Introduction to
Information Retrieval

CS276: Information Retrieval and Web Search

Pandu Nayak and Prabhakar Raghavan

Lecture 3: Dictionaries and tolerant retrieval
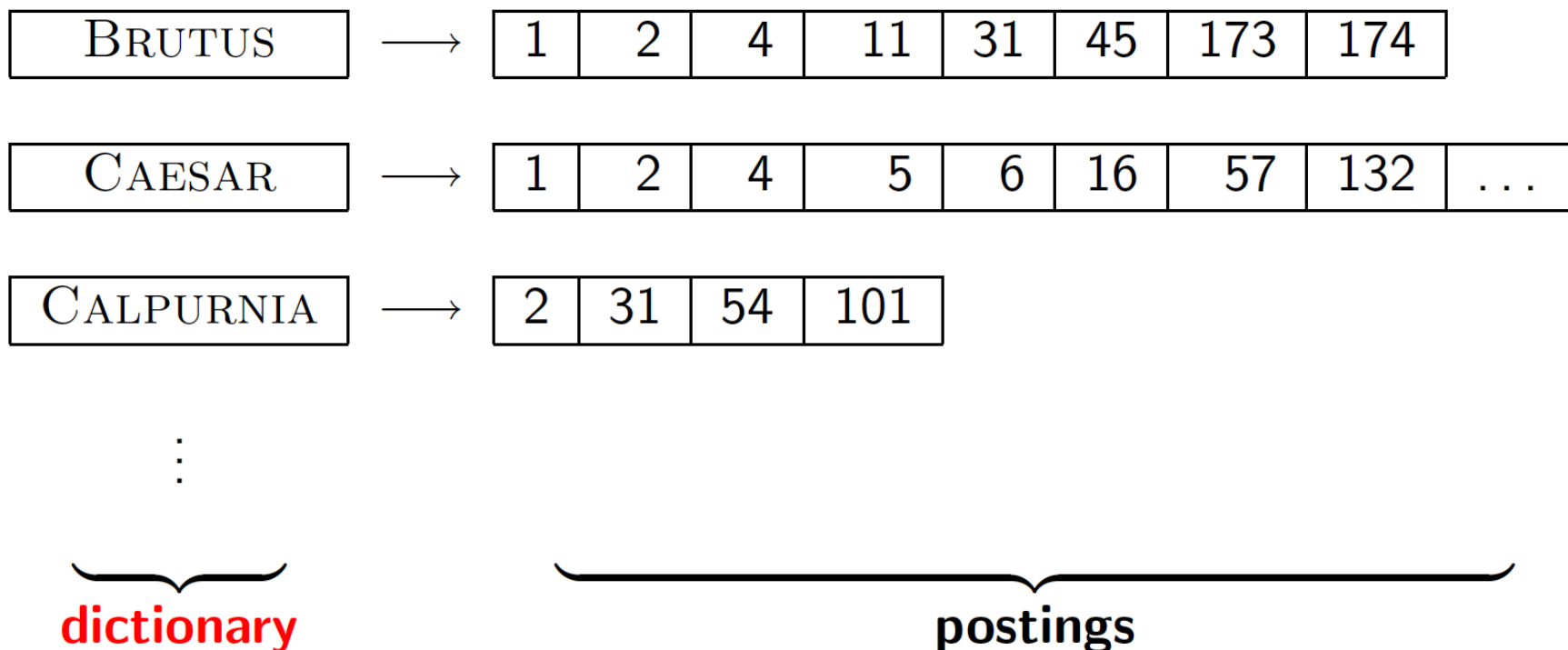
Recap of the previous lecture

- The type/token distinction
  - Terms are normalized types put in the dictionary
- Tokenization problems:
  - Hyphens, apostrophes, compounds, CJK
- Term equivalence classing:
  - Numbers, case folding, stemming, lemmatization
- Skip pointers
  - Encoding a tree-like structure in a postings list
- Biword indexes for phrases
- Positional indexes for phrases/proximity queries

This lecture

- Dictionary data structures

- "Tolerant" retrieval

    - Wild-card queries

    - Spelling correction

    - Soundex

Dictionary data structures for inverted indexes

- The dictionary data structure stores the term vocabulary, document frequency, pointers to each postings list ... in what data structure?

