Vybrané aktuální projekty Laboratoře NLP

Pavel Šmerk, Miloš Husák

E-mail: xsmerk@fi.muni.cz, xhusak@fi.muni.cz
http://nlp.fi.muni.cz/poc_lingv/

Obsah:

- Guesser – značkování českých neznámých slov
- GDEX – Good Dictionary Examples

Guesser

Značkování českých neznámých slov

Pavel Šmerk
xsmerk@fi.muni.cz

Problém

Není možné mít všechna slova, která lze při zpracování přirozeného jazyka "potkat", předem ve slovníku. Je tedy potřeba umět řešit situaci, kdy slovo ve slovníku není.

Potřebné informace o neznámém slovu mohu odhadnout buď to z jeho větného kontextu, nebo z jeho podoby, konkrétně z jeho konce. Optimální řešení bude pochopitelně kombinovat oba přístupy.

Dále bude řeč o druhé variantě. Tedy dostanu slovo *napipetovaná*, řeknu si, aha, to je jako *malovaná*, tedy lemma *napipetovaný* podle *malovaný*, značky stejné.

Produktivita

Čeština má relativně velké množství vzorů pro ohýbání i odvozování slov. Velká část z nich ale není v současnosti "produktivních", jsou to jen pozůstatky z minulosti (*den, husa, myš, kost; dřevěný vs. umakartový, ...*), případně jsou produktivní jen nějak omezeně (*pařba x *krvácba*).

Nově se objevující *česká* slova budou tedy patřit k nějaké podmnožině produktivních vzorů. Při odhadování konkrétního vzoru budou neproduktivní vzory nadbytečné, naopak dokonce budou tvořit šum a "překážet". My ale (zatím) nemáme informaci, které vzory jsou produktivní.

Předpoklad

Můžeme předpokládat, že slovníky zahrnují všechna nepravidelně se chovající slova, která lze reálně "potkat" (pokud bychom se dostali k nějakému historickému textu, nemuselo by to tak být). Zároveň byly jistě tvořeny se snahou o co největší pokrytí reálného jazyka. To znamená, že se přidávala "nová", pravidelně se chovající slova, např. z různých oborů.

Jde si to představit i jinak: máme-li 390 000 lemmat, není možné, aby naprostá většina z nich nebyla pravidelná, to by člověk nezvládl udržet v hlavě.

Pak si lze myslet, že to, co je produktivní, poznáme podle četnosti ve stávajících datech. Cílem je tedy z dat extrahovat jen jejich produktivní (častou) část a pouze na jejím základě pak odhadovat morfologické vlastnosti neznámých slov.

Konstrukce dat pro guesser

- ze slovníku ajky (.dic) vyhodím lemmata od vzorů, u kterých je méně než 20 lemmat (tedy cca 6000 lemmat z 390 000, 55 000 slovních tvarů)
- pro všechny tvary každého lemmatu vygeneruju trojice word-lemma-tag ve formátu
  `abaraks:a::k1gMnSc2,k1gMnSc4`,
  kde nejprve je obrácený slovní tvar, dále, co musím utrhnout, a za další dvojtečkou, co musím přidat, abych dostal lemma, nakonec značky
- beru jen slova délky alespoň 7, vyhazuju eN, d3 a věci mimo k[1256], neberu wH a pod.
- setřídím podle abecedy (nikoli česky)

Konstrukce dat pro guesser

- procházím a zaznamenávám do pole, min. kořen 3 znaky, max. konec 10 znaků
- pokud v daném místě pole bylo něco jiného, zkusím vypsat. Pro výpis to musí mít aspoň 20 výskytů, jednotlivé značky aspoň 10x, pokud je něčeho víc než 20x méně než zbytku, ignoruju.
- vypsané odečítám od levých podřetězců
- setřídím podle abecedy
- slučuju stejné věci (s ohledem na počty), ignoruju vid
- setřídím retrográdně kvůli snadnému nacpání do `$re (...|ab|a)`, viz dále

Použití

- hádáná slova otáčím a zkouším `=~ /($re).{$minroot}/`
  `# $minroot == 3`
- díky "uspořádání" `/^(ab|a)/` se pro *abxyz* namatchuje *ab*, a nikoli *a*, tedy obecně nejdelší možný konec
- k nalezenému konci slova mám v hashi info o změně konce na lemma a příslušných značkách

`/nlp/projekty/rule_ind/stat/guesser.pl`, očekává vertikál, případně označkovaný něčím jako `ajka -c -` (existující značkování ponechává beze změny)

Problém: zdaleka ne všechna neznámá slova jsou česká, tedy začleněná do systému jazyka. Zejména asi nesklonné názvy, jména, případně i cizojazyčná slova či celé citáty. Hodně z nich by asi bylo možné odhadovat pomocí velkého písmena na začátku a absence diakritiky. Zatím nijak netestováno/neřešeno.

