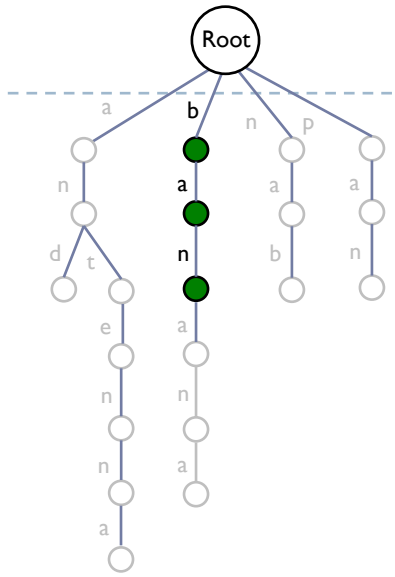
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



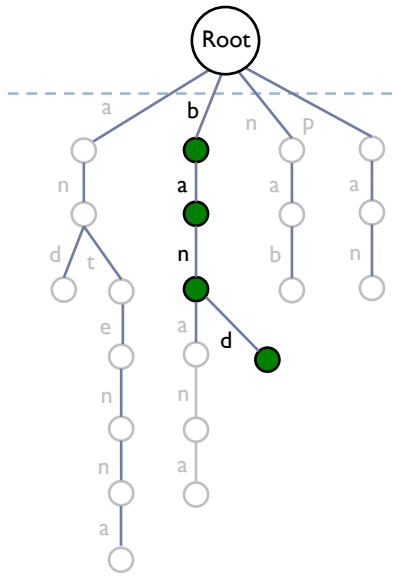
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



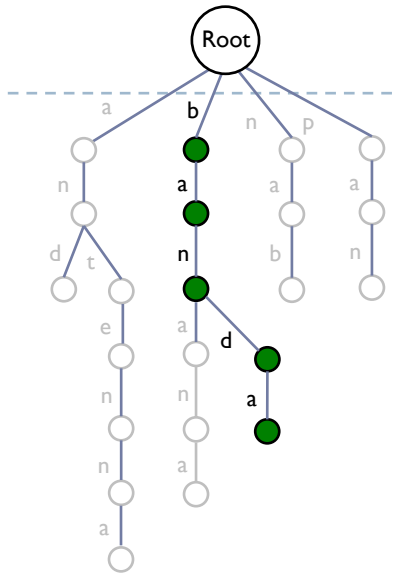
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



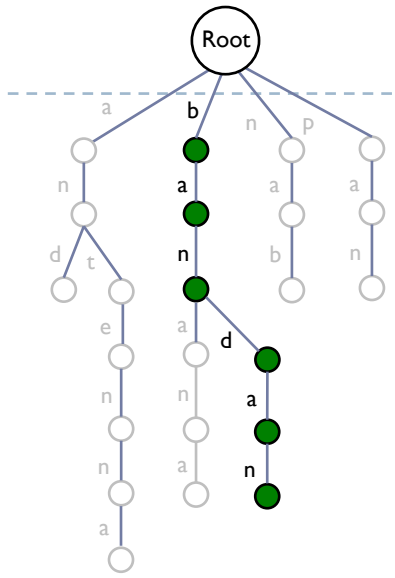
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



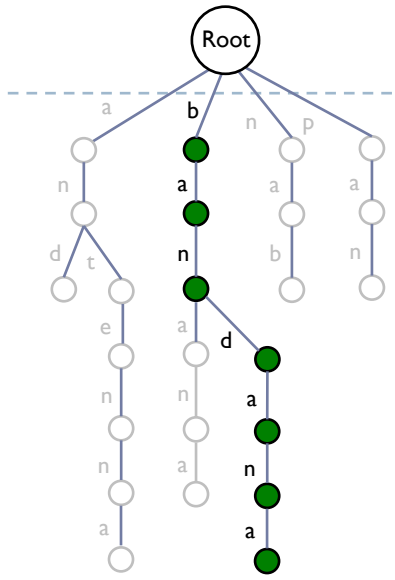
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



