# Hledání sémantické informace v textových datech s využitím latentní analýzy

Bc. Pavel Řezníček

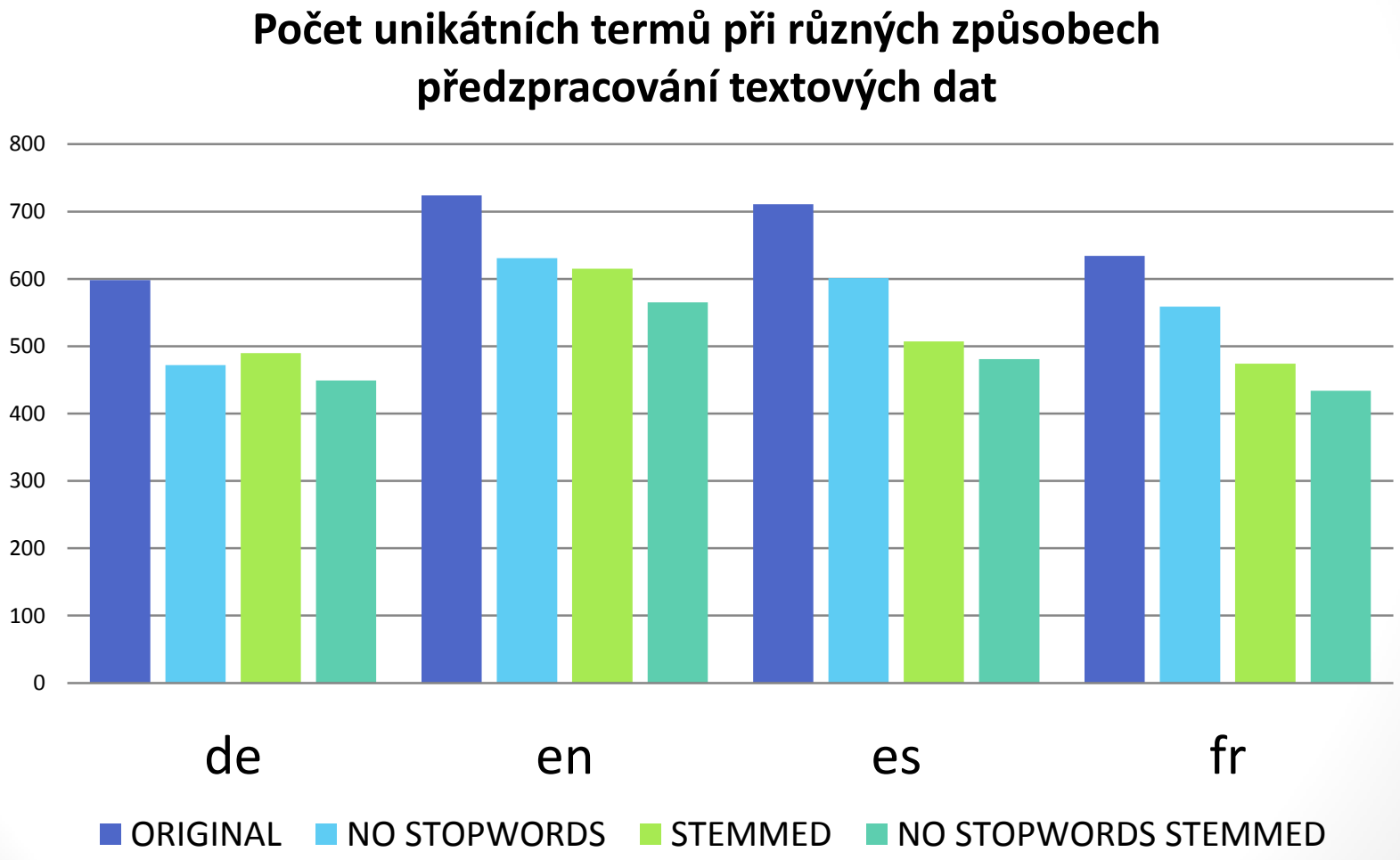
# Cíl práce

- analyzovat dostupnou sadu nestrukturovaných textových dat
- vhodnými metodami text miningu z dat vydolovat určité znalosti
- sledovat a porovnat vliv různého předzpracování dat
- měřit rozdíly ve výsledcích metod při využití klasického vektorového modelu a latentní analýzy