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

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?

Dictionary data structures

- Two main choices:
    - Hashtables
    - Trees
- Some IR systems use hashtables, some trees

Hashtables

- Each vocabulary term is hashed to an integer
    - (We assume you've seen hashtables before)
- Pros:
    - Lookup is faster than for a tree: O(1)
- Cons:
    - No easy way to find minor variants:
        - judgment/judgement
    - No prefix search   [tolerant retrieval]
    - If vocabulary keeps growing, need to occasionally do the expensive operation of rehashing *everything*

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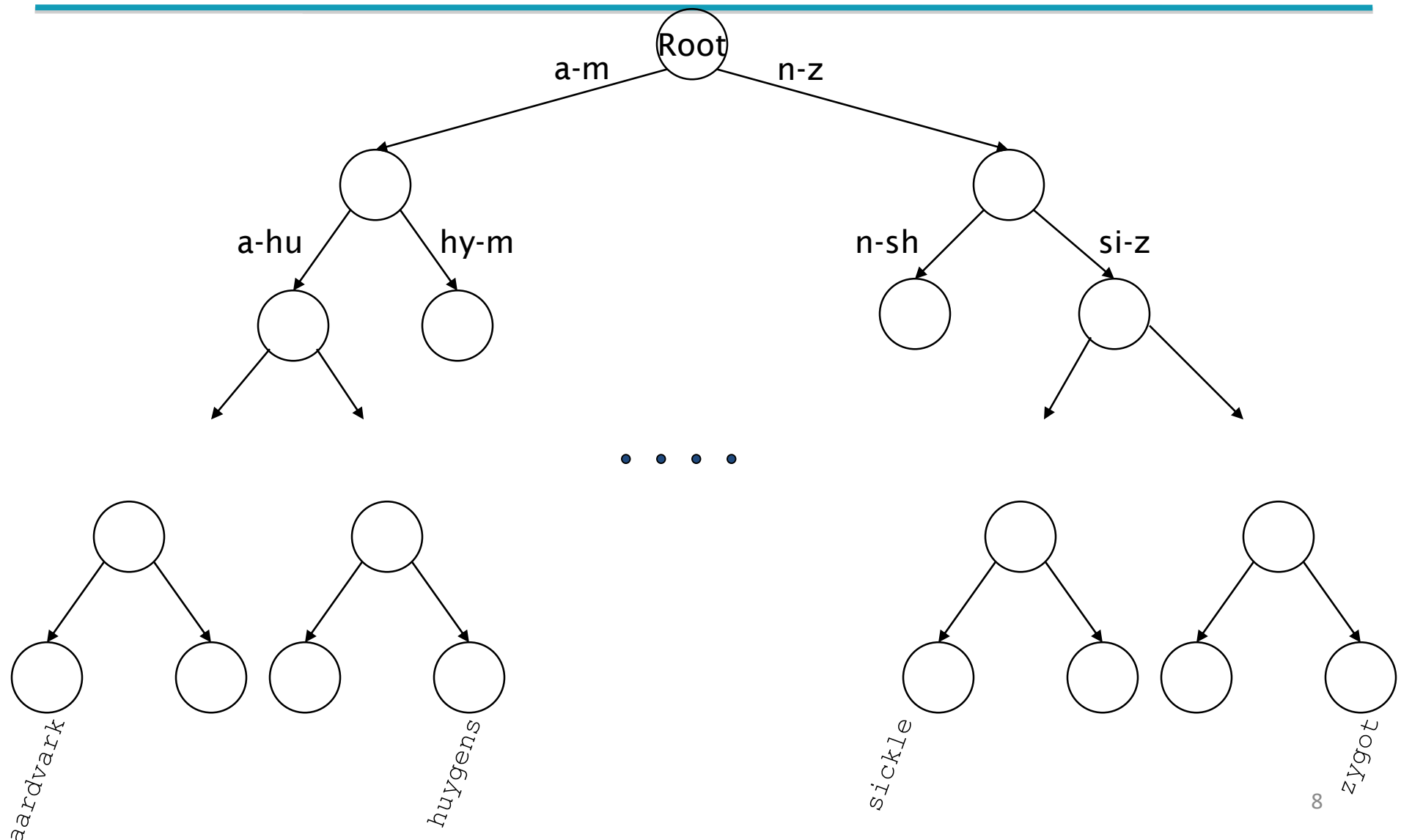
Tree: B-tree
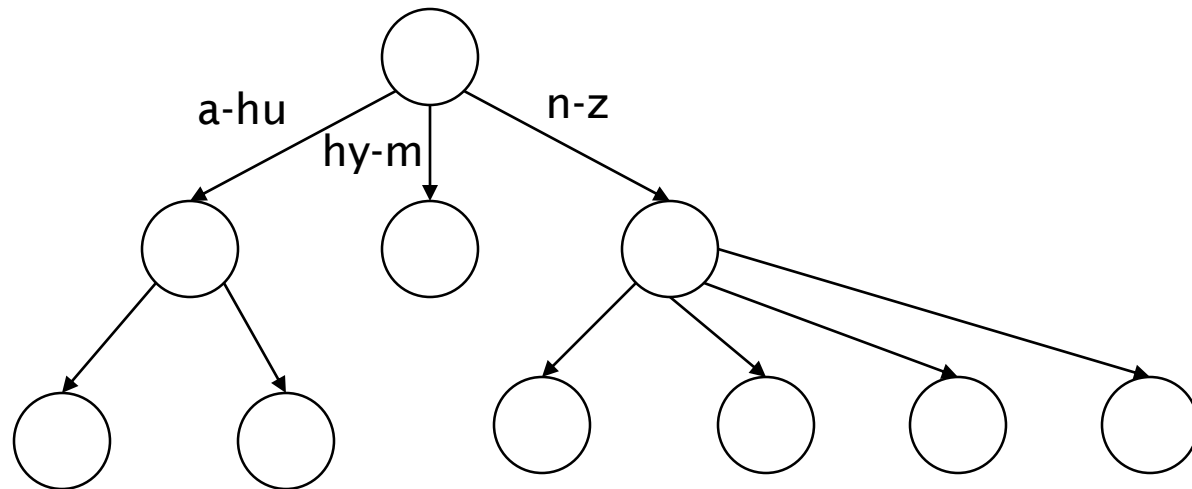
- Definition: Every internal nodel has a number of children in the interval [*a*,*b*] where *a, b* are appropriate natural numbers, e.g., [2,4].