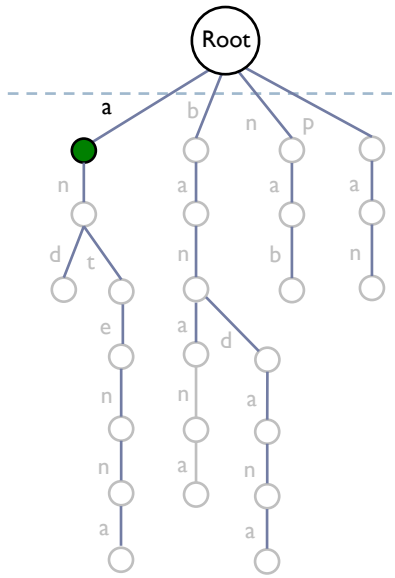
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



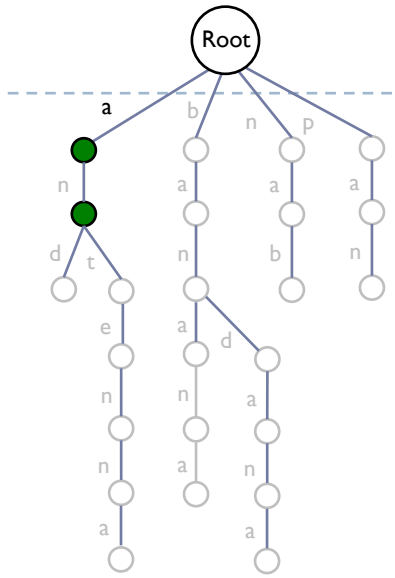
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



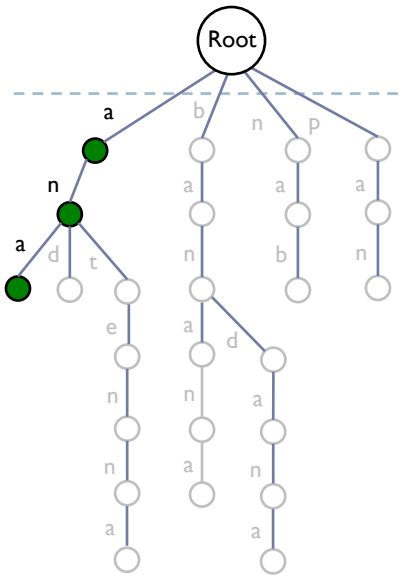
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



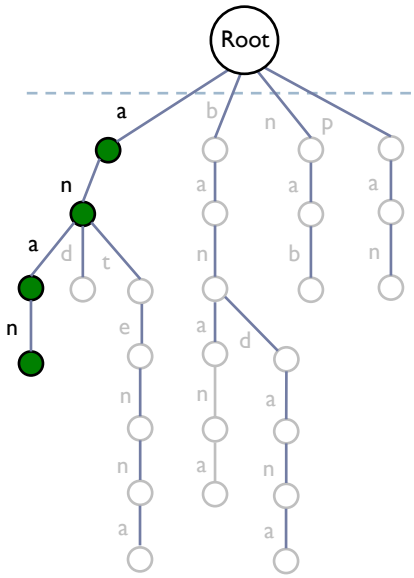
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



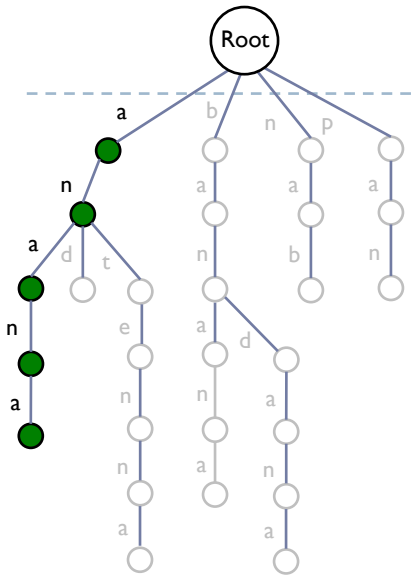
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



