They're not equal!

How the <<expletive>> do you expect me to find a match?

kyle.burton@gmail.com
http://asymmetrical-view.com/
PLUG August 6th 2008

Your Presenter

- Kyle Burton
- Algorithmics Inc.
- Philadelphia
- Geek

How Do *You* Spell ...?

- De Morgan
- Di Morgen
- D'Morgun
- Demorgyn
- De Murgen
- Dy Moregan
- Dy Murgan
- Da Myrgn

Er, So How Can You Find a Match?

Fuzzy Matching, That's how

- Partial Matching
- Phonetic Encodings
- String Similarity Metrics

How'd We Get Here?

- US Census Bureau
  - William Winkler (not the Fonz)

How'd We Get Here?

- Record Linkage, aka Duplicate Detection
  - I used to work for a Company that did this
  - (it is a complex problem domain)
- DNA Comparison and Sequence Alignment
  - (I don't do this, but it sounds cool on Tv)

- Partial Matching
- Phonetic Encodings
- String Similarity Metrics

Partial Matching

- ‘False’ Fuzziness: prefix, suffix, infix
- SQL’s ‘%’ operator
- n-grams (bi-grams, tri-grams)
  - foobar => foo, oob, oba, bar
  - This is infix in disguise

Partial Matching

- Indexable - fast lookup / search
- Fixed Degree of ‘Fuzziness’
- Doesn’t scale based on difference
  - Any hit and you have a match
  - Can’t Measure *Quality* of the match

- Partial Matching
- Phonetic Encodings
- String Similarity Metrics

Phonetic Encodings

- Soundex, NYSIIS, Double Metaphone
- ‘hash’ of input
- Indexable
- Fixed fuzziness, one or two degrees

Soundex

- Keep the First Character
- Convert Vowels (and some soft consonants) to a Zero [AEHIOUWY]
- [BFPV] => 1
- [CGJKQSXZ] => 2
- and so on...

Soundex

- B635 <= Burton, Barton
- G232 <= Gwozdziewycz, Gwozdz
- D562 <= De Morgen, Di Morgen, D'Morgun, Demorgyn, De Murgen, Dy Moregan, Dy Murgan, Da Morgan, Da Myrgn

NYSIIS

- New York State Immunization Information System
- Circa 1970
- 2.7% better than Soundex
- Targeted at Names

NYSIIS

- Drop Trailing SZs
- ^MAC => MC
- ^PF => F
- and so on (lots of special rules)

NYSIIS

BARTAN	Burton, Barton
GWASDSAC	Gwozdziewycz
GWASD	Gwozdz
DAGN	Da Myrgn, Demorgyn
DNARAGAN	Dy Moregan
DNARGAN	Da Morgan, De Morgen, De Murgen, Di Morgen, Dy Murgan
NARGAN	D'Morgun

Double Metaphone

- Lawrence Phillips, derived from Metaphone
- Primary and alternate encodings are possible
- Helps account for irregularities across multiple languages
  - eg: English, Slavic, Germanic, Celtic, Greek, French, Italian, Spanish...(atw)

Double Metaphone

PRTN	Burton, Barton
KSTS	Gwozdz
KSTSS	Gwozdziewycz
TMRJTMRK	De Morgen, De Murgen, Demorgyn, Di Morgen
TMRK	Da Morgan, Dy Moregan, Dy Murgan, D'Morgun
TMRNTMRK	Da Myrgn

How Do They Compare?

- Soundex, Metaphone, Nysiis
- US Census Name File
  - http://www.census.gov/genealogy/names/names_files.html
  - Useless Fact: 1% of the unique names cover 50% of population
  - Aalderink is the least frequent
  - Smith is the most frequent

US Census Name Files

- dist.all.last:
  - SMITH 1.006 1.006 1
  - JOHNSON 0.810 1.816 2
  - WILLIAMS 0.699 2.515 3
- dist.male.first
  - JAMES 3.318 3.318 1
  - JOHN 3.271 6.589 2
  - ROBERT 3.143 9.732 3

Phoneta-death-battle!

- Last Names: 88,799
- Soundex: 4,599 => 1/20th
- Metaphone: 18,317 => 1/5th
- NYSIIS: 31,149 => 1/3rd

(sorry, got a little carried away for a second there)

Phonetic Can't Catch Everything

- Transcription Errors
  - Typos
- Transmission Errors
  - Data Corruption
- Abbreviations, Contractions Acronyms (oh my!)