Trees

- Simplest: binary tree

- More usual: B-trees

- Trees require a standard ordering of characters and hence strings … but we typically have one

- Pros:

  - Solves the prefix problem (terms starting with *hyp*)

- Cons:

  - Slower: O(log *M*) [and this requires *balanced* tree]

  - Rebalancing binary trees is expensive

    - But B-trees mitigate the rebalancing problem

WILD-CARD QUERIES

- (1) the user is uncertain of the spelling of a query term (Sydney vs. Sidney = S*dney)

- (2) the user is aware of multiple variants of spelling a term (e.g., color vs. colour);

- (3) the user is unsure whether the search engine performs stemming (e.g., judicial vs. judiciary, leading to the wildcard query judicia*)

- (4) foreign word or phrase (e.g., the query Universit* Stuttgart).

Wild-card queries: *

- ***mon*:** find all docs containing any word beginning with "mon".

- Easy with binary tree (or B-tree) lexicon: retrieve all words in range: ***mon ≤ w < moo***

- ***mon:*** find words ending in "mon": harder

  - Maintain an additional B-tree for terms *backwards.*

  Can retrieve all words in range: ***nom ≤ w < non.***

  lemon => monel

Exercise: from this, how can we enumerate all terms meeting the wild-card query ***pro*cent*** ?

Query processing

- At this point, we have an enumeration of all terms in the dictionary that match the wild-card query.

- We still have to look up the postings for each enumerated term.

- E.g., consider the query:

se*ate *AND* **fil*er**

This may result in the execution of many Boolean *AND* queries.

B-trees handle *'s at the end of a query term

- How can we handle *'s in the middle of query term?
  - ***co*tion***
- We could look up ***co**** AND ***tion*** in a B-tree and intersect the two term sets
  - Expensive
- The solution: transform wild-card queries so that the *'s occur at the end
- This gives rise to the **Permuterm** Index.

Permuterm index

- For term **hello**, index under:
  - **hello\$, ello\$h, llo\$he, lo\$hel, o\$hell, \$hello**

  where \$ is a special symbol.
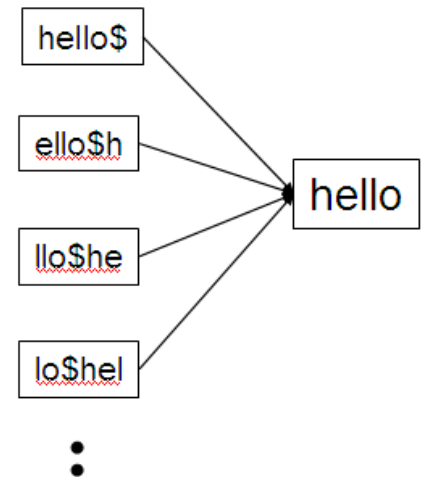
- m*n => n\$m* => man, moron, ….

- Queries:
  - **X**   lookup on **X\$**   **X***   lookup on   \$**X***
  - ***X**   lookup on **X\$***   ***X*** lookup on   **X***
  - **X*Y** lookup on **Y\$X***   **X*Y*Z**   ??? Exercise!

| hello\$ |
| ello\$h |
| llo\$he | → hello
| lo\$hel |

Query = **hel*o**
X=**hel**, Y=**o**
Lookup **o\$hel***

Permuterm query processing

- Rotate query wild-card to the right

- Now use B-tree lookup as before.

- *Permuterm problem: ≈ quadruples lexicon size*

  Empirical observation for English.

- fi*mo*er?

- er$fi*

- filter these out by exhaustive enumeration, checking each candidate to see if it contains mo

- fishmonger vs filibuster

Bigram (*k*-gram) indexes

- Enumerate all *k*-grams (sequence of *k* chars) occurring in any term

- castle => cas, ast , stl, tle are 3-grams

- $castle$ => $ca, cas, ast, stl, tle, le$.

- *e.g.,* from text "***April is the cruelest month***" we get the 2-grams (*bigrams*)

  a,ap,pr,ri,il,l,i,is,s,t,th,he,e,$c,cr,ru, ue,el,le,es,st,t$, m,mo,on,nt,h

  - $ is a special word boundary symbol

- Maintain a <u>second</u> inverted index <u>*from bigrams to dictionary terms*</u> that match each bigram.