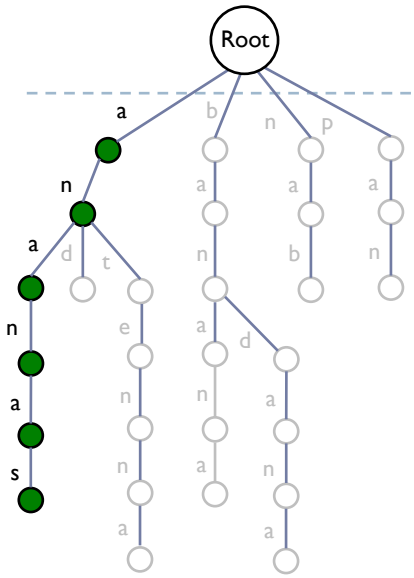
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



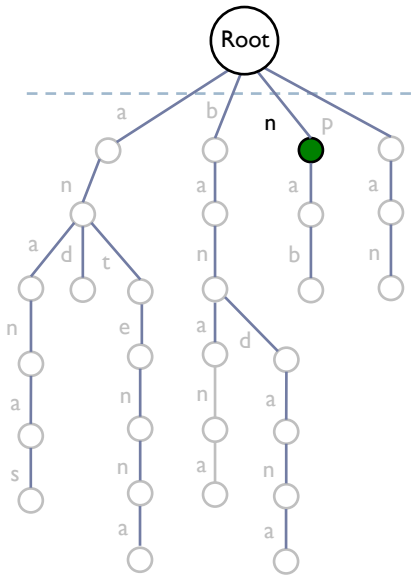
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



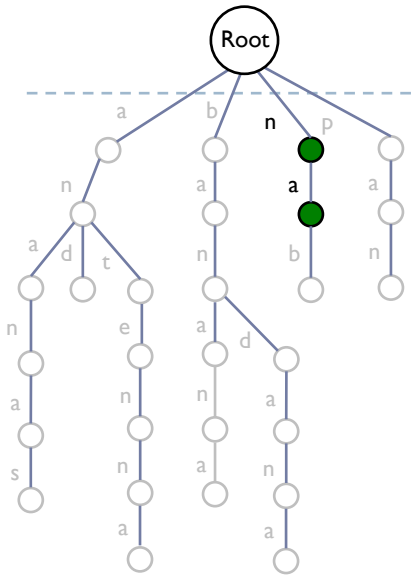
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



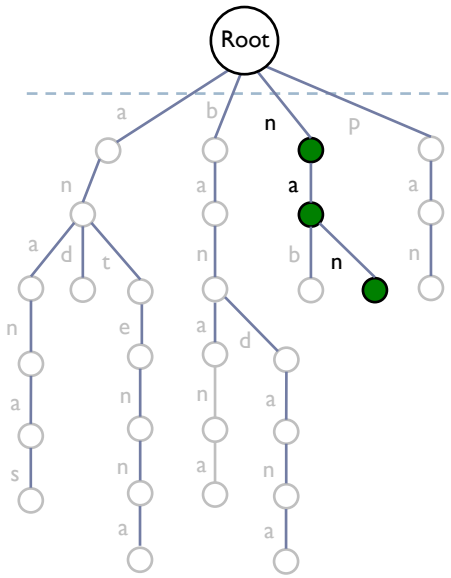
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



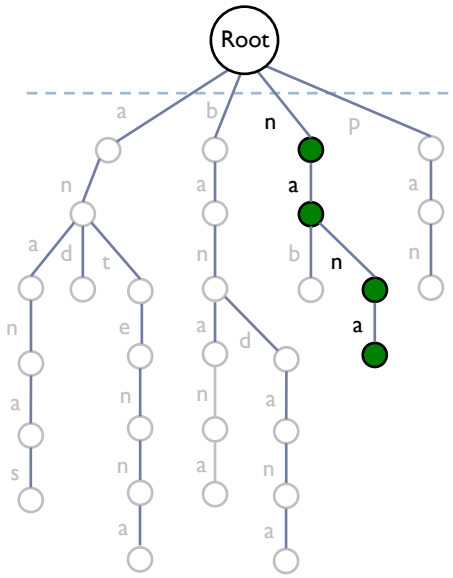
Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



