Workshop 2
(part 3)

Preparing to Index: Spelling correction

Goran Nenadic

with examples from the IIR book
(and Christopher Manning and Pandu Nayak)

MANCHESTER 1824
The University of Manchester

Spelling errors

(new) words or misspellings?

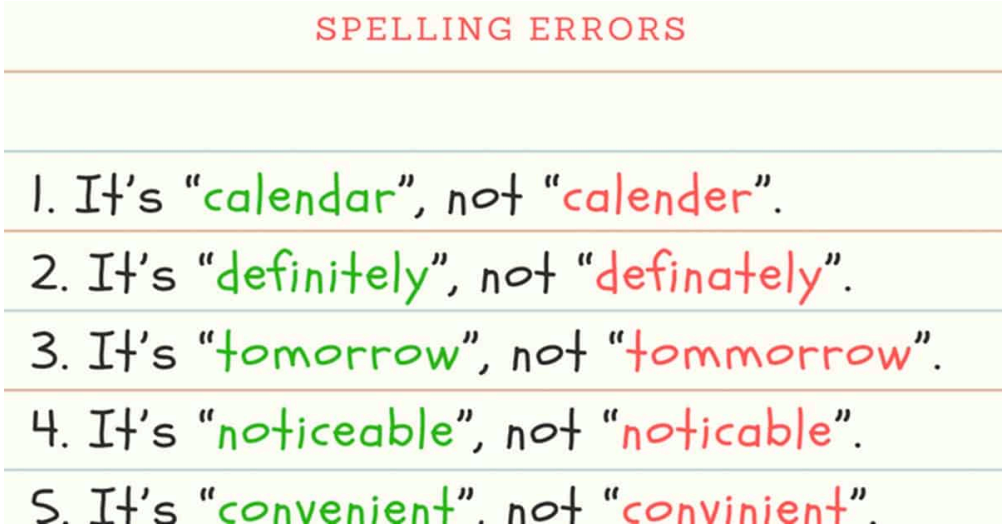
Spelling errors

- Frequency of spelling errors in human typed text varies from
  - 0.05% of the words in carefully edited newswire
  - 26% in Web queries
  - On average, 1 word in every tweet is misspelled
- Two types of spelling errors
  - Non-word errors
    - *graffe* → *giraffe*
  - Real-word errors
    - *piece* → *peace*

← Needs context

Spelling errors

- 80% of all misspelled words are caused by single-error misspellings (edit distance 1)
  - **Insertion** (ther)
  - **Deletion** (th)
  - **Substitution** (thw)
  - **Transposition** (teh)



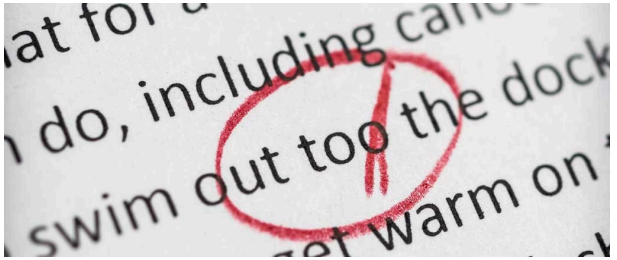
- Almost all errors within edit distance 2

Causes of Spelling Errors

- **Typographical** (keyboard-based)
  - 83% novice and 51% overall were keyboard related errors
  - Immediately adjacent keys in the same row of the keyboard (50% of the novice substitutions, 31% of all substitutions)

- **Cognitive**
  - Phonetic: seperate – separate
  - Homophones: there – their, piece – peace

- Optical character recognition
  - Make different kinds of errors
    "D"->"O"
    "ri"-> "n"



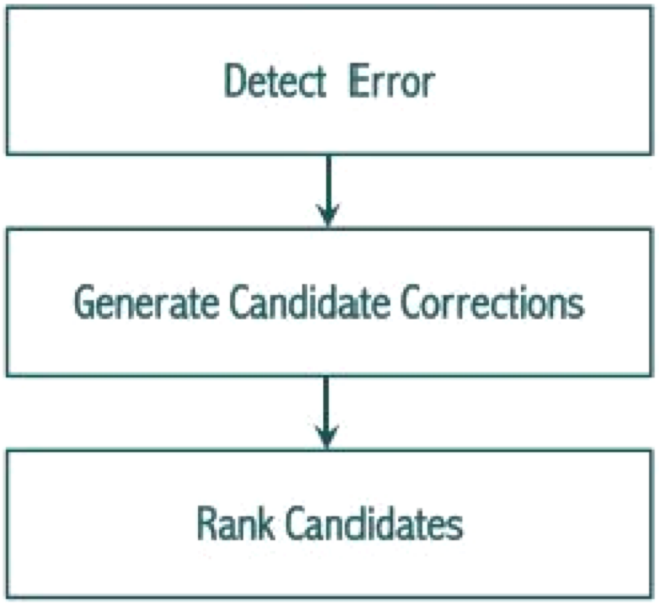
Spelling Errors in Queries

- Note: spelling correction is useful (i.e. needed) for querying too (not only in indexing)