18

Bigram index example

- The *k*-gram index finds *terms* based on a query consisting of *k*-grams (here *k*=2).
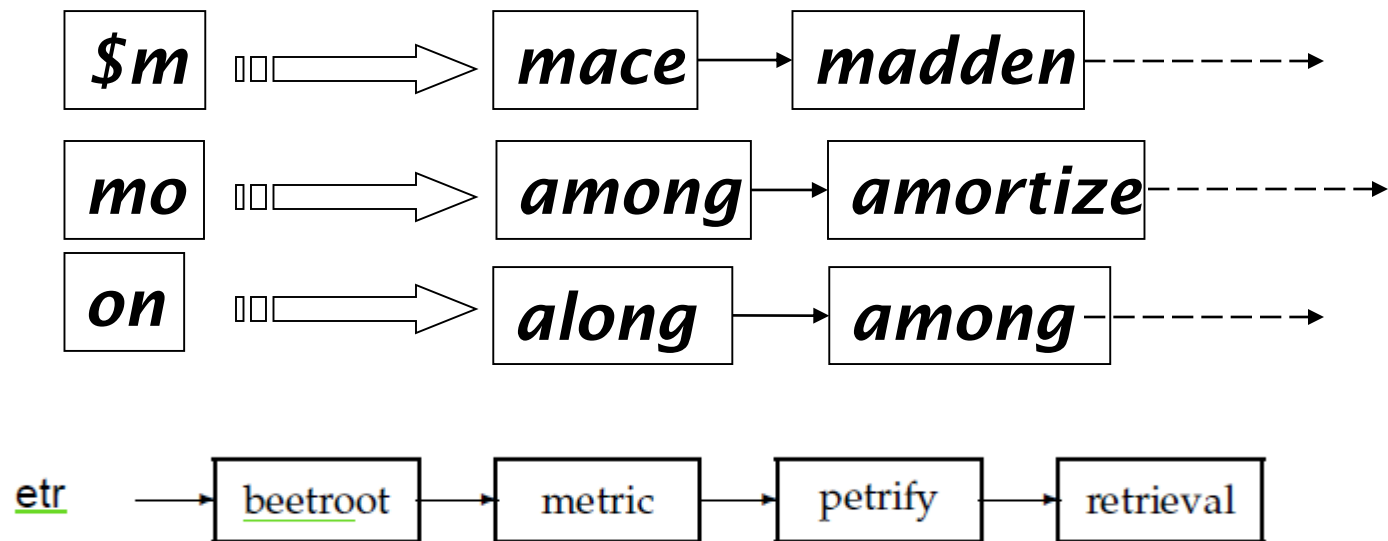
re*ve => $re AND ve$ => relive, remove and retrieve

Processing wild-cards

- Query *mon** can now be run as
  - *$m AND mo AND on*
- Gets terms that match AND version of our wildcard query.
- But we'd enumerate *moon*.
- Must post-filter these terms against query.
- red* => $re AND red => retired, ….
- Surviving enumerated terms are then looked up in the term-document inverted index.
- Fast, space efficient (compared to permuterm).

Processing wild-card queries

- As before, we must execute a Boolean query for each enumerated, filtered term.

- Wild-cards can result in expensive query execution (very large disjunctions…)
  - pyth* AND prog*

- If you encourage "laziness" people will respond!

[ ]	Search

Type your search terms, use '*' if you need to.
E.g., Alex* will match Alexander.

- Which web search engines allow wildcard queries? 21

SPELLING CORRECTION

Document correction

- Especially needed for OCR'ed documents
  - Correction algorithms are tuned for this: rn/m
  - Can use domain-specific knowledge
    - E.g., OCR can confuse O and D more often than it would confuse O and I (adjacent on the QWERTY keyboard, so more likely interchanged in typing).
- But also: web pages and even printed material have typos
- Goal: the dictionary contains fewer misspellings
- But often we don't change the documents and instead fix the query-document mapping

Query mis-spellings

- Our principal focus here
  - E.g., the query ***Alanis Morisett***
- We can either
  - Retrieve documents indexed by the correct spelling, OR
  - Return several suggested alternative queries with the correct spelling
    - *Did you mean … ?*

- Two principal uses
  - Correcting document(s) being indexed
  - Correcting user queries to retrieve "right" answers
- britney spears => britian spears, britney's spears, brandy spears and prittany spears.