Obsah

GDEX

Good Dictionary EXamples
Dobré slovníkové příklady

Miloš Husák
xhusak@mail.muni.cz

Slovníkový příklad

crawl 161 **credit**

PHRASE **be crawling with 1** to be too full of people: *The town was crawling with police.* **2** to be full of unpleasant insects or animals

crawl

crawl[2] /krɔːl/ noun [singular] **1** a very slow speed **2** a style of swimming in which you move one arm over your head and then the other while you are kicking your legs

crayon /ˈkreɪɒn/ noun [C] a stick of coloured WAX that is used for drawing

craze /kreɪz/ noun [C] something that suddenly becomes very popular for a short time

crazed /kreɪzd/ adj completely crazy and

crease[2] /kriːs/ verb [I/T] to make lines on cloth or paper by folding or crushing it, or to become covered in these lines

crazy /ˈkreɪzi/ adj *informal* ★★ not at all sensible or practical: *It's crazy. Who would do a thing like that?* ♦ **be crazy to do sth** *She knew she would be completely crazy to refuse.*

PHRASES **crazy about sb** very much in love with someone

crazy about sth very enthusiastic about something

drive sb crazy to make someone very annoyed

go crazy 1 to become very angry about something **2** to become very bored and upset **3** to become very excited

like crazy to a very great degree = MAD: *The games are selling like crazy.*

—crazily adv

C

Words that may cause offence: crazy

Avoid using words like **crazy**, **mad**, and **insane** about people who have mental illnesses or mental health problems. Instead, use an expression such as **mentally ill**.

Jak jsem ke slovníkům přišel/přešel

- brigáda “na léto”
- řecký korpus
- ohodnocování slovníkových příkladů
- Python, Manatee, SkE, WebBootCat ....
- bakalářská práce

Jak jsem ke slovníkům přišel/přešel

- brigáda “na léto”
- řecký korpus
- ohodnocování slovníkových příkladů
- Python, Manatee, SkE, WebBootCat ....
- bakalářská práce

Jak jsem ke slovníkům přišel/přešel

- brigáda “na léto”
- řecký korpus
- ohodnocování slovníkových příkladů
- Python, Manatee, SkE, WebBootCat ....
- bakalářská práce

Jak jsem ke slovníkům přišel/přešel

- brigáda “na léto”
- řecký korpus
- ohodnocování slovníkových příkladů
- Python, Manatee, SkE, WebBootCat ....
- bakalářská práce

Jak jsem ke slovníkům přišel/přešel

- brigáda “na léto”
- řecký korpus
- ohodnocování slovníkových příkladů
- Python, Manatee, SkE, WebBootCat ....
- bakalářská práce

Cíl práce

Zautomatizovat (alespoň částečně) vyhledávání slovníkových příkladů v korpusu.

- zjistit, jak se dobré příklady dají rozpoznat
- zkombinovat různé přístupy
- vytvořit snadno použitelné rozhraní

Cíl práce

Zautomatizovat (alespoň částečně) vyhledávání slovníkových příkladů v korpusu.

- zjistit, jak se dobré příklady dají rozpoznat
- zkombinovat různé přístupy
- vytvořit snadno použitelné rozhraní

Cíl práce

Zautomatizovat (alespoň částečně) vyhledávání slovníkových příkladů v korpusu.

- zjistit, jak se dobré příklady dají rozpoznat
- zkombinovat různé přístupy
- vytvořit snadno použitelné rozhraní

Cíl práce

Zautomatizovat (alespoň částečně) vyhledávání slovníkových příkladů v korpusu.

- zjistit, jak se dobré příklady dají rozpoznat
- zkombinovat různé přístupy
- vytvořit snadno použitelné rozhraní

Dobrý příklad je ...

- informativní, samopopisný
  - pozice slova ve větě
  - synonyma
  - redundance
- snadno čitelný
  - jednoduchá větná struktura
  - krátká slova
  - běžná jména (politické problémy)
  - častá slovní spojení

