# File Organization & Spelling Detection

Jacques Savoy
Université de Neuchâtel

P. M. Nugues: *An Introduction to Language Processing with Perl and Prolog*. Springer. Berlin

## Outline

- Dictionary Look-up
  - constraint
  - evaluation
- Hash functions for strings
  - Simple Examples
  - Effective Hash Function
- Spelling Error Detection & Correction



## Dictionary Look-up

How can we store *n* words (our dictionary) in the main memory (or part in the memory – word -, part on disk – definition -) and have fast (very fast) access to them?

Find if "cat" is in the list

| Word |
|---|
| an |
| in |
| money |
| … |
| Zipf |



## Dictionary Look-up

Using a sequential search (average)
$O(n) = (n+1)/2$ (worst $O(n)=n$)

Using a dichotomic search
$O(n) = (\log(n)+1)/2 = \log_2(n)-1$
(worst $O(n) = \log_2(n)$)

Requirements:
$O(1)$ access and update time
$O(n)$ space

| Word |
|---|
| an |
| in |
| money |
| … |
| Zipf |

## Dictionary Look-up

If we have *n* places (locations, bins) and if each string "choses" *randomly* one of this location, what is the probability of selecting a given *i*th place?

*Prob* = 1/*n*

And another place than the *i*th?

*Prob* = 1 – (1/*n*)

What is the probability that *k* out of *n* strings chose the same location?

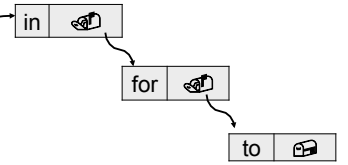
| | Word |
|---|---|
| 1 | an |
| 2 | in |
| 3 | |
| 4 | money |
| … | … |
| *n* | Zipf |

5

## Dictionary Look-up

If two (or more) strings select the same location, we store them using a linked chain

| | Word |
|---|---|
| 1 | an |
| 2 | |
| 3 | |
| 4 | money |
| … | … |
| *n* | Zipf |




6

## Dictionary Look-up

What is the probability that *k* out of *n* strings chose the same location?

In this case, *k* words select the same *i*th place, and *n*-*k* select other places, and for *k* = 0, 1, …, *n*

$$Prob[k] = \left(\frac{1}{n}\right)^k \left(1-\frac{1}{n}\right)^{n-k} \binom{n}{k}$$
$$= \left(\frac{1}{n}\right)^k \left(1-\frac{1}{n}\right)^{n-k} \frac{n!}{(n-k)!\cdot k!}$$

7

## Dictionary Look-up

If *n* is reasonably large, $Prob[k] = \frac{1}{(e \cdot k!)}$

and with *n* = 1,000,

Prob[*k*=0] = 0.37
Prob[*k*=1] = 0.37
Prob[*k*=2] = 0.18
Prob[*k*=3] = 0.06
Prob[*k*=4] = 0.015
Prob[*k*=5] = 0.003

Thus placing 1,000 random words into 1,000 bins, about 37% of the bins will be empty, 37% will have only one word, 18% will have two words, … only 0.3% will have 5 words.

Verify this using `www.random.org`.

8

## Values from Random.org

Here are your random numbers:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 2 | 3 | 4 | 5 | 6 | 1 | 4 | 6 | 6 |
| 1 | 4 | 3 | 4 | 5 | 6 | 5 | 1 | 3 | 3 |
| 3 | 5 | 6 | 3 | 4 | 2 | 3 | 2 | 6 | 6 |
| 5 | 4 | 3 | 5 | 6 | 1 | 1 | 5 | 5 | 2 |
| 4 | 3 | 1 | 5 | 6 | 5 | 1 | 3 | 2 | 6 |
| 4 | 2 | 3 | 5 | 3 | 2 | 2 | 6 | 4 | 2 |
| 5 | 6 | 6 | 6 | 6 | 3 | 6 | 5 | 4 | 5 |
| 2 | 1 | 4 | 3 | 4 | 5 | 3 | 2 | 2 | 5 |
| 4 | 6 | 3 | 5 | 3 | 4 | 4 | 6 | 3 | 2 |
| 2 | 6 | 2 | 3 | 4 | 5 | 5 | 6 | 6 | 6 |

9

## Dictionary Look-up

With 37% of empty space and a maximum chain length of 5, we may consider using fewer bins (or more strings).

