CollateX normalization

Computer-supported collation with CollateX

DH2015, Sydney, 2015-06-29

Overview

- CollateX default matching
- Why you may want to override it
- How to override it

CollateX default matching

- Exact string matching
  - Near matching
- Tokenize by splitting on white space
- Punctuation marks are individual tokens
- No case normalization
- No Unicode normalization

Sample normalization overrides

- Case folding
- Unicode normalization (precomposed characters)
- Strip punctuation
- Strip markup

Soundex

- English-language surnames, 1918
- Algorithm (simplified)
  - Retain first letter
  - Delete other vowels; degeminate
  - Conflate other letters according to phonetic similarity (e.g., t/d = 3; m/n = 5)
  - Truncate or zero-pad to four characters
- Examples
  - Birnbaum B-651 (also ✓ Barenboim; also ✗ Brumble)

Soundex assumptions

- More nuanced than generic edit distance
- Character differences are not all equivalent with respect to information load
  - Consonants carry more information than vowels
- Information load may be sensitive to position
  - Beginning of word carries more information than end
  - Especially true for lexical (not morphological) searching in inflected languages

Adapting Soundex to Church Slavonic

- Neutralize variant spellings of initial vowel
  - оу, у, ѹ = у
  - ѡ, ѿ, о, ѳ = о
- Case fold, neutralize consonantal variants
  - Not always one-to-one, e.g., ц = шт
- Degeminate, delete other vowels, delete diacritics
  - Keep two letters of two-letter words
    - Higher information load
- Other conflations?
  - Knowledge based vs machine learning
- Expand abbreviations?
  - бг҃а, бг҃а, б҃а = вога (вг)
- Truncate or zero-pad (to what length?)

Soundex sample

Ch397	и	вьзвра|тить	дьщерьше	своѥе.
Ch384	и	вьзвратит͡	дьщершоу	свою.
Nbkm298	и	вьзвратити	братанитцꙋ |	своѫ
Berlin	и	вьꙗ҆звратити |	братаницꙋ	свою.

Ch397	и	взвр	дштр	св
Ch384	и	взвр	дштр	св
Nbkm298	и	взвр	бртн	св
Berlin	и	взвр	бртн	св

Two types of normalization

- Collation
  - Find alignment points
  - Coarse adjustments
  - No harm in conflating, e.g., imperfect and aorist or infinitive and supine
- Evaluation
  - Alignment points are already known
  - Finer comparisons
  - Many need to distinguish on the basis of small details

Collation after Soundex

- Greatly improved results
- Utilize forced matches
  - A B C
  - A D C
- Misses
  - Gap in alignment (no forced match)
  - Imperfect match
    - фраки ~ фраци
  - CollateX recognizes only perfect matches
    - Unable to recognize *closest match* (until last week!)

3,5

3,5

Lav	гаръмати	тавр[і]|ани.	сиѹфыа.		фраци.
Tro	гаръмати	тавриани	скуфиа		фраки
Rad	сармати	тавриани	скоуфиа	и	фраци |
Aka	сармати.	таврїани	скоуфїа.	и	фра|ци
Ipa	сармати.	тавриани. |	скуфиа.		фраци.
Xle	сармати.	таврїани.	скѵфїа		фра|ци.
Vyč	Сарѣматн,	Таврнанн,	Скуфна,		Фрацн,
Šax	Сарматн,	Таврнанн,	Скуфня,		Фрацн,
Lix	Сарѣматн,	Таврнанн,	Скуфна,		Фрацн,
α	Сарматн,	Таврнанн,	Скуфня,		Фрацн,