488941 britney spears	… (many, many more)
40134 brittany spears	2 brynty spears
36315 brittney spears	2 brythey spears
24342 britany spears	2 bryttney spears
7331 britny spears	2 btiany spears
6633 briteny spears	2 btirtney spears
2696 britteny spears	2 btitiney spears
1807 briney spears	2 btittny spears
1635 brittny spears	2 btritany spears
1479 brintey spears	2 buttney spears
1479 britanny spears	2 grittney spears
1338 britiny spears	2 prietny spears
1211 britnet spears	2 pritany spears
… (many, many more)	2 prittany spears

Spelling Corrections

Peter Norvig



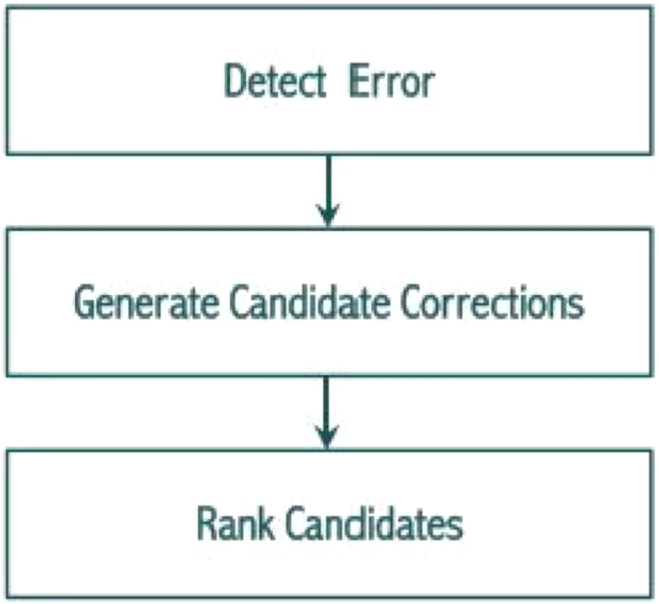
Python spelling corrector in 22 lines of code

- Simple, but still 80-90 correct

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

Spelling Corrections

Peter Norvig



- Probabilistic approach

One or two simple edits (delete, transpose, replace, insert)

Most likely

Note: candidates have to be known words!

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

Spelling Corrections

- Non-word spelling error detection:
  - Any word not in a ***dictionary*** is an error
  - The larger the dictionary the better … up to a point
  - (The Web is full of mis-spellings, so the Web isn't necessarily a great dictionary …)
- Non-word spelling error correction:
  - Generate ***candidates***: real words that are similar to error
  - Choose the one which is best (ranking)
    - Using some estimate

Generating Candidates

- Let's look only at misspelling by a single* error
  - **Insertion** – add a letter (ther)
  - **Deletion** – remove one letter (th)
  - **Substitution** – change one letter to another (thw)
  - **Transposition** – swap two adjacent letters (teh)
- Candidate corrections: generate all possible **known** words that are made of a single error
  - Can be few errors introduced – but we are looking at known words only
  - e.g. acress -> actress, cress, acres

* Two errors can be easily introduced on top of single errors

Generating Candidates

- How to find candidates (efficiently)? Some options:
  - Run through dictionary, check edit distance (e.g. using Levenshtein distance) with each word
  - Generate all words within edit distance ≤ k (e.g., k = 1 or 2) and then intersect them with dictionary
  - Use a character k-gram index and find dictionary words that share "most" k-grams with word (e.g., by Jaccard coefficient)
    - see IIR sec 3.3.4
  - Compute them fast with a Levenshtein finite state transducer
  - Have a precomputed map of words to possible corrections
    - e.g. Norvig's list at http://norvig.com/ngrams/spell-errors.txt