- for a mis-spelled query, choose the "nearest" one.
- Select the one that is more common. (grunt and grant for grnt)

Spell correction

- Two main flavors:
  - Isolated word
    - Check each word on its own for misspelling
    - carot => carrot or tarot
    - Will not catch typos resulting in correctly spelled words
    - e.g., *from → form*
  - Context-sensitive
    - Look at surrounding words,
    - e.g., *I flew <u>form</u> Heathrow to Narita.*

Isolated word correction

- Fundamental premise – there is a lexicon from which the correct spellings come
- Two basic choices for this
  - A standard lexicon such as
    - Webster's English Dictionary
    - An "industry-specific" lexicon – hand-maintained
  - The lexicon of the indexed corpus
    - E.g., all words on the web
    - All names, acronyms etc.
    - (Including the mis-spellings)

Isolated word correction

- Given a lexicon and a character sequence Q, return the words in the lexicon closest to Q

- What's "closest"?

- We'll study several alternatives
  - Edit distance (Levenshtein distance)
  - Weighted edit distance
  - *n*-gram overlap

Edit distance

- Given two strings S_1 and S_2, the minimum number of operations to convert one to the other

- Operations are typically character-level

  - Insert, Delete, Replace, (Transposition)

- E.g., the edit distance from ***dof*** to ***dog*** is 1

  - From ***cat*** to ***act*** is 2   (Just 1 with transpose.)

  - from ***cat*** to ***dog*** is 3.

- Generally found by dynamic programming.

- See http://www.merriampark.com/ld.htm for a nice example plus an applet.

Weighted edit distance

- As above, but the weight of an operation depends on the character(s) involved
    - Meant to capture OCR or keyboard errors
      Example: *m* more likely to be mis-typed as *n* than as *q*
    - Therefore, replacing *m* by *n* is a smaller edit distance than by *q*
    - This may be formulated as a probability model
- Requires weight matrix as input
- Modify dynamic programming to handle weights

Using edit distances

- Given query, first enumerate all character sequences within a preset (weighted) edit distance (e.g., 2)

- Intersect this set with list of "correct" words

- Show terms you found to user as suggestions

- Alternatively,

    - We can look up all possible corrections in our inverted index and return all docs … slow

    - We can run with a single most likely correction

- The alternatives disempower the user, but save a round of interaction with the user

Edit distance to all dictionary terms?

- Given a (mis-spelled) query – do we compute its edit distance to every dictionary term?
  - Expensive and slow
  - Alternative?
- How do we cut the set of candidate dictionary terms?
- restrict to terms beginning with the same letter
  - Permuterm index : mase => sema
  - semaphore or semantic but not mare or mane
- One possibility is to use *n*-gram overlap for this
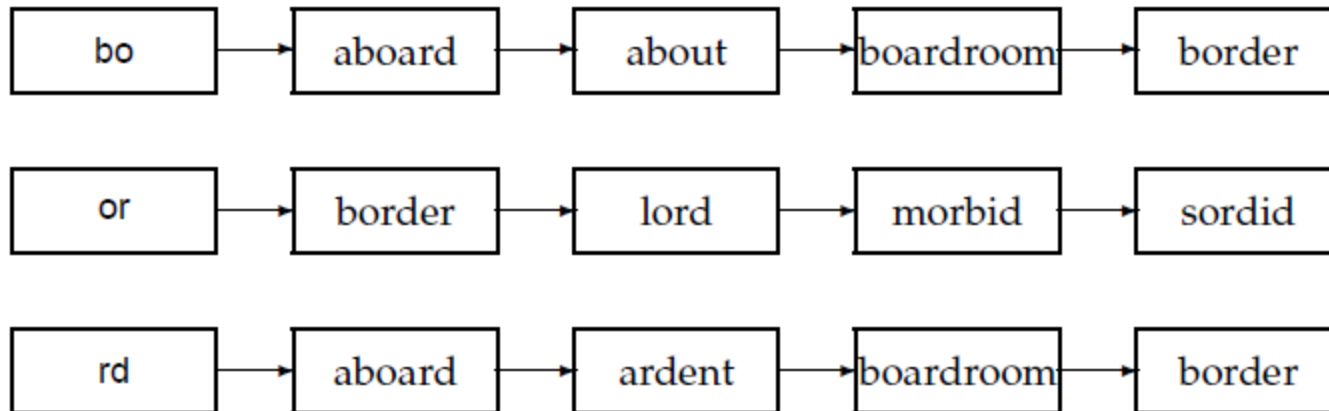- This can also be used by itself for spelling correction.

n-gram overlap

- Enumerate all the *n*-grams in the query string as well as in the lexicon

- Use the *n*-gram index (recall wild-card search) to retrieve all lexicon terms matching any of the query *n*-grams

- Threshold by number of matching *n*-grams
  - Variants – weight by keyboard layout, etc.