Define the load factor as $\alpha = m/n$,
with $m$ the number of elements ($m$ could be > $n$)
and $n$ the number of bins

$$Prob[k] = \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{m-k} \binom{m}{k} \approx \frac{\alpha^k}{e^{\alpha} \cdot k!}$$

$$\text{with } \left(1 - \frac{1}{m}\right)^n = \left(1 - \frac{\alpha}{n}\right)^n \approx e^{-\alpha} \quad \text{if } n \to \infty$$

10

## Dictionary Look-up

If $n$ = 1,000 and $m$ = 2,000 ($\alpha$ = 2), we obtained

| | |
|---|---|
| Prob[$k$=0] = 0.13 | Prob[$k$=4] = 0.09 |
| Prob[$k$=1] = 0.27 | Prob[$k$=5] = 0.03 |
| Prob[$k$=2] = 0.27 | Prob[$k$=6] = 0.01 |
| Prob[$k$=3] = 0.18 | Prob[$k$=7] = 0.003 |

Thus placing 2,000 random words into 1,000 bins, about 13% of the bins will be empty, 27% will have only one word, 27% will have two words, … only 0.3% will have 7 words.

11

## Dictionary Look-up

But … problem:

"It is *theoretically impossible* to define a hash function that creates random data from the non-random data in actual files". (D. Knuth)

Thus it seems that we cannot produce a good hashing (perfectly uniform random).

Really true?

From a word (string) S = $c_1,c_2,\ldots c_k$,
we must define a function that return a value hash(S) between 0 and $n$-1 (the location in our table)

12

## Hash Function

From a string S = $c_1,c_2,\ldots c_k$,
return a value hash(S) between 0 and *n*-1

Hash functions examples:

1. `[c1·26 + c2]·10 + Min {k-2; 9}`

2. `[c1·256^2 + c2 ·256^1 + c3] mod 256^2+1`

3.
```
h[0] := 0;
for i in 1..k loop
   h[i] := (h[i-1] + c[i]) mod n;
end loop
return h[k];
```



## Hash Function

- Example with function
  `[26·c1 + c2]·10 + Min {k-2; 9}`

  with "house"
  we obtain [26·8 + 15]·10 + Min {5-2; 9}
  = [208 + 15]·10 + 3 = 2,233

  with "horse"
  we obtain [26·8 + 15]·10 + Min {5-2; 9}
  = [208 + 15]·10 + 3 = 2,233



## Hash Function

Better Hash function (Pearson, 1990)
input: from a string S = $c_1,c_2,\ldots c_k$
output: return a value between 0-255

```
h[0] := 0;
for i in 1..k loop
  h[i] := T[ h[i-1] xor c[i] ];
end loop
return h[k];
```

Where `T[0..255]` is a table of 256 randomish bytes.
(generated using $x_{i+1} \equiv a \cdot x_i + c \bmod m$)



## Random Number Generator

Generate a random sequence of values
$x_{i+1} \equiv (a \cdot x_i + c) \bmod m$
If m = 10, $x_0$ = 7; c = 7; a = 7
we generate → 6, 9, 0, 7, 6, 9, 0, 7, …

Conditions:
*c* and *m* are relatively prime;
*b*=*a*-1 is multiple of *p*, for every prime *p* dividing *m*;
*b* is a multiple of 4 if *m* is a multiple of 4
Ex: m = 256, $x_0$ = 7; c = 71; a = 53

## Hash Function

Better Hash function (Pearson, 1990)
from a string S = $c_1,c_2,\ldots c_k$
return a value between 0-255
If you need a hash value greater than 255?

From the original string S, we have H1
Slightly modify S, and we obtain H2
Then concatenate H1 and H2 to obtain

| H1 | H2 |
|---|---|

and a hash value of 16 bits

17

## Scrabble

Can you find all possible words for a Scrabble player?

Represent the seven (or less) letters according to their alphabetic order. For example, from the list RBTIOOC, we form the word BCIOORT.

We use this word to search into the dictionary… because to create the dictionary we have applied the same reordering method (but keep the correct form)

Possible answer: ROBOTIC.

| | Word |
|---|---|
| 1 | adijnr (jardin) |
| 2 | bcioort (robotic) |
| 3 | |
| 4 | emnoy (money) |
| … | … |
| *n* | fipz (Zipf) |

18

## Scrabble

How to generate list of possible words with less than seven letters?