# Metodika

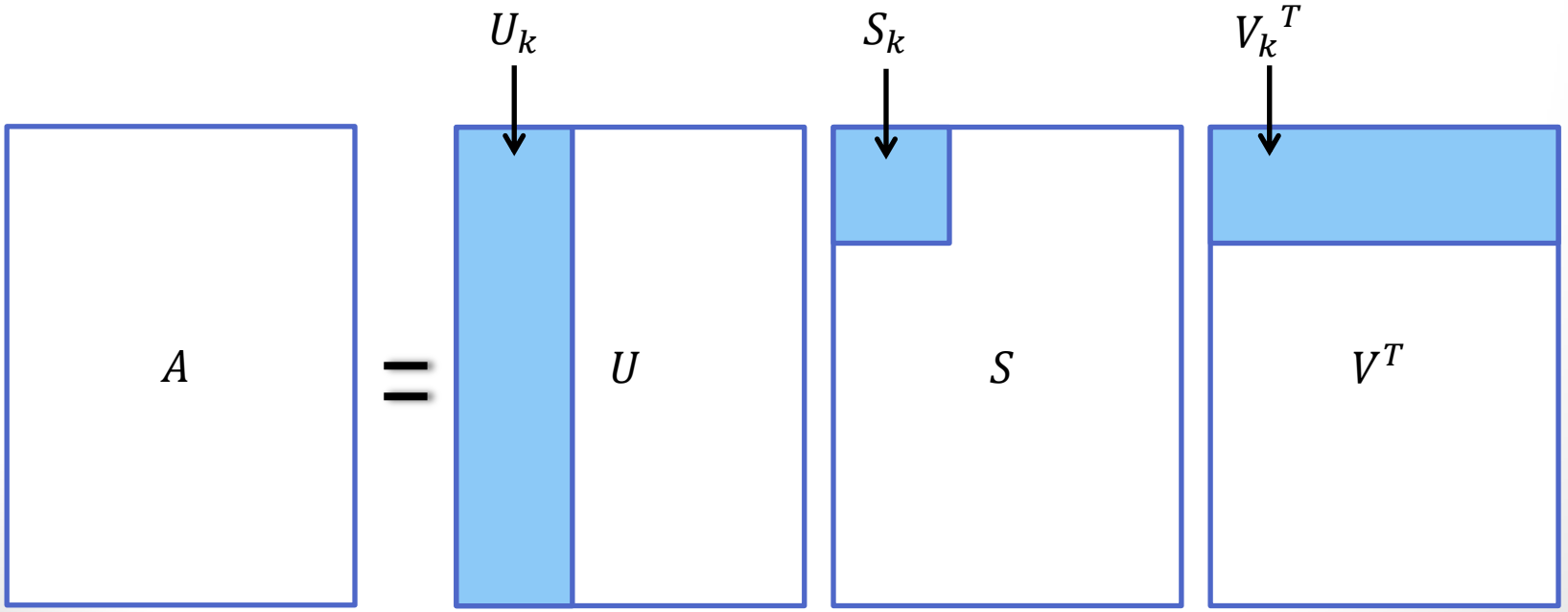
- zdrojová data – recenze návštěvníků hotelů v několika jazycích
- předzpracování – case-folding, odstranění stop-slov, stemming
- převod do numerické reprezentace
  - Vektorový model
  - LSI databáze

