GDEX in Sketch Engine

Jan Michelfeit

jan.michelfeit@sketchengine.co.uk

February 12, 2015

Definitions

- GDEX = Good Dictionary EXamples
- EXamples = example sentences
- Good = clean and comprehensible out of context

A bad sentence is

- too short or too long
  Hungry?
- *too specific – names, numbers*
  The new Ford Focus gives you 29 mpg.
- *too vague – anaphors*
  He didn't want to do it and she did neither.

Types of noise

- common web corpus noise
  ÅřClick here >> :::
- *typos, slang and other non-words*
  Dood, check out my nwe ride!
- *profanities, sensitive topics. . .*
  We did the Upper Decker Double Blumpkin at McDonald's.

A brief history

- 2008 – GDEX v1
  - each sentence gets a score between 0 and 1
  - rudimentary, English only, not configurable
- 2010 – GDEX v2
  - configurable, but quite complicated
  - web interface not very flexible
- 2014 – GDEX v3
  - simplified
  - human-readable/writable configuration

New configuration

```
formula: >
 (50 * is_whole_sentence()
  * blacklist(words, illegal_chars)
  * blacklist(lemmas, parsnips)
  * (min([word_frequency(w) for w in words]) > 3)
 + 20 * optimal_interval(length, 10, 14)
 + 15 * greylist(words, rare_chars, 0.1)
 + 15 * greylist(tags, pronouns, 0.1)
 ) / 100
variables:
  illegal_chars: ([<|\]\[>/\\^@])
  rare_chars: ([A-Z0-9'.,!?)(;:-])
  pronouns: PRON.*
  parsnips: ^(tory|jesus|bacon)$
```

The new format

- a simple arithmetic expression in Python
- predefined variables like length, words, tags. . .
- useful classifiers implemented as callable functions
- regular expressions

Features in Beta

- Upload your own configuration
- Check GDEX score in concordance view
- TickBox lexicography comparison