Fast Levenshtein

Uwe Quasthoff

Universität Leipzig
Institut für Informatik
quasthoff@informatik.uni-leipzig.de

Fast Levenshtein

Calculates Levenstein distance of <3 in O(n log(n))

In reality, unly useful for words of length >6

Example: *mein*, *Stein*, *eine*

1. step: For any word: Remove any zero, one or two characters and store the result together with the starting word.

Result: Pair (hash value, key)

mein	mein	mn	mein	ein	eine
ein	mein	mi	mein	…	…
min	mein	me	mein	ei	eine
men	mein	Stein	Stein	…	…
mei	mein	…	…	in	eine
in	mein	ein	Stein	…	…
en	mein	…	…		
ei	mein	eine	eine		

Fast Levenshtein (2)

2. step: Identical hash values come from similar keys. Sort table and group by hash value. Ignore groups of size 1. The keys in a group are candidates for Levenshtein distance 1 or 2.

```
ei    eine
ei    mein
ein   eine
ein   mein
ein   Stein
in    eine
in    mein
```

3. step: Check candidate pairs with usual Levenshtein algorithm:

```
eine – mein      distance = 2
eine - Stein     distance = 3, dropped!
mein - Stein     distance = 2
```

Which words have most similar words?

(Please guess!)

Which words have most similar words?

1-jährige, *2-jährigen* etc. have more than 600

Surnames like *Schieler, Schweier, Schieder* have more than 400:

Example: *Schieler*

Schwuler, Schöller, Scheer, Schoeller, Scheele, Schieben, Schweer, Schuller, Schaeder, Scheeßler, Schelper, Schiebler, Schieker, Schielin, Schielo, Schielt, Schicker, Schilfer, Schinkeler, Schittler, Schlegler, Schneer, Schnizler, Schoeder, Schouler, Schuhler, schiefes, scheener, scheller, schnieker, schweer, Schneier, Schnieder, Schweller, Scheerer, schienen, Schipper, Schiewe, schier, Schienen, Schmelter, schneller, Schilder, Schleyer, Schmelzer, Schielen, Schießer, Stieler, …

Observation: These are not the interesting words

Affixe of length 1 and 2

Differences at the beginning or the end of words are candidates for affixes:

affix	freq
-n	16706
-en	12460
-s	11617
-e	9739
-er	5026
-r	4300
-es	3880

affix	freq
un-	1316
ge-	1152
ab-	917
be-	880
an-	856
er-	487
zu-	412

Levenshtein Distances

- Up to distance 3 for words of a certain minimum length
- Comparison of any word with a given list of 1M word
- i.e. for small languages: For all word pairs

Outlook:

- Weighted differences according to frequency of the observations
- Iteration for longer affixes