02.05 Spelling Similarity: Edit Distance

NLP

Text Similarity

Spelling Similarity: Edit Distance

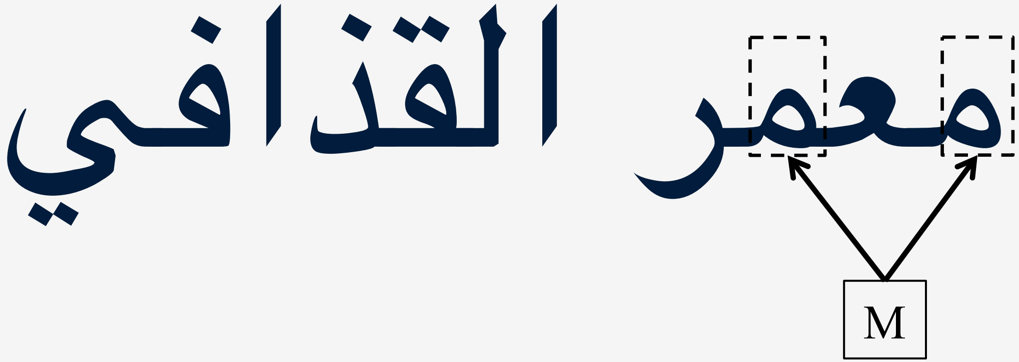
Spelling Similarity

- Typos:
  - Brittany Spears -> Britney Spears
  - Catherine Hepburn -> Katharine Hepburn
  - Reciept -> receipt
- Variants in spelling:
  - Theater -> theatre

Who Is This?

معمر القذافي

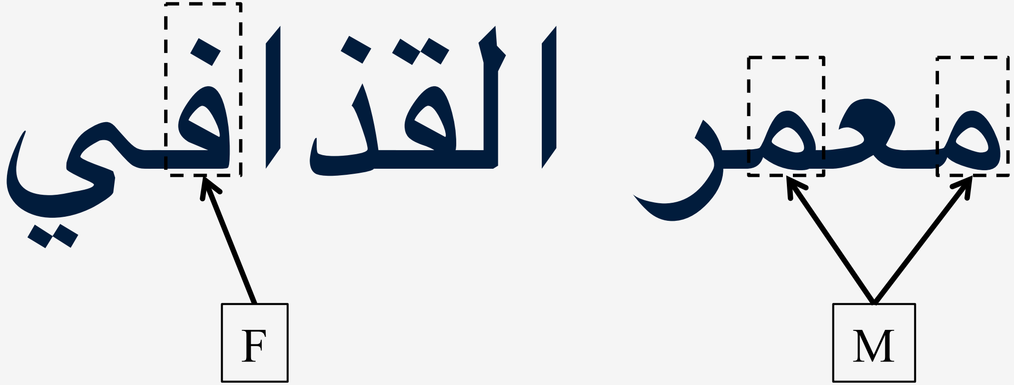
Hints



معمر القذافي

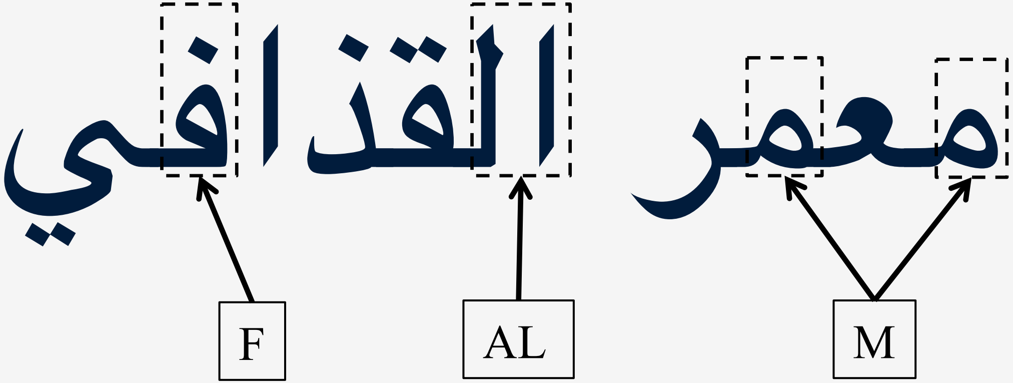
M

Hints



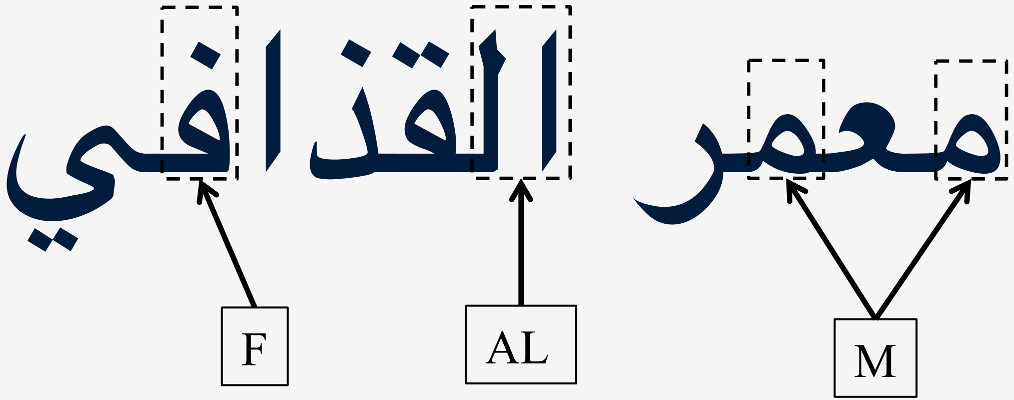
معمر القذافي

Hints



معمر القذافي

Hints



Muammar (al-)Gaddafi, or **Moamar Khadafi,** or …

Quiz

How many different transliterations can there be?

