SentiSquare

Tvoříme chatbota na základě předchozích interakcí

Josef Steinberger a Tomáš Brychcín

SentiSquare

Výzkum zpracování přirozeného jazyka (NLP)

- ➔ Západočeská univerzita v Plzni
- ➔ Spin-off SentiSquare (2014)

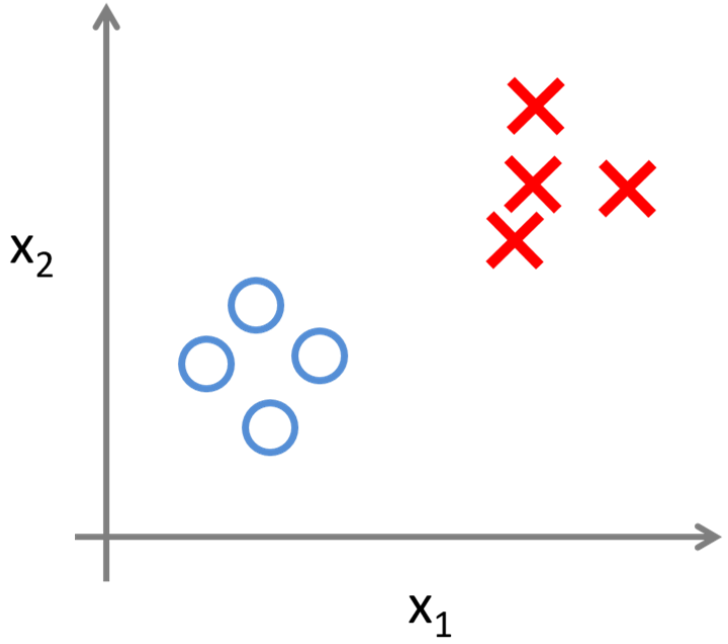
Zaměření

- ➔ Analýza textů: online zdroje, emaily, poznámky v CRM, přepisy hovorů, chaty
- ➔ Machine learning - supervised vs. unsupervised
- ➔ Nezávislost metod na jazyku textů

Automatické přeposílání emailů

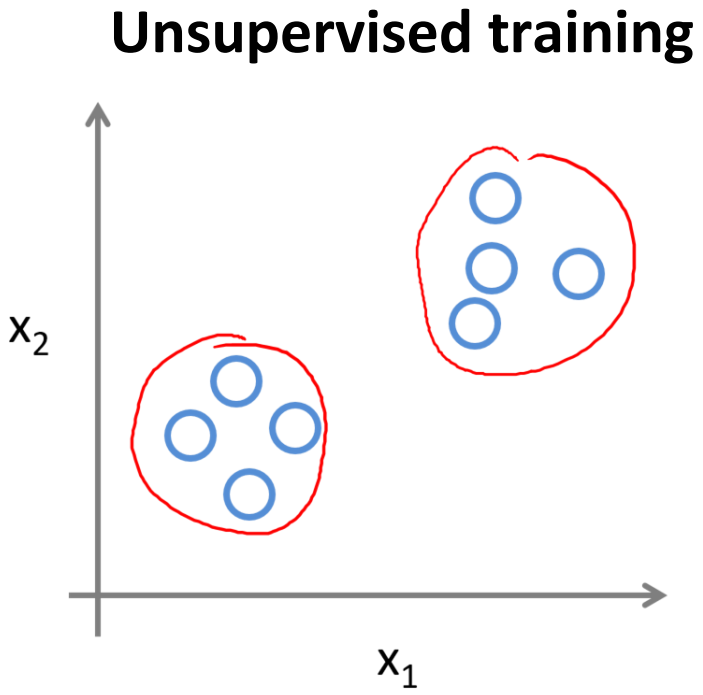
- ➔ Klasifikace emailů do cca 300 tříd
- ➔ Machine learning vs. klasifikace klíčovými slovy
- ➔ Ušetření 2 FTE /rok
- ➔ Obecně: klasifikace interních dokumentů