Remove one by one the letters you have.

More than seven?

Eight letters but in the second position a "P"

| | Word |
|---|---|
| 1 | adijnr (jardin) |
| 2 | bcioort (robotic) |
| 3 | |
| 4 | emnoy (money) |
| … | … |
| *n* | fipz (Zipf) |

19

## Spelling Error Detection

- Very simple to have a spelling error detection, simple dictionary look-up
- Just one dictionary? Consider the Zipf's law

| word | freq | rel freq | cumul freq |
|---|---|---|---|
| 1. the | 69,975 | 0.0700 | 0.0700 |
| 2. to | 65,365 | 0.0654 | 0.1353 |
| 3. be | 39,175 | 0.0392 | 0.1745 |
| 4. of | 36,432 | 0.0364 | 0.2109 |
| 5. and | 28,872 | 0.0289 | 0.2398 |
| 6. a | 23,129 | 0.0231 | 0.2629 |
| 7. in | 21,337 | 0.0213 | 0.2843 |
| **8. he** | **19,427** | **0.0194** | **0.3037** |
| 9. have | 12,458 | 0.0125 | 0.3162 |
| 10. it | 10,943 | 0.0109 | 0.3271 |
| 11. for | 9,495 | 0.0095 | 0.3366 |
| 12. i | 8,388 | 0.0084 | 0.3450 |

20

## Spelling Error Detection

Just one dictionary?

- small list of very frequent words (e.g., 200-500) with 8 words, we may cover 30% of all tokens
- list of document-specific words (e.g., 2,000)
- complete dictionary (e.g., 20,000-100,000)

In the Brown corpus (50,406 words), 50% of them appear only once or twice.

A dictionary of 8,000 words would represent 90% of the words

21

## Spelling Error Correction

- Given a word, check if it is correctly spelled, not problem, otherwise suggest some alternatives
  "hosre" → "horse"
- In mean, 1 word out of every 200 mistyped words would accidentally become another word.
  But
- Frequent words tend to be shorter
  Short words are more likely to be undetected
- Example: from the 784 (28x28) possible two-letter words (A..Z'-), 431 are valid (55%).
- Thus most undetected errors come from the short frequent words (Peterson, 1986)

22

## Spelling Error Correction

Example:

| | |
|---|---|
| "afte" (*) → | "after", "fate", "aft", "ate", "ante" |
| "dialy"(*) → | "daily", "diary", "dials", "dial", "dimly" |
| "poice"(*) → | "police", "price", "voice", "poise", "pice", "ponce", "poire" |

Focus on performance errors
not on competence
not on syntax errors

23

## Type of Errors

Assuming 28 characters (A..Z-'), and a word of *n* letters,

| | |
|---|---|
| Transposition | *n*-1 |
| Remove one letter | *n* |
| Replace one letter | 27 · *n* |
| Add one extra letter | 28 · (*n*+1) |
| Total | 57 · n + 27 |

(Peterson, 1986)

24

## Spelling Error Correction

- Given an input word,
  If it is in the dictionary, return OK
  otherwise
    generate all alternatives considering
      removing a letter,
      replacing a letter,
      transposing two adjacent letters,
      adding one extra letter
  and check if these alternatives are in the dictionary,
  return the list of possible alternatives

25

## Soundex (1918)

Transform a word (name) into a Soundex code (1 letter + 3 digits)

1. The first character is the first letter of the name
2. Each letter are assigned the following digit

| | | | |
|---|---|---|---|
| 0 | A E I O U H W Y | 1 | B F P V |
| 2 | C G J K Q S X Z | 3 | D T |
| 4 | L | 5 | M N |
| 6 | R | | |

3. The '0' are ignored
4. Runs of the same digit are reduced to a single digit
5. Truncated the code to one letter and three digits

