Jak se řeší tvarosloví se zaměřením na češtinu

Richard Salač

DATERA s.r.o.

17. Ledna 2019

EMMA
powered by DATERA

Fulltextové vyhledávání

Musím vědět, co hledám, jak to pojmenovat a jak to někdo popsal

Fulltextové vyhledávání

EMMA
powered by DATERA

- Fulltextový vyhledávač najde to, co v textu je
- Jak najít něco, co tam není?
  - Skloňování
  - Časování
  - Synonyma, hyperonyma, hyponyma, antonyma

Podpora jazyka

EMMA
powered by DATERA

- Skloňování
- Časování
- Práce s diakritikou i bez
- Co největší množství podporovaných jazyků

EMMA
powered by DATERA

Živá ukázka

Systém, který všechno splňuje

Vál

EMMA
powered by DATERA

- ová**lný**
- vál**cem**
- val**ník**
- **s**val**ový**

- val**era**
- **festi**val
- val**ley**

- vál**ení**
- vál**eček**

EMMA
powered by DATERA

Stemming

Nejhorší avšak nejpoužívanější způsob podpory tvarosloví

Stemming

EMMA
powered by DATERA

- Nejjednodušší, nejrozšířenější a **nejvíce chybový** přístup
- Slova redukuje na **stem - kořen**
- Seznam předpon a přípon z jazyka nebo více jazyků