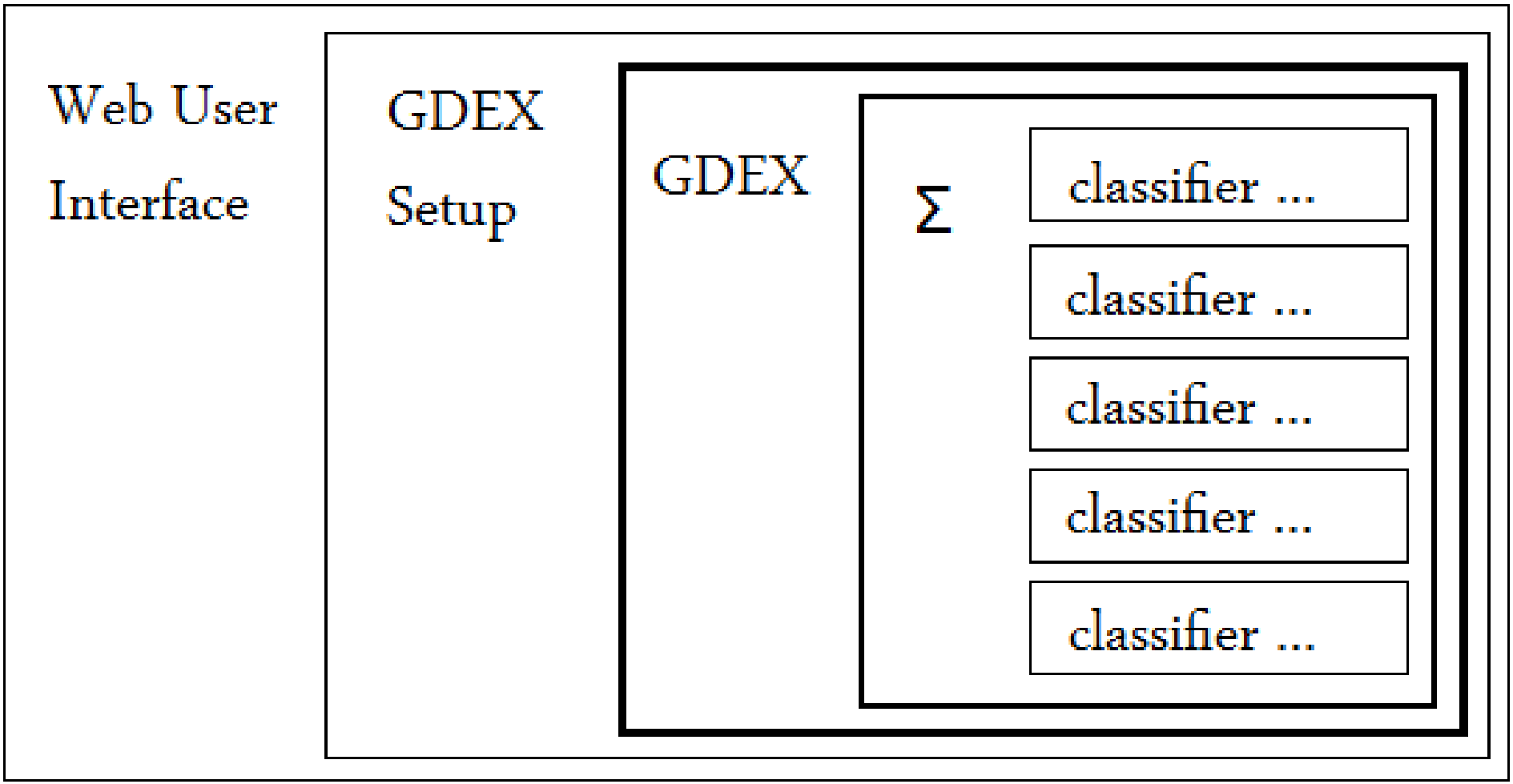
Dobrý příklad je ...

- informativní, samopopisný
  - pozice slova ve větě
  - synonyma
  - redundance
- snadno čitelný
  - jednoduchá větná struktura
  - krátká slova
  - běžná jména (politické problémy)
  - častá slovní spojení

Dobrý příklad je ...

- informativní, samopopisný
  - pozice slova ve větě
  - synonyma
  - redundance
- snadno čitelný
  - jednoduchá větná struktura
  - krátká slova
  - běžná jména (politické problémy)
  - častá slovní spojení

GDEX


Web User
Interface
GDEX
Setup
GDEX
Σ
classifier ...
classifier ...
classifier ...
classifier ...
classifier ...

GDEX

GDEX je srdcem a tělem ohodnocování vět.

- hostí hlavní klasifikátor
- poskytuje jednotné rozhraní všem klasifikátorům
- spravuje dostupné prostředky

GDEX

GDEX je srdcem a tělem ohodnocování vět.

- hostí hlavní klasifikátor
- poskytuje jednotné rozhraní všem klasifikátorům
- spravuje dostupné prostředky

GDEX

GDEX je srdcem a tělem ohodnocování vět.

- hostí hlavní klasifikátor
- poskytuje jednotné rozhraní všem klasifikátorům
- spravuje dostupné prostředky

GDEX

GDEX je srdcem a tělem ohodnocování vět.

- hostí hlavní klasifikátor
- poskytuje jednotné rozhraní všem klasifikátorům
- spravuje dostupné prostředky

Klasifikátory

Klasifikátory jsou mozkem ohodnocování vět.

- každou větu ohodnotí číslem z $< 0, 1 >$
- silné a slabé klasifikátory
- délka vět, délka slov, četnost slov, počty významů, ...
- vážený průměr, progresivní skóre, přibližné progresivní skóre

Klasifikátory

Klasifikátory jsou mozkem ohodnocování vět.