26

## Soundex (1918)

Examples

Dickson → D022205 → D2225 → D25
Dixon → D0205 → D25

Bush → B020 → B2
Busch → B0220 → B22 → B2

Muller → M04406 → M446 → M46
Miller → M04406 → M446 → M46

Rodgers → R032062 → R3262 → R326
Rogers → R02062 → R262

Hodgson → H032205 → H3225 → H325
Dodgson → D032205 → D3225 → D325

27

## New Spelling in French

| Old spelling | new spelling | Old spelling | new spelling |
|---|---|---|---|
| pique-nique | piquenique | a priori | apriori |
| porte-clé | porteclé | statu quo | statuquo |
| auto-stop | autostop | hot-dog | hotdog |
| mille-feuille | millefeuille | cow-boy | cowboy |
| ping-pong | pingpong | week-end | weekend |
| train-train | traintrain | | |
| québecois | québécois | crémerie | crèmerie |
| événement | évènement | diesel | diésel |
| edelweiss | édelweiss | revolver | révolver |
| veto | véto | île | ile |
| chariot | charriot | combatif | combattif |
| eczéma | exéma | nénuphar | nénufar |
| relais | relai | | |

28

## Examples in French

Accord du participe passé

| | |
|---|---|
| sans auxiliaire | les chats **lavés** mangent. |
| avoir (c.o.d.) | Vous avez **envoyé** une lettre. Je l'ai bien **reçue.** |
| et infinitif | L'histoire que j'ai **entendu** raconter.<br>La cantatrice que j'ai **entendue** chanter.<br>(les deux sont correctes, dès 1990) |
| et en (invar.) | J'ai **cueilli** des fraises et j'en ai **mangé**.<br>Ses ordres, s'il en a **donnés** ne me sont pas **parvenus**. |
| être (sujet) | Les chats sont **lavés.** |
| "se" est c.o.d. | Les chats se sont **lavés.** |
| "se" est c.o.ind. | Les chats se sont **lavé** les pattes. |
| que faire ? | Les chats de Marie que Jean a **vus** (**vue**) sont **lavés.**<br>Jean a vu les chats (ou Marie)? |

29

## Examples in French

"The correct spelling is French"

"faux témoignage" & "faux fuyant"

"chausse-pied" &"portemanteau"

"coupe-cigare**s**" &"coupe-circuit"

"ratio**n**aliste" &"ratio**nn**el" (rationalis)

"Chinois" & "Anglois" → "Anglais" (Voltaire)

"arcs-boutants" (1694)

"arc-boutants" (1718)

"arcs-boutants" (1740)

"arc-boutants" (1762)

"arcs-boutants" (1835)

30

## Examples in French

- No clear relationship with the etymology
  - hauteur → altus
  - pomme→ poma
  - huile→ olea
- Diacritics used to specify the pronunciation?
  - pôle and police → no real difference
  - pêcheurs and pécheurs → no real difference
- Diacritics used to indicate a missing letter?
  - âne → asinus
  - théâtre → theatrum
- Arbitrary (apercevoir vs. apparaître) , no logic

31

## Examples in French

Private spelling vs. correct spelling

Madame de Sévigné (XVIIIe)

"Monsieur vous me permettres de souhaitter la paix car ie trouve uec vostre permission quune heure de Conuersation vaut mieux que cinquante lettres, quand vous seres icy etque iauray lhonneur devous voir ievous ferey demeurer dacort quela querre est vne fort sottechose …"

32

## Examples in French

**Mais qu'es ce que tu parles**

Mais qu'es ce que tu parles tu connais l'équipe pour parlé comme ça franchement gardes tes commentaires pour toi tu connais rien alors la rammène pas trop merci !

**Toi t'est un qui n'a rien vue**

Toi t'est un qui n'a rien vue contre le FCC. Je te rappel que contre le fcc sur 11 joueur sur le terrain il y avait que 4 joueurs licencié. Sinon c'était des essais pour voir le niveau des nouveaux arrivées. Ce qui s'est passé dans le vestiaire est un autre problème et les sanctions ont été prise par le club.

**FC Bosna**

fc bosna j aisspaire que momo revienne sest le seule qui peux nous faire monté comme entreneure



## Conclusion

- Efficient access to a list of words is crucial for many NLP applications
- Efficient hash functions for strings exist
- A simple spelling detection & correction is easy to build on top of a dictionary
- Remind the Zipf's law



## References


P.A.V. Hall, G.R. Dowling: Approximate String Matching. *ACM Computing Surveys*, 12(4), 1980, 381-402.

K. Kukich: Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys*, 24(4), 1992, 377-439.

G. Navarro: A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, 33(1), 2001, 31-88.

Pearson, P.K. (1990). Fast Hashing of Variable-Length Text Strings. *Communications of the ACM*, 33(6), 677-680.