Stemming – příklad

EMMA
powered by DATERA

- Un**iverse**
- Un**iversity**
- Un**iversal**
- Un**iversality**

- **Vál**
- O**vál**
- O**válný**
- S**val**ový

EMMA
powered by DATERA

Analýza a stemming

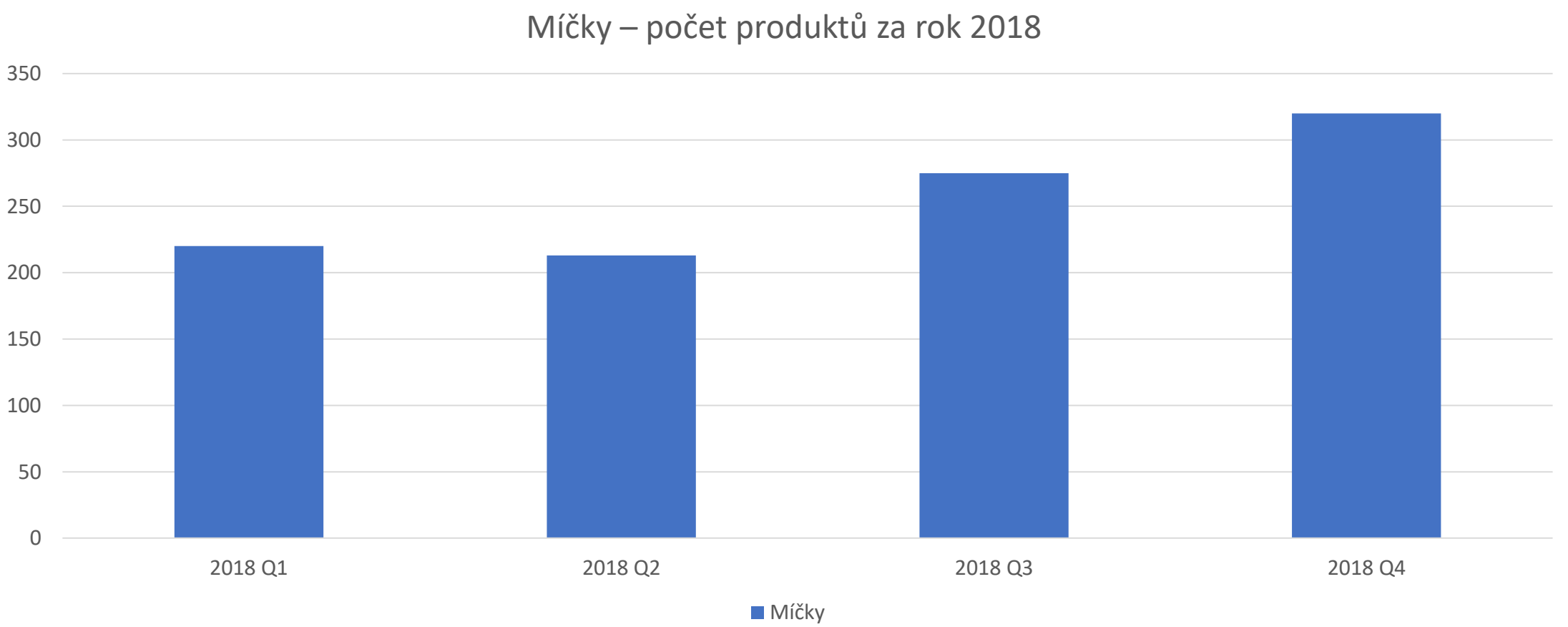
Živá ukázka

Průzkum trhu - míčky

EMMA
powered by DATERA

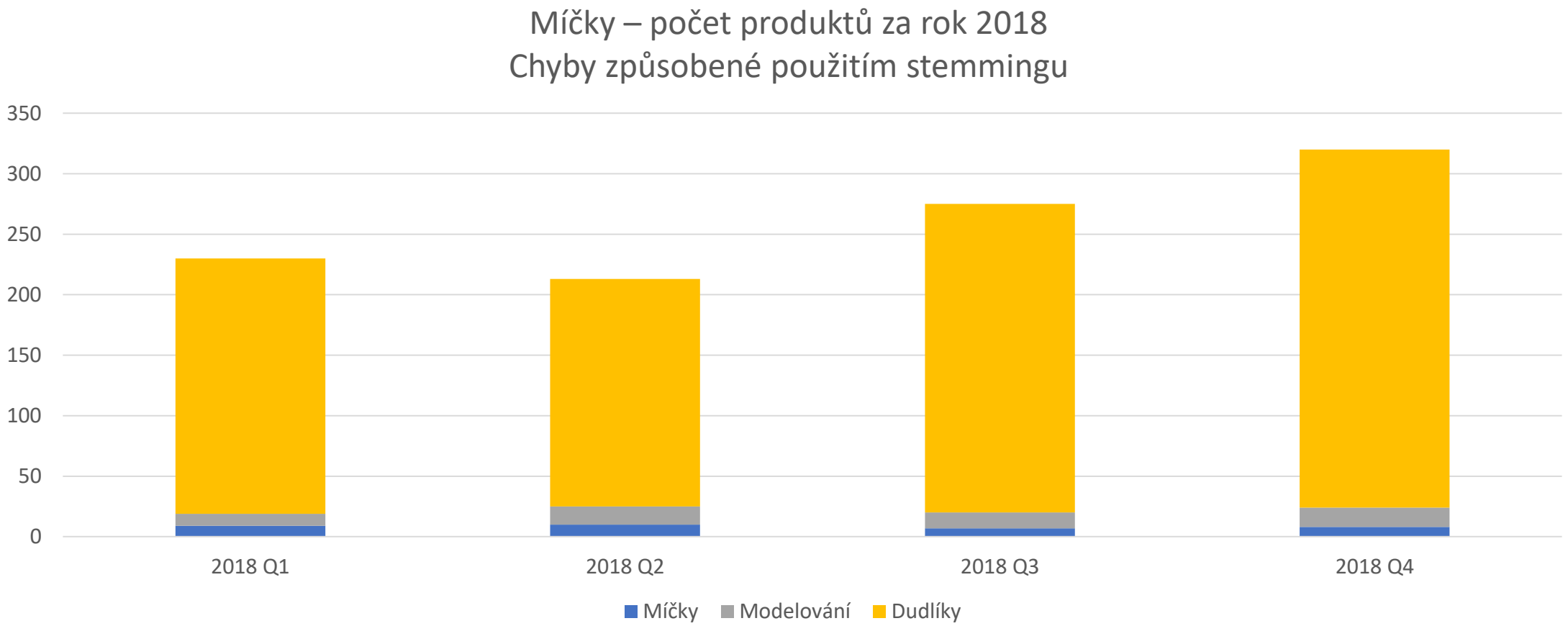
- Vyrábíme míčky a sledujeme trh
- Zajímá nás, jaká je naše konkurence
- Míčky – celkem 213 výsledků (cca 10 relevantních)

Průzkum trhu – míčky



Průzkum trhu – míčky

EMMA
powered by DATERA



Stemming

EMMA
powered by DATERA

- Funguje na bázi seznamu předpon a přípon
- Velká chybovost
- Spojuje slova, která k sobě nepatří
- Rozděluje slova, která k sobě patří
- Neřeší nepravidelné tvary (stem, kořen, se mění)

EMMA
powered by DATERA

Slovníkový stemming

Slovníkový stemming

EMMA
powered by DATERA

- Databáze pravidel pro aplikaci předpon a přípon
- Obsahuje slovník, ke každému slovu jsou přiřazena pravidla
- ~~Spojuje slova, která k sobě nepatří~~
- Rozděluje slova, která k sobě patří
- Neřeší nepravidelné tvary
- Pro češtinu nepoužitelný, pro některé jazyky omezeně, např. angličtinu