- každou větu ohodnotí číslem z $< 0, 1 >$
- silné a slabé klasifikátory
- délka vět, délka slov, četnost slov, počty významů, ...
- vážený průměr, progresivní skóre, přibližné progresivní skóre

Klasifikátory

Klasifikátory jsou mozkem ohodnocování vět.

- každou větu ohodnotí číslem z $< 0, 1 >$
- silné a slabé klasifikátory
- délka vět, délka slov, četnost slov, počty významů, ...
- vážený průměr, progresivní skóre, přibližné progresivní skóre

Klasifikátory

Klasifikátory jsou mozkem ohodnocování vět.

- každou větu ohodnotí číslem z $< 0, 1 >$
- silné a slabé klasifikátory
- délka vět, délka slov, četnost slov, počty významů, ...
- vážený průměr, progresivní skóre, přibližné progresivní skóre

Klasifikátory

Klasifikátory jsou mozkem ohodnocování vět.

- každou větu ohodnotí číslem z $< 0, 1 >$
- silné a slabé klasifikátory
- délka vět, délka slov, četnost slov, počty významů, ...
- vážený průměr, progresivní skóre, přibližné progresivní skóre

Měření efektivity

$$benchmark = 1 - \frac{\sum_{i \in concordances} \frac{\sum_{j \in goodexamples_i} n_{i,j}}{n_{last\_good\_example_i}+1}}{\sum_{i \in concordances} |goodexamples_i|}$$

Měření efektivity

- průměrná pozice dobrých slovníkových příkladů
- ručně značkované konkordance
- optimalizace parametrů klasifikátorů
- vyvažování důležitostí klasifikátorů

Měření efektivity

- průměrná pozice dobrých slovníkových příkladů
- ručně značkované konkordance
- optimalizace parametrů klasifikátorů
- vyvažování důležitostí klasifikátorů

Měření efektivity

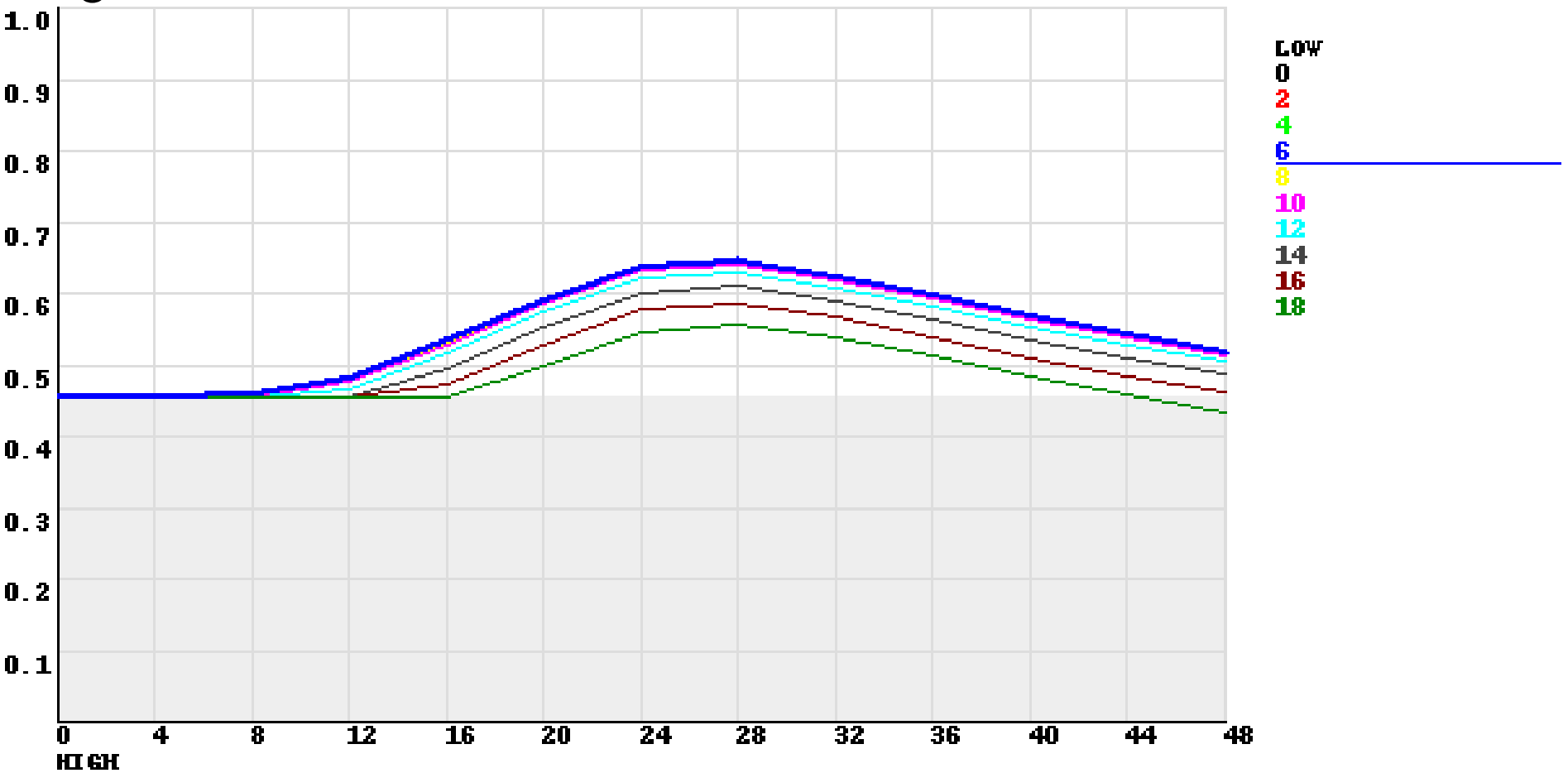
- průměrná pozice dobrých slovníkových příkladů
- ručně značkované konkordance
- optimalizace parametrů klasifikátorů
- vyvažování důležitostí klasifikátorů