Generating Candidates

- Have a precomputed map of words to possible corrections
  - e.g. Norvig's list at http://norvig.com/ngrams/spell-errors.txt

```
sufficient: sersishant, suffitionent, suficient, sirfishant, suffecient, cerfistiont, surficiant, surfishant, surfishate, surficsut, sufient, sefisent
serficint, serfisent, surfishiat, serficiant, shoofishtion, serfisemete, sherfishent, sufishunt, surfend, serficent, servishant, surfeshement, shefint
sirfishint, suffiecent, serfishont, soffoant, surficant, suffisent, servishent, surfistion, surfeshant, surficent, sufeciant, surfieshent, serfisant,
suffient, sufficant, sufficent*2, surefish, suffeciant, sefishant, sersfinhet, serficant, surfitiont, suficant, serfercute, saficient, sufishent,
serfishent, sufishant, suffichent, sufficiant, sefilshont, sifishant, sufisant, sufficuint, surfficant, survition, surfinent, suffishant, serfichant,
surficient, sirfisant, suficent, surfishent, shurfistent, serfishant, sutticant, suficiant
ensuring: i
last: lars, laste, lass, lost, lats, lates, lnst, list, lorst, late, lorts, lust, lasd, lase, lot, lis, lat, lsat, larst
repetition: repatition, repetion, repitition*2
present: prosent, preasent, presant
abandoned: abondoned
fearful: fearfull

underwear: underware
language: langwage, languge, langauge, laguage
ministry: ministery
listings: listsings
```

- How to collect these?

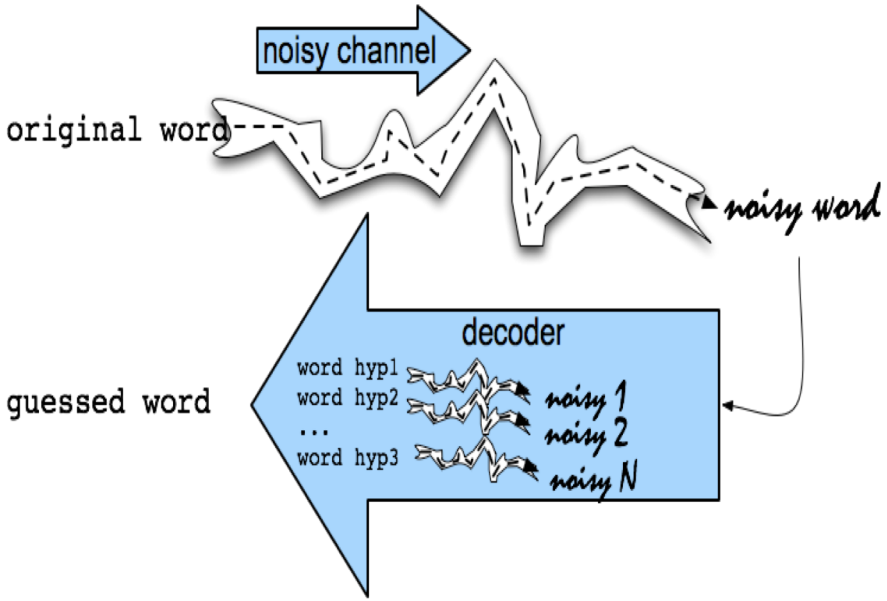
IR paradigm

- We'd like to get the best spell corrections
- Instead of finding the very best, we
  - Find a subset of pretty good corrections (say, edit distance at most 2)
  - Find the best amongst them
- *These may not be the actual best*

- This is a recurring paradigm in IR including finding the best docs for a query, best answers, best ads …
  - Find a (very) good candidate set
  - Find the top K amongst them and return them as the best

Spelling Corrections

- How to find the best candidate?
  - The most frequent of known words (in a large corpus)?
  - Model common/possible errors in typing and pronunciation
- Relies on 'noisy channel model'



Noisy channel – Bayes' Rule

- We see an observation x of a misspelled word
- Find the correct word $\hat{w}$ from vocabulary V

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w \mid x)$$

$$= \underset{w \in V}{\operatorname{argmax}} \frac{P(x \mid w) P(w)}{P(x)}$$

Bayes

$$= \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)$$

Noisy channel – Bayes' Rule

- We see an observation x of a misspelled word
- Find the correct word $\hat{w}$ from vocabulary V