Supervised training



SMS feedback

- ➔ Analýza >10,000 sms / měsíc
- ➔ Automatická detekce témat
- ➔ Nalezení nespokojených zákazníků a důvodů nespokojenosti
- ➔ Sémantické vyhledávání a filtrování v datech



Cesta k chatbotovi

Krok 1: Analýza chatů – unsupervised

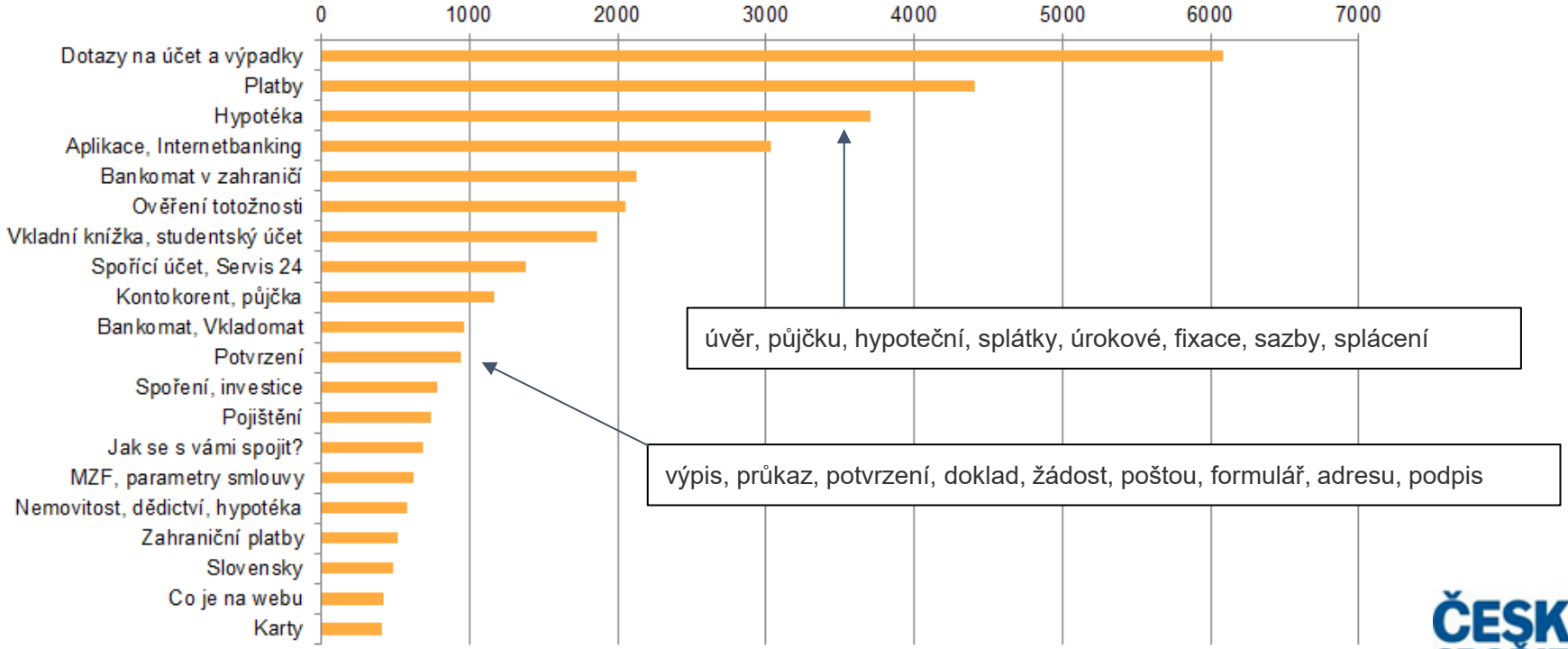
- ➔ Analytika – témata, aspekty, vhledy do dat
- ➔ Směrování na vhodného operátora
- ➔ Kategorie do pro učení chatbota