# Výsledky předzpracování



# Výsledky redukce dimenzí

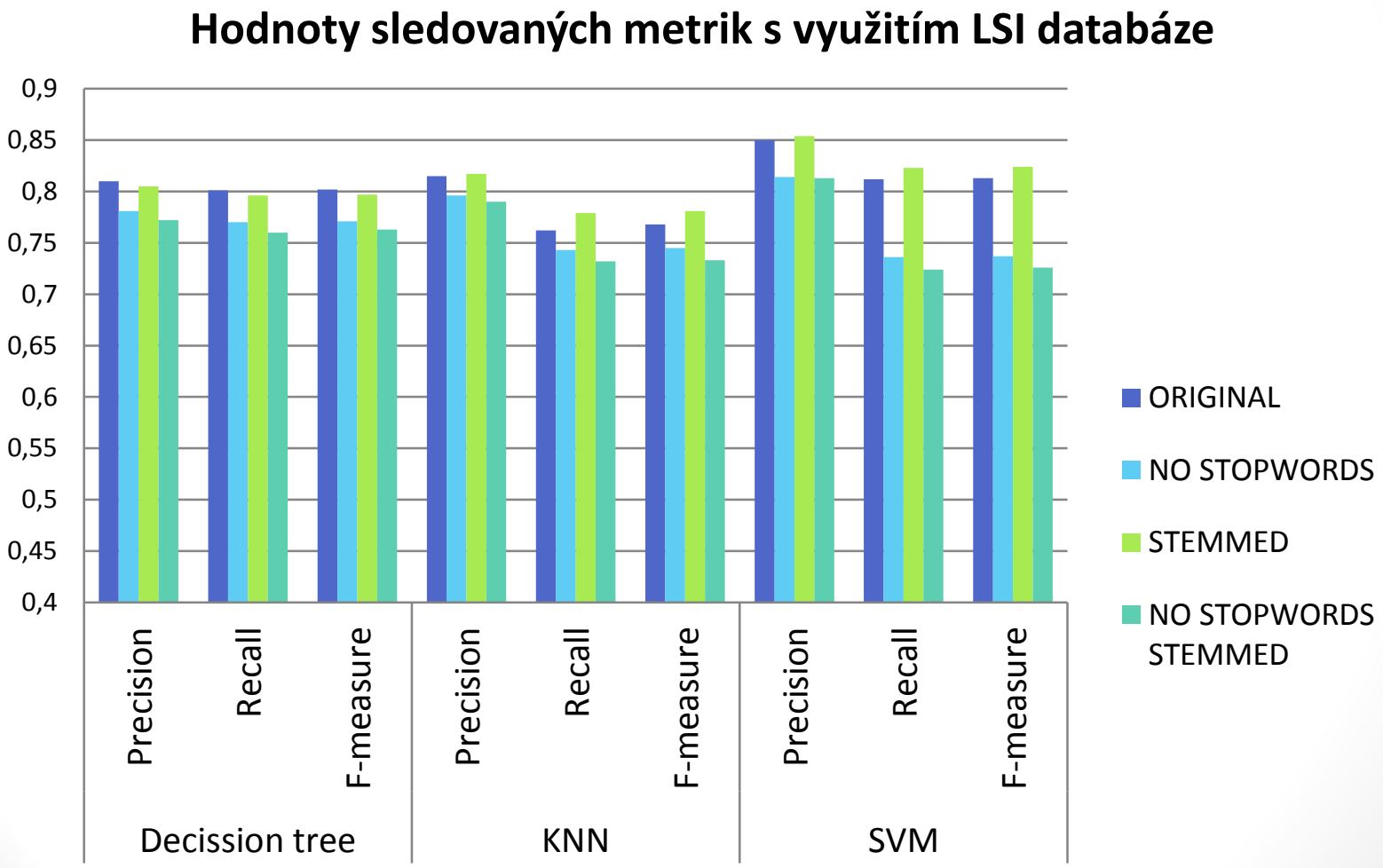
- zmenšení prostoru pro provádění výpočtů metodou SVD
- odstranění šumu
- odhalení skrytých vazeb mezi termy a dokumenty