Peterson, J.L. (1980). Computer Programs for Detecting and Correcting Spelling Errors. *Communications of the ACM*, 23(12), 676-687.

Peterson, J.L. (1986). A Note on Undetected Typing Errors. *Communications of the ACM*, 29(7), 633-637.




## Soundex (1918)

Transform a word (name) into a Soundex code (1 letter + 3 digits)

1. The first character is the first letter of the name
2. Each letter are assigned the following digit

| | | | |
|---|---|---|---|
| 0 | A E I O U H W Y | 1 | B F P V |
| 2 | C G J K Q S X Z | 3 | D T |
| 4 | L | 5 | M N |
| 6 | R | | |

3. The '0' are ignored
4. Runs of the same digit are reduced to a single digit
5. Truncated the code to one letter and three digits

## Soundex (1918)

Examples

Dickson → D022205 → D2225 → D25
Dixon → D0205 → D25

Bush → B020 → B2
Busch → B0220 → B22 → B2

Muller → M04406 → M446 → M46
Miller → M04406 → M446 → M46

Rodgers → R032062 → R3262 → R326
Rogers → R02062 → R262

Hodgson → H032205 → H3225 → H325
Dodgson → D032205 → D3225 → D325

37

## Examples in French

Accord du participe passé

| | |
|---|---|
| sans auxiliaire | les chats **lavés** mangent. |
| avoir (c.o.d.) | Vous avez **envoyé** une lettre. Je l'ai bien **reçue.** |
| et infinitif | L'histoire que j'ai **entendu** raconter. |
| | La cantatrice que j'ai **entendue** chanter. |
| | (les deux sont correctes, dès 1990) |
| et en (invar.) | J'ai **cueilli** des fraises et j'en ai **mangé**. |
| | Ses ordres, s'il en a **donnés** ne me sont pas **parvenus**. |
| être (sujet) | Les chats sont **lavés.** |
| "se" est c.o.d. | Les chats se sont **lavés.** |
| "se" est c.o.ind. | Les chats se sont **lavé** les pattes. |
| que faire ? | Les chats de Marie que Jean a **vus** (**vue**) sont **lavés.** |
| | Jean a vu les chats (ou Marie)? |

38

## Examples in French

"The correct spelling is French"
"faux témoignage" & "faux fuyant"
"chausse-pied" &"portemanteau"
"coupe-cigare**s**" &"coupe-circuit"
"ratio**n**aliste" &"ratio**nn**el" (rationalis)
"Chinois" & "Anglois" → "Anglais" (Voltaire)
"arcs-boutants" (1694)
"arc-boutants" (1718)
"arcs-boutants" (1740)
"arc-boutants" (1762)
"arcs-boutants" (1835)

39

## Examples in French

- No clear relationship with the etymology
  - hauteur → altus
  - pomme→ poma
  - huile→ olea
- Diacritics used to specify the pronunciation?
  - pôle and police → no real difference
  - pêcheurs and pécheurs → no real difference
- Diacritics used to indicate a missing letter?
  - âne → asinus
  - théâtre → theatrum
- Arbitrary (apercevoir vs. apparaître) , no logic

40

## Examples in French

Private spelling vs. correct spelling

Madame de Sévigné (XVIIIe)

"Monsieur vous me permettres de souhaitter la paix car ie trouve uec vostre permission quune heure de Conuersation vaut mieux que cinquante lettres, quand vous seres icy etque iauray lhonneur devous voir ievous ferey demeurer dacort quela querre est vne fort sottechose …"

41

## Examples in French

**Mais qu'es ce que tu parles**

Mais qu'es ce que tu parles tu connais l'équipe pour parlé comme ça franchement gardes tes commentaires pour toi tu connais rien alors la rammène pas trop merci !

**Toi t'est un qui n'a rien vue**

Toi t'est un qui n'a rien vue contre le FCC. Je te rappel que contre le fcc sur 11 joueur sur le terrain il y avait que 4 joueurs licencié. Sinon c'était des essais pour voir le niveau des nouveaux arrivées. Ce qui s'est passé dans le vestiaire est un autre problème et les sanctions ont été prise par le club.

**FC Bosna**

fc bosna j aisspaire que momo revienne sest le seule qui peux nous faire monté comme entreneure

42