Krok 2: Chat asistent (otázka - odpověď)

Krok 3: Dialog, scénáře, omezená doména

Krok 4: Odvozování faktů, komplexní chatbot

Krok 1: Analýza chatů



ČESKÁ SPOŘITELNA

Krok 2: Chat asistent

Předpoklady:

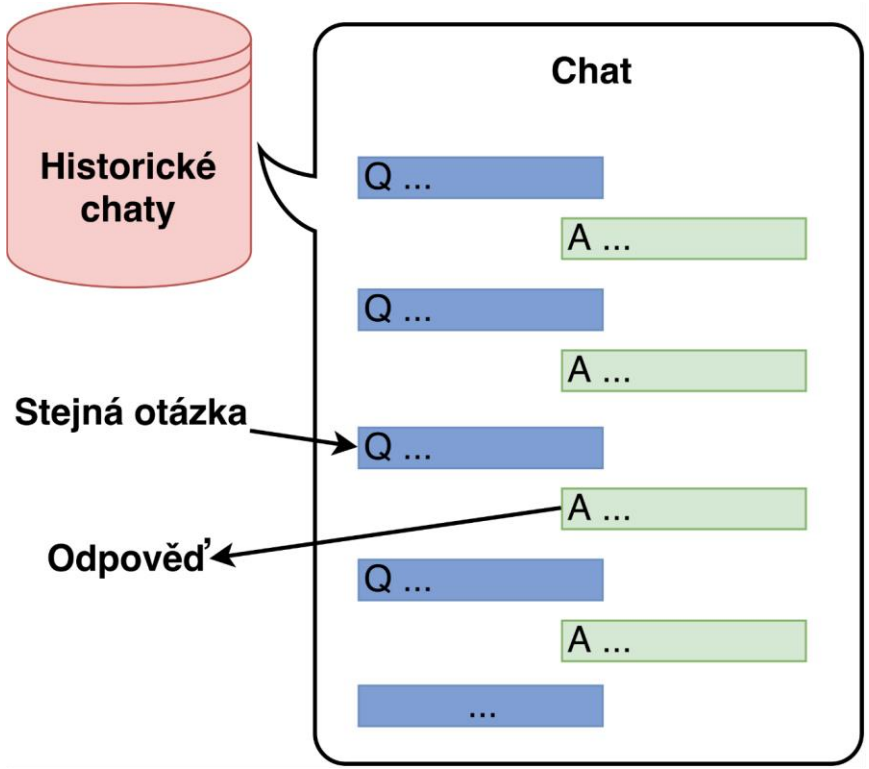
➔ Mít historická data (rozumné velikosti)

Řešení:

➔ Najít stejnou (podobnou) otázku (požadavek) v historii
➔ Vrátit odpověď co následovala

Problémy:

➔ Jak porovnávat texty?
➔ Jak poznat, že odpověď byla správná?



Proč naivní metody nefungují?

- ➔ Význam lze vyjádřit různými slovy.
- ➔ I při použití morfologické analýzy (lemmatizace, stemming) velikost běžného slovníku > 100 000
- ➔ Překlepy, zkratky, slangové výrazy
- ➔ Počet kombinací roste exponenciálně s délkou textu.

Chtěl bych si půjčit peníze.
Chci si pučit peníze.
Pučíte mi peníze?
Rád bych si pučil.
Je možná půjčka?
Je možné si pučit?

Chtěl bych hypotéku.
Potreboval bych vyridit hypoteku.
Mám zájem o úvěr.
Uvažuji o hypotecním úveru.
Mohu si u vás vzít hypotéku?
Mohl bych si vzít americkou hypoteku (uver)?