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w \mid x)$$

$$= \underset{w \in V}{\operatorname{argmax}} \frac{P(x \mid w) P(w)}{P(x)}$$

Bayes

$$= \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)$$

frequency of error

word frequency

Noisy channel – Bayes' Rule

- $P(w)$ – probability of a word
- Take a big supply of words (your document collection with T tokens); let $C(w)$ = # occurrences of w

$$P(w) = \frac{C(w)}{T}$$

  - In other applications – you can take the supply to be typed queries (suitably filtered) – when a static dictionary is inadequate

Noisy channel – Bayes' Rule

- $P(w)$ – probability of a word

Prior probabilities

word	Frequency of word	*P(w)*
actress	9,321	.0000230573
cress	220	.0000005442
caress	686	.0000016969
access	37,038	.0000916207
across	120,844	.0002989314
acres	12,874	.0000318463

Counts from 404,253,213 words in Corpus of Contemporary English (COCA)

Error probability model

- $P(x \mid w)$ – probability of error x given word w.

  P(*acress* | *across*)

- Probability of making error(s) to type x when we wanted w.
  - by deletion/insertion/substitution/transposition
- How to compute that? Use a corpus of errors

```
del[x,y]:    count(xy typed as x)
ins[x,y]:    count(x typed as xy)
sub[x,y]:    count(y typed as x)
trans[x,y]:  count(xy typed as yx)
```

Confusion Matrix

Frequency of substitution of x (incorrect) by y (correct)

$y \backslash x$	A	B	C	D	E	F	G	H	. . .
A	168	1	0	2	5	5	1	3	. . .
B	0	136	1	0	3	2	0	4	. . .
C	1	6	111	5	11	6	36	5	. . .
D	1	17	4	157	6	11	0	5	. . .
E	2	10	0	1	98	27	1	5	. . .
F	1	0	0	1	9	73	0	6	. . .
G	1	3	32	1	5	3	127	3	. . .
H	2	0	0	0	3	3	0	4	
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

- We saw G when the intended character was C 36 times.
- Full matrix at: https://norvig.com/ngrams/count_1edit.txt

Error probability model

- Common spelling errors
  - Wikipedia's list of common English misspelling
  - Aspell filtered version of that list
  - Birkbeck spelling error corpus
  - Peter Norvig's list of errors (includes Wikipedia and Birkbeck, for training or testing)

Error probability model

- Other/additional methods to estimate $P(x \mid w)$
  - Simple local factors: consider the most important factors predicting an insertion, deletion, transposition
  - Unequal weightings attached to different editing operations.
  - Insertion and deletion probabilities are conditioned on context. The probability of inserting or deleting a character is conditioned on the letter appearing immediately to the left of that character.
  - Positional information is also useful (the position in the string in which the edit occurs)

artifact/artefact;
correspondance/correspondence

Google: User query spelling mistakes

488941 britney spears
40134 brittany spears
36315 brittney spears
24342 britany spears
7331 britny spears
6633 briteny spears
2696 britteny spears
1807 briney spears
1635 brittny spears
1479 brintey spears
1479 britanny spears
1338 britiny spears
1211 britnet spears
… (many, many more)

… (many, many more)
2 brynty spears
2 brythey spears
2 bryttney spears
2 btiany spears
2 btirtney spears
2 btitiney spears
2 btittny spears
2 btritany spears
2 buttney spears
2 grittney spears
2 prietny spears
2 pritany spears
2 prittany spears

Google figures out possible misspellings and their likely correct spellings by using words it finds while searching the web and processing user queries

- **user-behavior based**

Real word spelling errors

- For each word *w*, generate candidate set:
  - Find candidate words with similar ***pronunciations***
  - Find candidate words with similar ***spellings***
  - Include *w* in candidate set
- Choose best candidate
  - Noisy Channel view of spell errors
  - Context-sensitive – so have to consider whether the surrounding words "make sense"
  - *Flying form Heathrow to LAX → Flying from Heathrow to LAX*

Summary

- Preparing for indexing
  - Document conversion
  - Language/domain identification
  - Tokenization
  - Normalisation
    - Case folding
    - Lemmatization
    - Stemming
  - Spelling corrections

Language/domain dependent

MANCHESTER 1824
The University of Manchester

Materials to read

- Sections 2.1 and 2.2 in Manning et al., **Introduction to Information Retrieval** (https://nlp.stanford.edu/IR-book/)

- Peter Norvig: **How to Write a Spelling Corrector** https://norvig.com/spell-correct.html

- Additional reading for those more interested in this topic are available on Blackboard.