Example with trigrams

- Suppose the text is **november**

  - Trigrams are *nov, ove, vem, emb, mbe, ber*.

- The query is **december**

  - Trigrams are *dec, ece, cem, emb, mbe, ber*.

- So 3 trigrams overlap (of 6 in each term)

- How can we turn this into a normalized measure of overlap?

query bord

bo	→	aboard	→	about	→	boardroom	→	border

or	→	border	→	lord	→	morbid	→	sordid

rd	→	aboard	→	ardent	→	boardroom	→	border

boardroom

One option – Jaccard coefficient

- A commonly-used measure of overlap

- Let *X* and *Y* be two sets; then the J.C. is

$$|X \cap Y| / |X \cup Y|$$

- Equals 1 when *X* and *Y* have the same elements and zero when they are disjoint

- *X* and *Y* don't have to be of the same size

- Always assigns a number between 0 and 1
  - Now threshold to decide if you have a match
  - E.g., if J.C. > 0.8, declare a match

- *q* = bord

- *t* = boardroom.

- Matching bigrams = 2 (from bo and rd)

- 2/(8+3−2)

Matching trigrams

- Consider the query ***lord*** – we wish to identify words matching 2 of its 3 bigrams (***lo, or, rd***)

lo	⇒	*alone* → *lore* → *sloth*
or	⇒	*border* → *lore* → *morbid*
rd	⇒	*ardent* → *border* → *card*
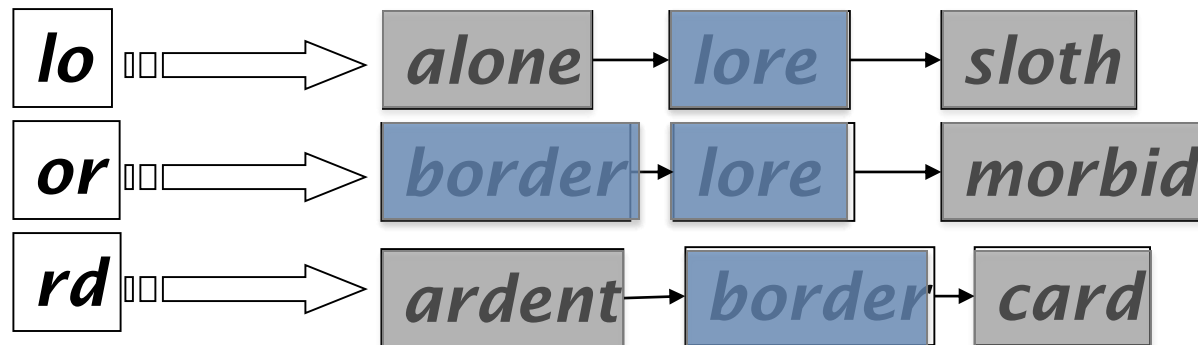
Standard postings "merge" will enumerate …

Adapt this to using Jaccard (or another) measure.

Context-sensitive spell correction

- Text: ***I flew <u>from</u> Heathrow to Narita.***

- Consider the phrase query ***"flew <u>form</u> Heathrow"***

- We'd like to respond

    Did you mean "***flew from Heathrow***"?

because no docs matched the query phrase.