Měření efektivity

- průměrná pozice dobrých slovníkových příkladů
- ručně značkované konkordance
- optimalizace parametrů klasifikátorů
- vyvažování důležitostí klasifikátorů

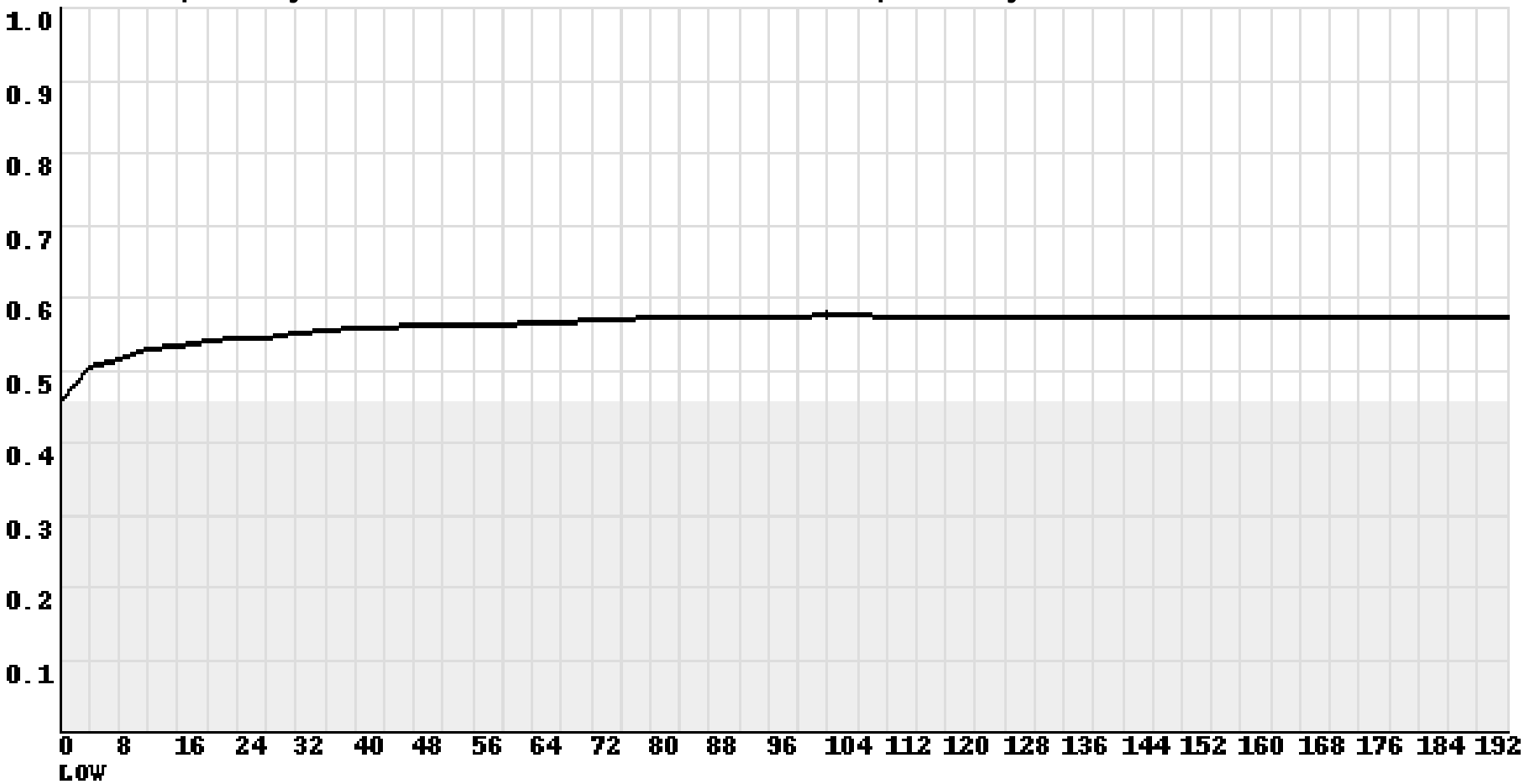
Dobrý klasifikátor I.

