

CS317

Information Retrieval

Week 03

Muhammad Rafi

February 10, 2020

Spelling & Phonetic
Corrections

Spelling Corrections

- Two principal uses
 - Correcting document(s) being indexed
 - Correcting user queries to retrieve “right” answers
- Two main flavors:
 - Isolated word
 - Check each word on its own for misspelling
 - Will not catch typos resulting in correctly spelled words
 - e.g., **from** → **form**
 - Context-sensitive
 - Look at surrounding words,
 - e.g., **I flew form Heathrow to Narita.**

Spelling Corrections

[Return to Google's jobs pages](#)

488941 britney spears	29 britent spears	9 brinttany spears	5 brney spears	3 britiy spears	2 britreny spears
40134 brittany spears	29 brittany spears	9 britany spears	5 broitney spears	3 britmeny spears	2 brittany spears
36315 britney spears	29 brittany spears	9 britany spears	5 broitney spears	3 britneey spears	2 brittany spears
24342 britany spears	29 britney spears	9 britn spears	5 brutney spears	3 britnehy spears	2 brittney spears
7331 britny spears	26 birttney spears	9 britnew spears	5 btiyney spears	3 britnely spears	2 britain spears
6633 briteny spears	26 breitney spears	9 britney spears	5 btrictney spears	3 britnesy spears	2 britane spears
2656 brittney spears	26 brinity spears	9 britney spears	5 gritney spears	3 britnetty spears	2 britanhy spears
1807 britney spears	26 britvay spears	9 britny spears	5 spritney spears	3 britnex spears	2 britania spears
1635 brittney spears	26 britneyt spears	9 brittney spears	4 bittny spears	3 britneykx spears	2 britann spears
1479 britney spears	26 brittan spears	9 brtny spears	4 hritney spears	3 britnity spears	2 britanna spears
1479 britany spears	26 brittne spears	9 brytny spears	4 brandy spears	3 britntey spears	2 britannie spears
1338 britiny spears	26 brittany spears	9 britney spears	4 brritney spears	3 britnyey spears	2 britannt spears
1211 britnet spears	24 beltney spears	8 britiny spears	4 breaciny spears	3 brittorny spears	2 britannu spears
1096 britney spears	24 birtney spears	8 bithney spears	4 breecney spears	3 britttney spears	2 britanyl spears
991 britaney spears	24 brightney spears	8 brattany spears	4 bretiney spears	3 britttney spears	2 britanyt spears
991 britany spears	24 brintiny spears	8 breitny spears	4 brifitney spears	3 britttney spears	2 brittney spears
811 britney spears	24 britvany spears	8 brettany spears	4 brittany spears	3 brittjen spears	2 britteny spears
811 britney spears	24 britenny spears	8 brightny spears	4 brieteny spears	3 briytney spears	2 britnet spears
664 britney spears	24 britini spears	8 brintay spears	4 briety spears	3 britney spears	2 briteniy spears
664 britney spears	24 brittney spears	8 brinttney spears	4 briltny spears	3 broteny spears	2 britenya spears
664 britney spears	24 britini spears	8 briottney spears	4 brittany spears	3 bruaney spears	2 britianey spears
601 britney spears	24 brittne spears	8 britanya spears	4 brinie spears	3 britsiany spears	2 britin spears
601 britny spears	21 brittney spears	8 britley spears	4 brintcney spears	3 britiny spears	2 britinary spears
544 brittany spears	21 britany spears	8 britneyb spears	4 brintne spears	3 britinney spears	2 britiny spears
544 brittany spears	21 biteny spears	8 britney spears	4 britaby spears	3 brittany spears	2 britanay spears
364 britey spears	21 breatney spears	8 brittany spears	4 britney spears	3 brittney spears	2 britnat spears
364 brittany spears	21 britani spears	8 brittner spears	4 britainey spears	3 britnet spears	2 britnhey spears
329 britney spears	21 britanie spears	8 brottany spears	4 britinie spears	3 brytiny spears	2 britndy spears
269 britney spears	21 britcney spears	7 baritney spears	4 britinney spears	3 btney spears	2 britneh spears
269 britney spears	21 brittany spears	7 britney spears	4 brittney spears	3 brittney spears	2 brittney spears
244 britne spears	21 brittany spears	7 bitcney spears	4 britnear spears	3 pretney spears	2 brittney spears
244 britney spears	21 brtany spears	7 britny spears	4 britnel spears	3 zbritney spears	2 brittney spears
220 breatney spears	21 britany spears	7 breateny spears	4 britney spears	2 barittany spears	2 brittney spears
220 brittany spears	19 birney spears	7 brianty spears	4 britney spears	2 bbbritney spears	2 brittney spears
159 brittney spears	19 brittney spears	7 brinty spears	4 britney spears	2 bbitney spears	2 brittney spears
163 britny spears	19 brittany spears	7 brittany spears	4 brittaby spears	2 bbbritny spears	2 brittney spears
147 britney spears	19 britnee spears	7 britny spears	4 brittney spears	2 bbbritny spears	2 brittney spears
147 brittney spears	19 britney spears	7 britney spears	4 brittney spears	2 brittany spears	2 brittney spears
147 brittney spears	19 brittany spears	7 brittney spears	4 brittney spears	2 brittany spears	2 brittney spears