Context-sensitive correction

- Need surrounding context to catch this.
- First idea: retrieve dictionary terms close (in weighted edit distance) to each query term
- Now try all possible resulting phrases with one word "fixed" at a time
    - *flew from heathrow*
    - *fled form heathrow*
    - *flea form heathrow*
- **Hit-based spelling correction:** Suggest the alternative that has lots of hits.

Exercise

- Suppose that for ***"flew <u>form</u> Heathrow"*** we have 7 alternatives for flew, 19 for form and 3 for heathrow.

How many "corrected" phrases will we enumerate in this scheme?

Another approach

- Break phrase query into a conjunction of biwords (Lecture 2).

- Look for biwords that need only one term corrected.

- Enumerate only phrases containing "common" biwords.

General issues in spell correction

- We enumerate multiple alternatives for "Did you mean?"

- Need to figure out which to present to the user

  - The alternative hitting most docs

  - Query log analysis

- More generally, rank alternatives probabilistically

$$\text{argmax}_{corr}\ P(corr \mid query)$$

  - From Bayes rule, this is equivalent to

$$\text{argmax}_{corr}\ P(query \mid corr) * P(corr)$$

Noisy channel   Language model

SOUNDEX

Soundex

- Class of heuristics to expand a query into phonetic equivalents

  - Language specific – mainly for names

  - E.g., ***chebyshev*** → ***tchebycheff***

- Invented for the U.S. census … in 1918

Soundex – typical algorithm

- Turn every token to be indexed into a 4-character reduced form

- Do the same with query terms

- Build and search an index on the reduced forms
  - (when the query calls for a soundex match)

- http://www.creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm#Top

Soundex – typical algorithm

1. Retain the first letter of the word.

2. Change all occurrences of the following letters to '0' (zero):
    'A', E', 'I', 'O', 'U', 'H', 'W', 'Y'.

3. Change letters to digits as follows:

-   B, F, P, V $\rightarrow$ 1

-   C, G, J, K, Q, S, X, Z $\rightarrow$ 2

-   D,T $\rightarrow$ 3

-   L $\rightarrow$ 4

-   M, N $\rightarrow$ 5

-   R $\rightarrow$ 6

Soundex continued

4. Remove all pairs of consecutive digits.

5. Remove all zeros from the resulting string.

6. Pad the resulting string with trailing zeros and return the first four positions, which will be of the form <uppercase letter> <digit> <digit> <digit>.

E.g., **Herman** becomes H655.

Will **hermann** generate the same code?

Soundex

- Soundex is the classic algorithm, provided by most databases (Oracle, Microsoft, …)

- How useful is soundex?

- Not very – for information retrieval

- Okay for "high recall" tasks (e.g., Interpol), though biased to names of certain nationalities

- Zobel and Dart (1996) show that other algorithms for phonetic matching perform much better in the context of IR

What queries can we process?

- We have
  - Positional inverted index with skip pointers
  - Wild-card index
  - Spell-correction
  - Soundex
- Queries such as

(SPELL(moriset) /3 toron*to) OR *SOUNDEX***(chaikofski)**

Exercise

- Draw yourself a diagram showing the various indexes in a search engine incorporating all the functionality we have talked about

- Identify some of the key design choices in the index pipeline:

  - Does stemming happen before the Soundex index?

  - What about *n*-grams?

- Given a query, how would you parse and dispatch sub-queries to the various indexes?

Resources

- IIR 3, MG 4.2

- Efficient spell retrieval:

  - K. Kukich. Techniques for automatically correcting words in text. ACM Computing Surveys 24(4), Dec 1992.

  - J. Zobel and P. Dart. Finding approximate matches in large lexicons. Software - practice and experience 25(3), March 1995. http://citeseer.ist.psu.edu/zobel95finding.html

  - Mikael Tillenius: Efficient Generation and Ranking of Spelling Error Corrections. Master's thesis at Sweden's Royal Institute of Technology. http://citeseer.ist.psu.edu/179155.html

- **Nice, easy reading on spell correction:**

  - Peter Norvig: How to write a spelling corrector

  http://norvig.com/spell-correct.html