Length score - délka vět



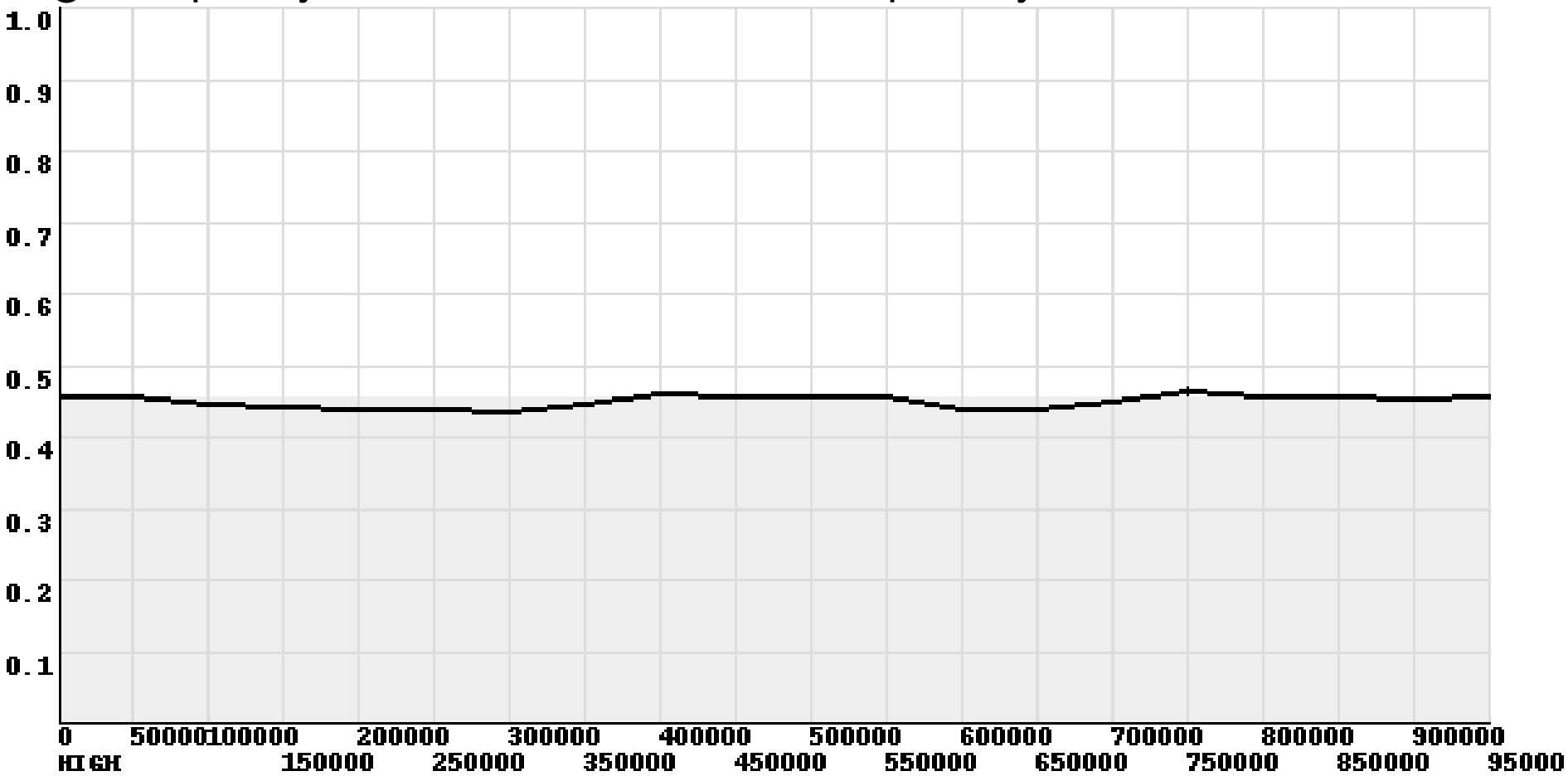
Dobrý klasifikátor II.