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 has typos
- Goal: the dictionary contains fewer misspellings

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.

Edit Distance – Levenshtein

```

EDITDISTANCE( $s_1, s_2$ )
1   $\text{int } m[i, j] = 0$ 
2  for  $i \leftarrow 1$  to  $|s_1|$ 
3  do  $m[i, 0] = i$ 
4  for  $j \leftarrow 1$  to  $|s_2|$ 
5  do  $m[0, j] = j$ 
6  for  $i \leftarrow 1$  to  $|s_1|$ 
7  do for  $j \leftarrow 1$  to  $|s_2|$ 
8       $\text{do } m[i, j] = \min\{m[i-1, j-1] + \text{if } (s_1[i] = s_2[j]) \text{ then } 0 \text{ else } 1, \text{ if}$ 
9           $m[i-1, j] + 1,$ 
10          $m[i, j-1] + 1\}$ 
11 return  $m[|s_1|, |s_2|]$ 

```

► **Figure 3.5** Dynamic programming algorithm for computing the edit distance between strings s_1 and s_2 .

Edit Distance – Levenshtein

		f	a	s	t
	0	1 1	2 2	3 3	4 4
c	1 1	1 2	2 3	3 4	4 5
	1	2 1	2 2	3 3	4 4
a	2 2	2 2	1 3	3 4	4 5
	2	3 2	3 1	2 2	3 3
t	3 3	3 3	3 2	2 3	2 4
	3	4 3	4 2	3 2	3 2
s	4 4	4 4	4 3	2 3	3 3
	4	5 4	5 3	4 2	3 3

► **Figure 3.6** Example Levenshtein distance computation. The 2×2 cell in the $[i, j]$ entry of the table shows the three numbers whose minimum yields the fourth. The cells in *italics* determine the edit distance in this example.

Using Edit Distance

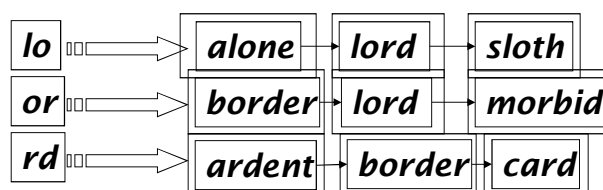
- Given query, first enumerate all character sequences within a preset (weighted) edit distance (e.g., 2)
- Intersect this set with list of “correct” words
- Show terms you found to user as suggestions
- Alternatively,
 - We can look up all possible corrections in our inverted index and return all docs ... slow
 - We can run with a single most likely correction

n-gram Overlaps

- 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.

2-grams for match

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



Standard postings “merge” will enumerate ...

Context-Sensitive Spelling Corrections

- Text: *I flew from Heathrow to Narita.*
- Consider the phrase query “**flew form Heathrow**”
- We’d like to respond
 Did you mean “**flew from Heathrow**”?
 because no docs matched the query phrase.

Context-Sensitive Spelling Corrections

- 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.

Issues in Spelling Corrections

- We enumerate multiple alternatives for “Did you mean?”
- Need to figure out which to present to the user
- Use heuristics
 - The alternative hitting most docs
 - Query log analysis + tweaking
 - For especially popular, topical queries
- Spell-correction is computationally expensive
 - Avoid running routinely on every query?
 - Run only on queries that matched few docs

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 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 → 1
 - C, G, J, K, Q, S, X, Z → 2
 - D, T → 3
 - L → 4
 - M, N → 5
 - R → 6

Soundex Algorithm

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.

Soundex

- Soundex is the classic algorithm, provided by most databases (Oracle, Microsoft, ...)
- How useful is soundex?
 - Not very – for information retrieval
- Zobel and Dart (1996) show that other algorithms for phonetic matching perform much better in the context of IR

Soundex Exercise

- Find two differently spelled proper nouns (different to the course example) whose soundex codes are the same and give their soundex code.
 - Mary, Nira (Soundex code = 5600).
- Find two phonetically similar proper nouns whose soundex codes are different.
 - Chebyshev, Tchebycheff
 - Rafi, Rafee