Dictionary

banana
ban
pan
and
nab
antenna
bandana
ananas
nana



Dictionary

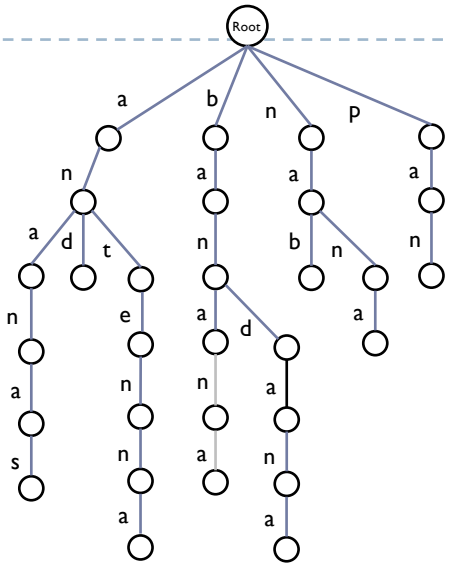
banana
ban
pan
and
nab
antenna
bandana
ananas
nana

Complexity of Trie Construction:

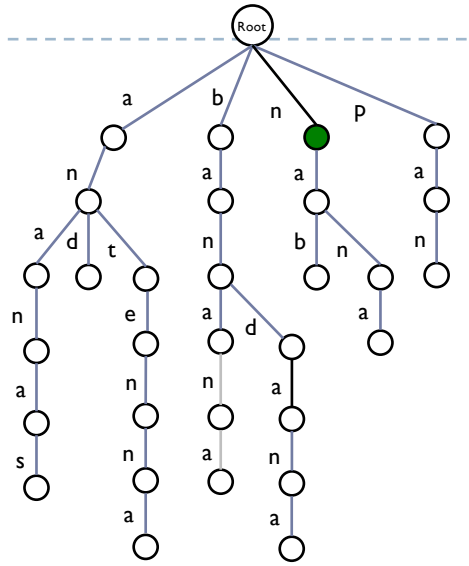
- O(|*Dictionary*|) or O(*m***N*)
  - *m* longest word in the Dictionary
  - *N* total number of words in the Dictionary