Low frequency score - minimální četnost použitých slov



Špatný klasifikátor

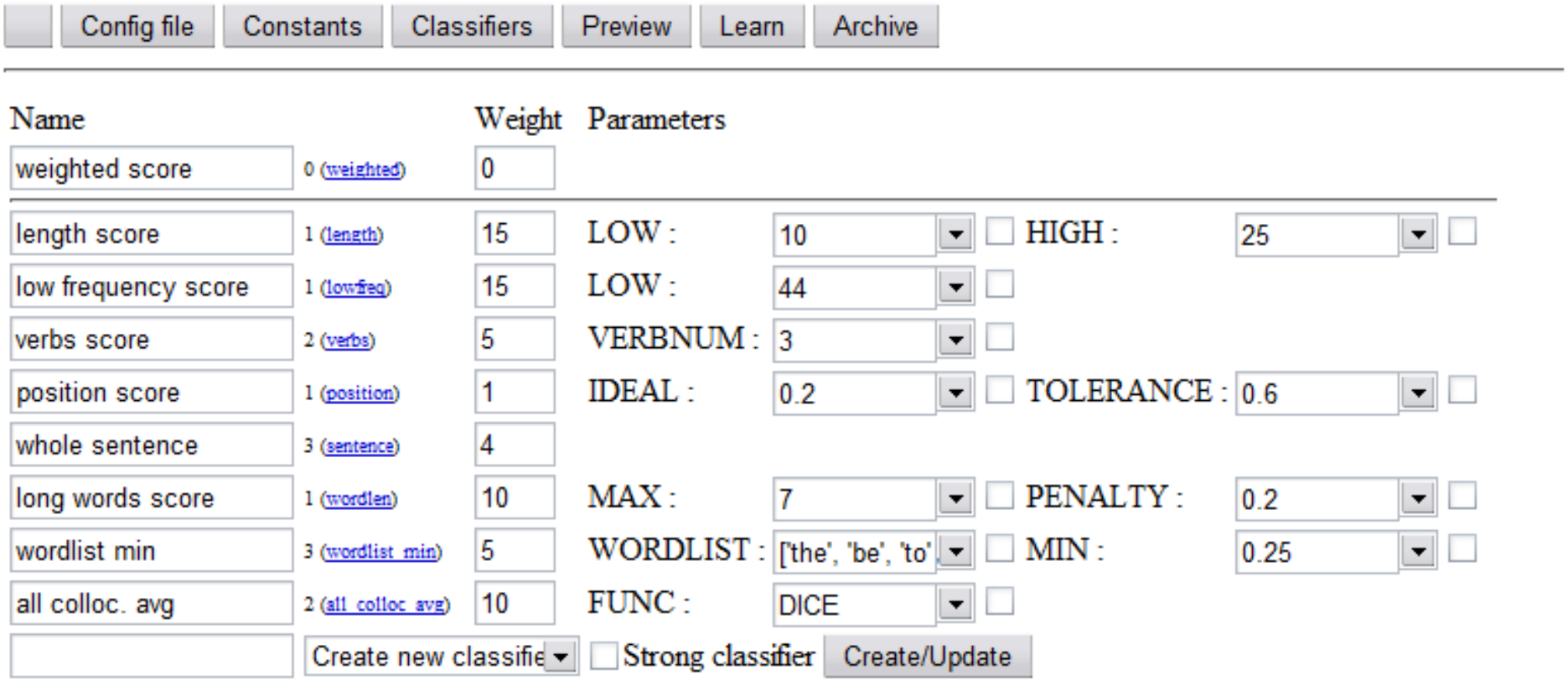
High frequency score - maximální četnost použitých slov



Ukázky

- GDEX Setup
- Sketch Engine
- GDEX Demo Dictionary

GDEX Setup


Config file
Constants
Classifiers
Preview
Learn
Archive
Name
Weight
Parameters
weighted score
0 (weighted)
0
length score
1 (length)
15
LOW :
10
HIGH :
25
low frequency score
1 (lowfreq)
15
LOW :
44
verbs score
2 (verbs)
5
VERBNUM :
3
position score
1 (position)
1
IDEAL :
0.2
TOLERANCE :
0.6
whole sentence
3 (sentence)
4
long words score
1 (wordlen)
10
MAX :
7
PENALTY :
0.2
wordlist min
3 (wordlist_min)
5
WORDLIST :
['the', 'be', 'to'
MIN :
0.25
all colloc. avg
2 (all_colloc_avg)
10
FUNC :
DICE
Create new classifie
Strong classifier
Create/Update

SketchEngine

http://beta.sketchengine.co.uk/

Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff | View

Corpus: British National Corpus
conc description

View options

Attributes

- [x] word
- [ ] ambtag
- [ ] lempos
- [ ] tag
- [ ] lemma
- [ ] lc

Display attributes
- ( ) For each token
- (•) KWIC tokens only

Structures

<bncdoc>
<stext>
<text>
<s>
<p>
<align>
<caption>
<hi>
<lb>
<bibl>
<body>

<div1>

References

Token number
bncdoc.id
bncdoc.date
bncdoc.author
bncdoc.title
bncdoc.info
Text availability
Text type
Publication date
David Lee's classification
Domain for written corpus texts
Domain for context-governed spoken material
Medium for written corpus texts