SentiSquare řešení – část 1

Distribuční sémantika

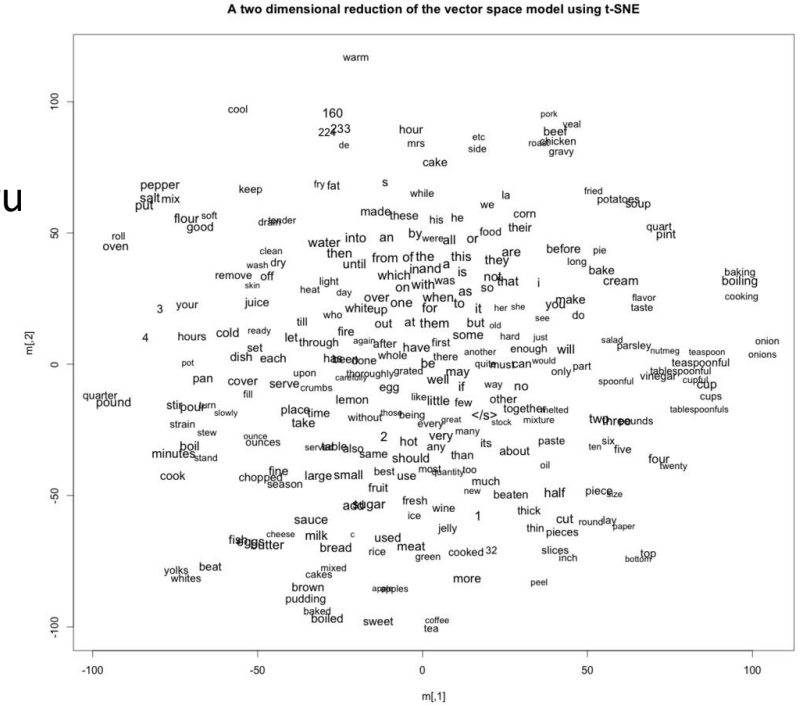
- ➔ Slova v podobném kontextu mají podobný význam
- ➔ Reprezentace významu v sémantickém prostoru s vysokou dimenzí
- ➔ LSA, LDA, word2vec, doc2vec, a další

Výhody:

- ➔ unsupervised/weakly supervised
- ➔ Jazykově a doménově nezávislé

Nevýhody:

- ➔ Nedokáže rozlišit drobné niance ve významu



Jaké dokumenty a potvrzení musím dodat, pokud si chci sjednat hypotéku na dům?
Jaké dokumenty a potvrzení musím dodat, pokud si chci sjednat pojištění na dům?

SentiSquare řešení – část 2

Přístupy:

- ➔ Distribuční sémantika: orientuje na význam / pouze abstraktní
- ➔ Slova: příliš mnoho kombinací / jsou konkrétní

Řešení:

- ➔ Extrakce podstatných informací
  - Významová slova
  - Pojmenované entity (čísla, produkty, datum, atd.)

- ➔ Kombinace obou přístupů!!

SentiSquare řešení – část 3

Další problémy:

1) Produkty se vyvíjí –> odpovědi na otázky také
2) Jak poznat že odpověď je OK?
3) Co dělat, když podobný požadavek v datech není?

Jak na to…

1) Vzít nejnovější odpověď
2) Pokud je uživatel spokojený, pravděpodobně se už nedoptává (poděkuje).
3) Semantic reasoning – odvozování faktů – vyšší verze chatbota ☺

SemEval 2016 shared task

Semantic Evaluation (SemEval) 2016 shared task na sémantickou podobnost textů:

➔ Vstupem dvě věty - výstup jejich sémantická podobnost

Výsledky:

➔ Monolinguální (angličtina) - **#2** ze 113 algoritmů

How do I pump up water pressure in my shower?
How can I fix low water pressure in one shower? **80%**

➔ Bilinguální (angličtina vs. španělština) - **#1** z 26 algoritmů

How do I fix a hole/gap between my shower tile and the dry wall next to it?
¿Cómo reparo un heuco entre la bañera y la pared? **92%**

Krok 3 a 4: Chatbot

O tom až příště ☺