Searching in a Trie

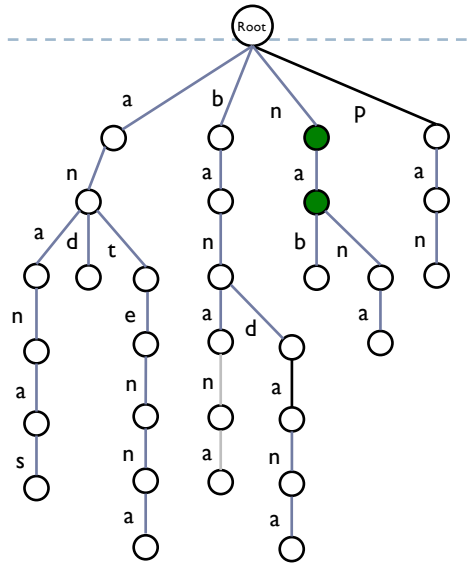
Search for: nar



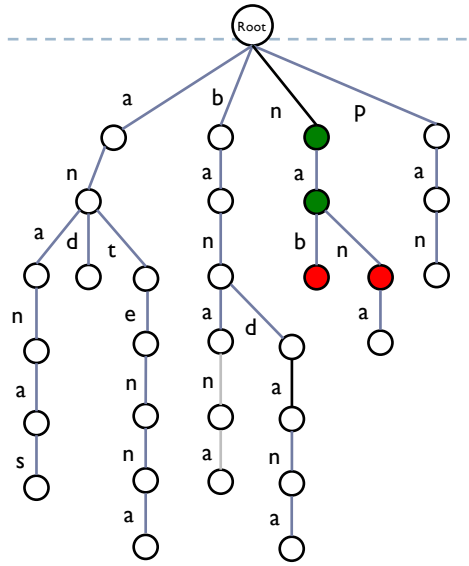
n a r


Root
a
b
n
P
n
a
a
a
a
d
t
n
b
n
n
n
e
a
d
a
a
n
n
a
s
n
a
n
a
a

n a r

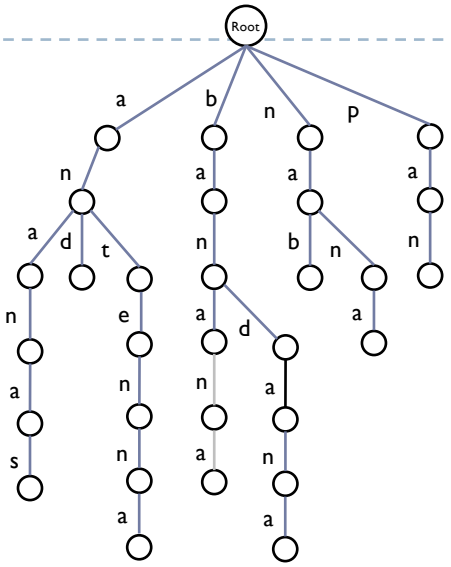

Root
a
b
n
P
n
a
a
a
a
d
t
n
b
n
n
n
e
a
d
a
a
n
n
a
s
n
a
n
a
a

n a r

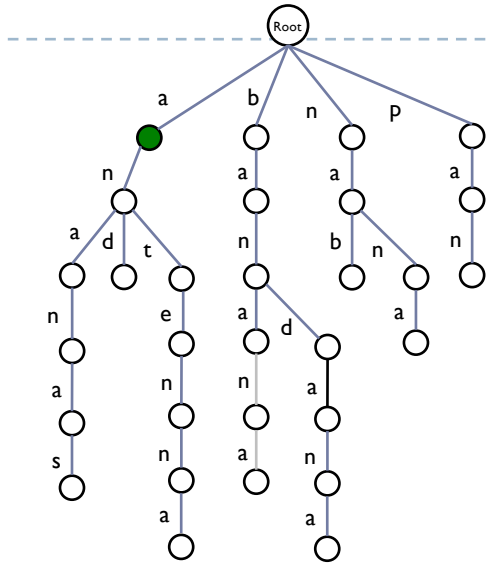


No match.