m
u o
a
m mm
a e
r

el al El Al ø

Q G Gh K Kh
a e u
d dh ddh dhdh th zz
a
f ff
i y

A Lot!

m u o a m mm a e r		el al El Al ø		Q G Gh K Kh a e u d dh ddh dhdh th zz a f ff i y		
8	x	5	x	360	=	14,400

Edit Operations

- behaviour – behavior (insertion/deletion) (“al”)
- string – spring (substitution) (“k”–”q”)
- sleep – slept (multiple edits)

Levenshtein Method

- Based on dynamic programming
- Insertions, deletions, and substitutions usually all have a cost of 1.

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	5								

Recurrence Relation

- Definitions
  - $s_1(i)$ - ith character in string s_1
  - $s_2(j)$ - jth character in string s_2
  - $D(i,j)$ - edit distance between a prefix of s_1 of length i and a prefix of s_2 of length j
  - $t(i,j)$ - cost of aligning the ith character in string s_1 with the jth character in string s_2

- Recursive dependencies

```
D(i,0)=i
D(0,j)=j
D(i,j)=min[
    D(i-1,j)+1
    D(1,j-1)+1
    D(i-1,j-1)+t(i,j)
    ]
```

- Simple edit distance:

  $t(i,j)=0$ *iff* $s_1(i)=s_2(j)$

  $t(i,j)=1$, *otherwise*

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1							
r	2								
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1						
r	2								
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2						
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	**1** ↘	**2** ↓	3	4	5	6	7
r	2	2	**2** →						
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	4

Edit Transcript

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	4

Other Costs

- Damerau modification
  - Swaps of two adjacent characters also have a cost of 1
  - E.g., Lev(“cats”,”cast”) = 2, Dam(“cats”,”cast”) = 1

Quiz

- Some distance functions can be more specialized.
- Why do you think that the edit distances for these pairs are as follows?
  - Dist (“sit clown”,“sit down”) = 1
  - Dist (“qeather”,”weather”) = 1, *but* Dist (“leather”,”weather”) = 2

Quiz Answers

- Dist(“sit down”,”sit clown”) is lower in this example because we want to model the type of errors common with optical character recognition (OCR)
- Dist(“qeather”,”weather”) < Dist(“leather”,”weather”) because we want to model spelling errors introduced by “fat fingers” (clicking on an adjacent key on the keyboard)

Quiz: Guess the Language

AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAAACT
TTCAACAATGGATCTCTTGGTTCCGGC

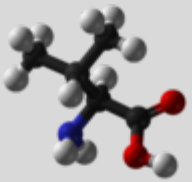
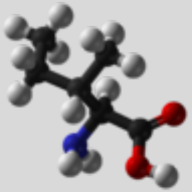
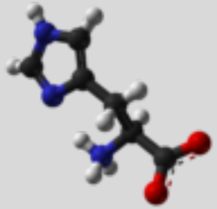
Quiz Answer

- This is a genetic sequence (nucleotides AGCT)

```
>U03518 Aspergillus awamori internal transcribed spacer 1 (ITS1)
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAAACT
TTCAACAATGGATCTCTTGGTTCCGGC
```

Other Uses of Edit Distance

- In biology, similar methods are used for aligning non-textual sequences
  - Nucleotide sequences, e.g., GTTCGTGATGGAGCG, where A=adenine, C=cytosine, G=guanine, T=thymine, U=uracil, “-”=gap of any length, N=either one of ACGTU, etc.
  - Amino acid sequences, e.g., FMELSEDGIEMAGSTGVI, where A=alanine, C=cystine, D=aspartate, E=glutamate, F=phenylalanine, Q=glutamine, Z=either glutamate or glutamine, X=“any”, etc. The costs of alignment are determined empirically and reflect evolutionary divergence between protein sequences. For example, aligning V (valine) and I (isoleucine) is lower-cost than aligning V and H (histidine).

Valine	Isoleucine	Histidine
		

External URLs

- Levenshtein demo
  - http://www.let.rug.nl/~kleiweg/lev/
- Biological sequence alignment
  - http://www.bioinformatics.org/sms2/pairwise_align_dna.html
  - http://www.sequence-alignment.com/sequence-alignment-software.html
  - http://www.ebi.ac.uk/Tools/msa/clustalw2/
  - http://www.animalgenome.org/bioinfo/resources/manuals/seqformats

NACLO Problem

- “Nok-Nok”, NACLO 2009 problem by Eugene Fink:
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-B.pdf

Solution to the NACLO Problem

- “Nok–Nok”
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-BS.pdf

NACLO Problem

- "The Lost Tram", NACLO 2007 problem by Boris Iomdin:
  - http://www.naclo.cs.cmu.edu/problems2007/N2007-F.pdf

Solution to the NACLO problem

- “The Lost Tram”
  - http://www.naclo.cs.cmu.edu/problems2007/N2007-FS.pdf

NLP