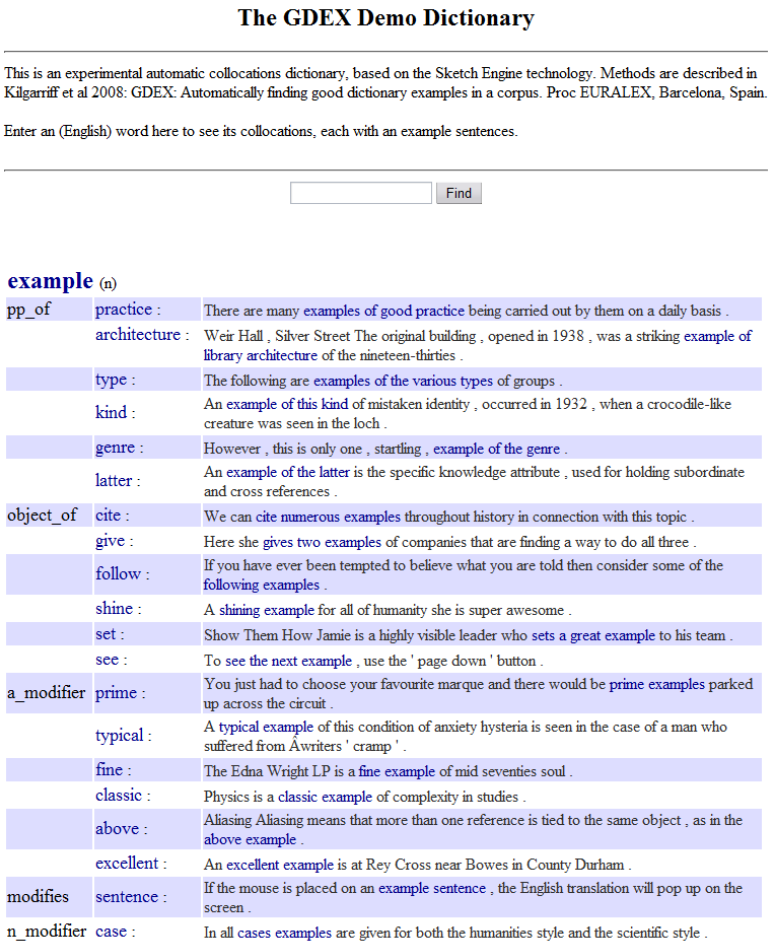
Page size (number of lines): 20
KWIC Context size (number of characters): 40

- [x] Icon for one-click sentence copying
- [ ] Sort concordance lines best-first. Number of lines to be sorted: 100

Change View Options

GDEX Demo Dictionary

http://forbetterenglish.sketchengine.co.uk/



The GDEX Demo Dictionary

This is an experimental automatic collocations dictionary, based on the Sketch Engine technology. Methods are described in Kilgarriff et al 2008: GDEX: Automatically finding good dictionary examples in a corpus. Proc EURALEX, Barcelona, Spain.

Enter an (English) word here to see its collocations, each with an example sentences.

Find

example (n)

pp_of	practice :	There are many examples of good practice being carried out by them on a daily basis .
	architecture :	Weir Hall , Silver Street The original building , opened in 1938 , was a striking example of library architecture of the nineteen-thirties .
	type :	The following are examples of the various types of groups .
	kind :	An example of this kind of mistaken identity , occurred in 1932 , when a crocodile-like creature was seen in the loch .
	genre :	However , this is only one , startling , example of the genre .
	latter :	An example of the latter is the specific knowledge attribute , used for holding subordinate and cross references .
object_of	cite :	We can cite numerous examples throughout history in connection with this topic .
	give :	Here she gives two examples of companies that are finding a way to do all three .
	follow :	If you have ever been tempted to believe what you are told then consider some of the following examples .
	shine :	A shining example for all of humanity she is super awesome .
	set :	Show Them How Jamie is a highly visible leader who sets a great example to his team .
	see :	To see the next example , use the ' page down ' button .
a_modifier	prime :	You just had to choose your favourite marque and there would be prime examples parked up across the circuit .
	typical :	A typical example of this condition of anxiety hysteria is seen in the case of a man who suffered from Áwriters ' cramp ' .
	fine :	The Edna Wright LP is a fine example of mid seventies soul .
	classic :	Physics is a classic example of complexity in studies .
	above :	Aliasing Aliasing means that more than one reference is tied to the same object , as in the above example .
	excellent :	An excellent example is at Rey Cross near Bowes in County Durham .
modifies	sentence :	If the mouse is placed on an example sentence , the English translation will pop up on the screen .
n_modifier	case :	In all cases examples are given for both the humanities style and the scientific style .

Budoucí práce

- implementovat další klasifikátory
- naučit GDEX i na jiných datech/jazycích
- reimplementovat GDEX v C/C++

Budoucí práce

- implementovat další klasifikátory
- naučit GDEX i na jiných datech/jazycích
- reimplementovat GDEX v C/C++

Budoucí práce

- implementovat další klasifikátory
- naučit GDEX i na jiných datech/jazycích
- reimplementovat GDEX v C/C++