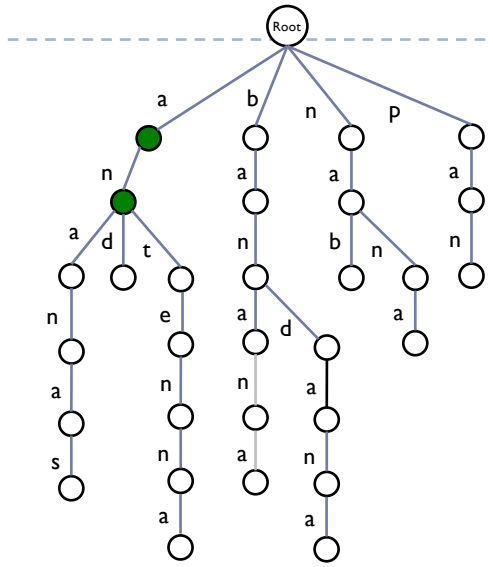
Search for: and



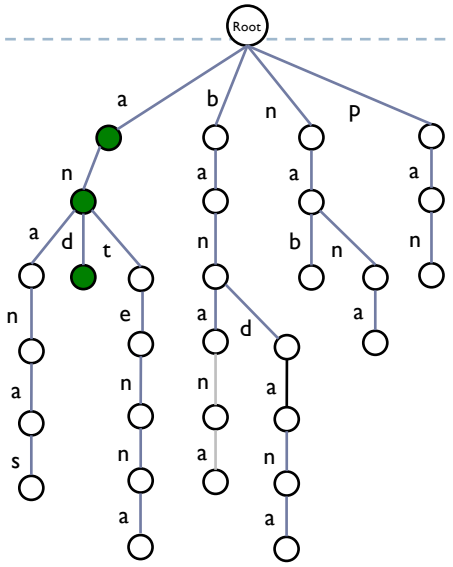
and


Root
a
b
n
p
n
a
a
a
a
d
t
n
b
n
n
n
e
a
d
a
a
n
n
a
s
n
a
n
a
a

and

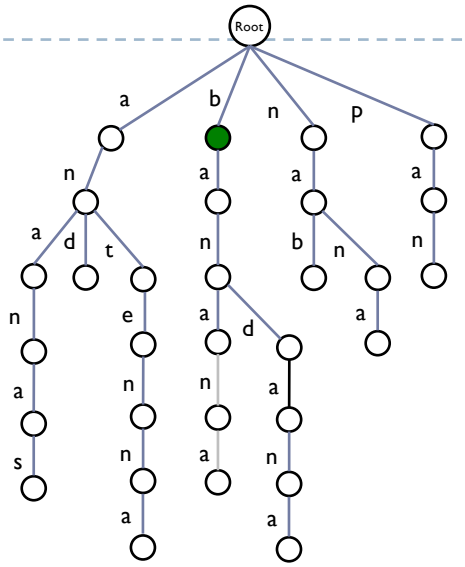

Root
a
b
n
p
n
a
a
a
a
d
t
n
b
n
n
n
e
a
d
a
a
n
n
a
s
n
a
n
a
a

and

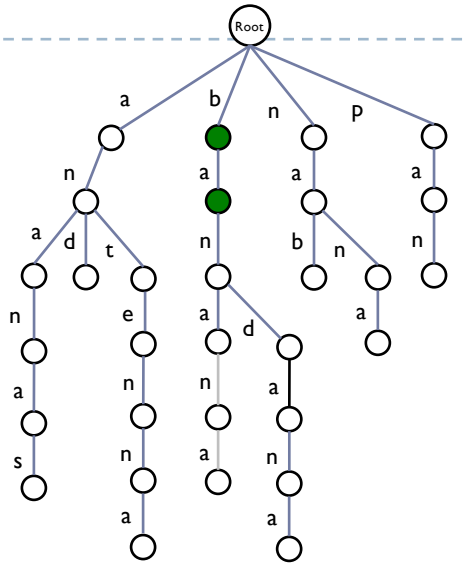


Match found, retrieve the postings for the keyword.

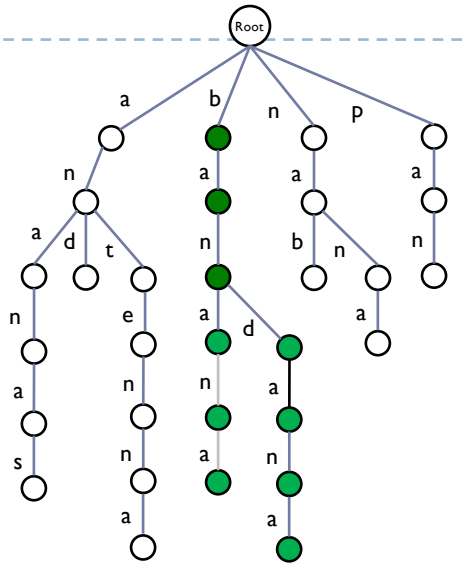
prefix search: ban*



ban*

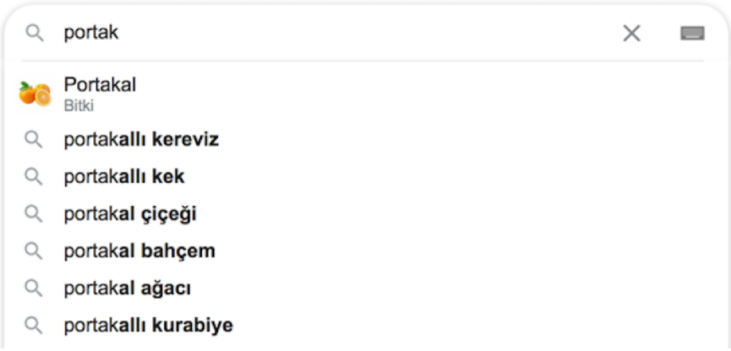


ban*



Matches:
ban
banana
bandana

Auto-Complete



Complexity of Trie Matching:

- Trie Construction: O(|*Dictionary*|) or O(*m***N*)
  - *m* longest word in the Dictionary
  - *N* total number of words in the Dictionary
- Search for keyword: O(*m*)
  - *m* longest word in the Dictionary

References

- *Introduction to Information Retrieval*, chapter 3
  - http://nlp.stanford.edu/IR-book/information-retrieval-book.html
- Some slides adapted from Dr. Christopher Manning and Dr. Pandu Nayak
- Trie slides adapted from the lecture notes of Bioinformatics Algorithms: An Active Learning Approach, Volume 2, Chapter 9.