- Partial Matching
- Phonetic Encodings
- String Similarity Metrics
  - Indexing Strategies

Get Your Fuzzy On

- Edit Distance and Variants
  - Levenshtein
  - Wu-Manber
    - Jaro-Winkler
- Ascii Frequency
- Keyboard Distance
- Many, Many Others

Edit Distance

- Vladimir Levenshtein 1965
- “the minimum number of operations needed to transform one string into the other”
- An operation is an insertion, deletion, or substitution of a single character

Edit Distance

- Given S1 and S2
- Initialize a Matrix of S1.len+1 x S2.len+1
- Initialize First Row With Default Costs:
  - (0,1,2,3,...,S1.len)
- Initialize First Column With Defaults:
  - (0,1,2,3,...,S2.len)
- Then...er, it'll be easier to just show you

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1						
A	2						
R	3						
T	4						
O	5						
N	6						

There, that's better

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0					
A	2						
R	3						
T	4						
O	5						
N	6						

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0	1				
A	2						
R	3						
T	4						
O	5						
N	6						

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0	1	2			
A	2						
R	3						
T	4						
O	5						
N	6						

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0	1	2	3		
A	2						
R	3						
T	4						
O	5						
N	6						

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0	1	2	3	4	
A	2						
R	3						
T	4						
O	5						
N	6						

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0	1	2	3	4	5
A	2						
R	3						
T	4						
O	5						
N	6						

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0	1	2	3	4	5
A	2	1	1	2	3	4	5
R	3						
T	4						
O	5						
N	6						

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0	1	2	3	4	5
A	2	1	1	2	3	4	5
R	3	2	2	1	2	3	4
T	4						
O	5						
N	6						

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0	1	2	3	4	5
A	2	1	1	2	3	4	5
R	3	2	2	1	2	3	4
T	4	3	3	2	1	2	3
O	5						
N	6						

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0	1	2	3	4	5
A	2	1	1	2	3	4	5
R	3	2	2	1	2	3	4
T	4	3	3	2	1	2	3
O	5	4	4	3	2	1	2
N	6						

Edit Distance

		B	U	R	T	O	N
	0	1	2	3	4	5	6
B	1	0	1	2	3	4	5
A	2	1	1	2	3	4	5
R	3	2	2	1	2	3	4
T	4	3	3	2	1	2	3
O	5	4	4	3	2	1	2
N	6	5	5	4	3	2	1

Voila!

Edit Distance

- Wanna see it again?

Edit Distance

		B	A	B	Y
	0	1	2	3	4
B	1	0	1	2	3
O	2	1	1	2	3
B	3	2	2	1	2
B	4	3	3	2	2
Y	5	4	4	3	2

Edit Distance

- De Morgan vs De Morgan 0 100%
- De_Morgan vs D'Morgun 3 64%
- De_Morgan vs Demorgyn 3 64%
- De Morgan vs De Murgen 2 77%
- De Morgan vs Dy Moregan 2 78%

Text Brew

- Edit Distance Variant
- Configurable Costs:
  - Match, Insert, Delete, Substitute
- Computes Edit Path
  - You Can Do Interesting Things with the Edit Path

Text Brew

		B	A	B	Y
	0.0	1.0 INS	2.0 INS	3.0 INS	4.0 INS
B	1.0 DEL	0.0 MAT B=B	1.0 INS B+A	2.0 MAT B=B	3.0 INS B+Y
O	2.0 DEL	1.0 DEL O-B	1.0 SUB O/A	2.0 SUB O/B	3.0 SUB O/Y
B	3.0 DEL	2.0 MAT B=B	2.0 SUB B/A	1.0 MAT B=B	2.0 INS B+Y
B	4.0 DEL	3.0 MAT B=B	3.0 SUB B/A	2.0 MAT B=B	2.0 SUB B/Y
Y	5.0 DEL	4.0 DEL Y-B	4.0 SUB Y/A	3.0 DEL Y-B	2.0 MAT Y=Y

Text Brew

- What Else Can you Do with Text Brew?

(why, I'm glad you asked!)

Text Brew

MATCH	0
INSERT	0.1
DELETE	15
SUBSTITUTE	1
TRANSPOSE	2

Tune For Abbreviations

Text Brew

- "Hosp" vs "Hospital"

  93% Similar Brew

  67% Similar: Levenshtein

- "Clmbs Blvd" vs "Columbus Boulevard"

  94% Similar: Brew

  57% Similar: Levenshtein

Text Brew

MATCH	0
INSERT	1
DELETE	1
SUBSTITUTE	2
TRANSPOSE	0.1

Tune For Typos

Text Brew

- “Harrisburg” vs “Harrsibugr”

  98% Similar: Brew

  60% Similar: Levenshtein

- “Burton” vs “Bruton”

  98% Similar Brew

  67% Similar: Levenshtein

Text Brew

- More Ideas?
- Use the Edit Path to Score the Edits
- "Scrabble Scores"
  - Cheap: E, A, I, O, N, R, T, L, S , U
  - Costly: K, J, X, Q, Z
- Create Your Own Training Set That Fits your Data Domain

References

- http://en.wikipedia.org/wiki/Soundex
- http://en.wikipedia.org/wiki/New_York_State_Identification_and_Intelligence_System
- http://en.wikipedia.org/wiki/Double_Metaphone
- http://en.wikipedia.org/wiki/Levenshtein_distance
- http://norvig.com/spell-correct.html
- http://en.wikipedia.org/wiki/Jaro-Winkler
- http://search.cpan.org/~kcivey/Text-Brew-0.02/lib/Text/Brew.pm
- http://github.com/kyleburton/fuzzy-string

Conclusion

- You Too Can Match Fuzzily
- Partial Matches
- Phonetic Encodings
- Edit Distance Family

Fin

Thank You!

(Questions? Examples?)