Slovníkový stemming

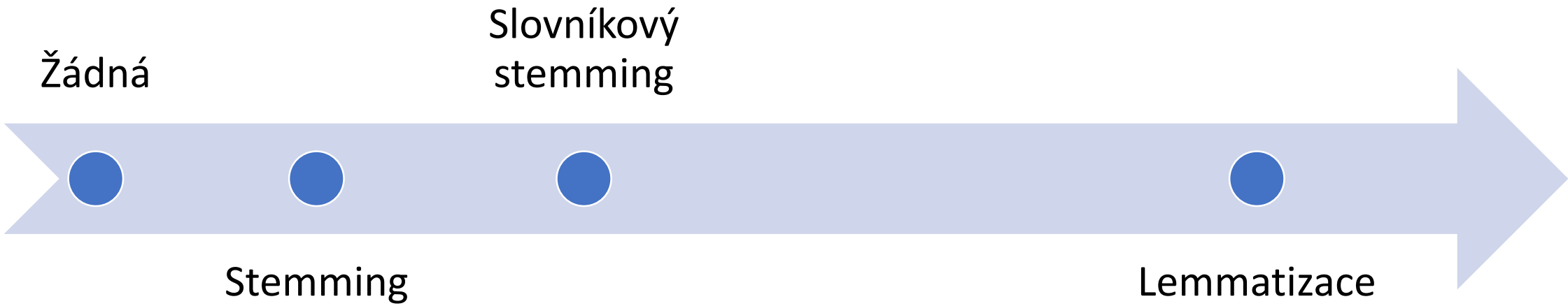
EMMA
powered by DATERA

- Být -> býti, nebýt, nebýti
  - Byl, byli, byla, jsem, jsi, je, jsme, jste, jsou,

    budu, budeš, bude, budeme, budete, budou, bych, bys, bychom, byste, by + negativní tvary +...
- Podepsat -> podepsati, nepodepsat, nepodepsati
- Dítě -> dítěte, dítěti, dítětem
  - Děti, dětí, dětem, dětech, dětmi
- Osobní zájmena
- Kmen slova se mění, význam nikoliv

Kvalita podpory tvarosloví

EMMA
powered by DATERA



Lemmatizace

EMMA
powered by DATERA

- **Nejpřesnější** způsob práce se slovy
- Podporuje skloňování i časování vč. nepravidelných tvarů
- Slova redukuje na **lemma** – skutečná základní forma slova

Lemmatizace - ukázka

EMMA
powered by DATERA

☐	1	Michálek: Post ministra bych přijal, pokud by obsazování pozic náměstků ne...	2014-01-09 21:42

... V médiích se dnes objevila informace, že mi pan Babiš marně nabízel post ministra životního ... Tak tomu není. Při prvním oslovení jsem uvedl, že bych tento post přijal, pokud by obsazování pozic ... Před Vánoci jsem potom panu Babišovi poslal seznam lidí, kteří by mohli tvořit můj tým, pokud bych ... byl ministrem, ale odpověď jsem již neobdržel. Babiš, Michálek, MŽP ...

☐	2	Triumf před domácím kotlem? Jako když jsem byl poprvé na Matějské, rozplýv...	2017-05-28 07:01

... říkal, že mě to nenervuje, tak to bylo docela těžký. Dneska jsem nervózní byl, lidí tu byla spousta ... Tušil jste před příjezdem k tribunám, že máte vítězství v kapse a mohl si tak triumf o to víc užít ... jsem si to užil. Z evropských šampionátů to bylo určitě nejlepší a i když mistrovství světa to není ... Je pravda, že loni jsem nevyhrál žádný závod, byl jsem vždycky druhý nebo třetí, letos první závod ...

☐	3	ILL BILL – How to Survive the Apocalypse	2014-01-09 11:05

... někteří by plakali a modlili se k bohu Svět je nebezpečnej, je na nestabilné ose Většina z nás není ... kdyby se společnost zhroutila ...

☐	4	MessenJah – Neni Mi Jedno	2012-12-04 11:13

... , Není mi jedno, že školy budou pro bohatý, Není mi jedno, že co se třpytí není zlatý, Není mi jedno ...

☐	5	Dojatá Špotáková: Tak silný pocit štěstí jsem dlouho nezažila!	2017-08-08 06:52

Diskuze

EMMA
powered by DATERA

Děkujeme za pozornost