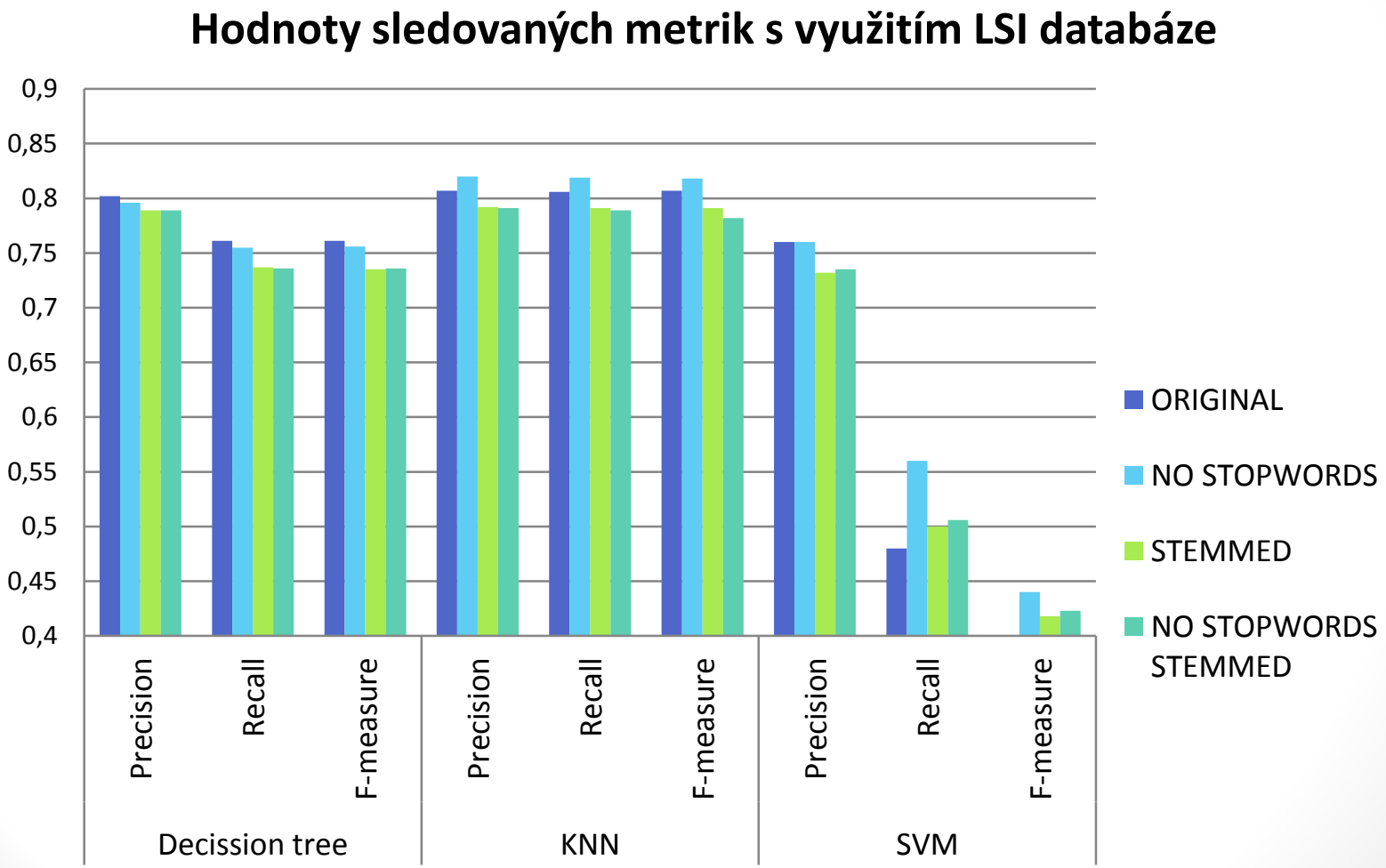
# Klasifikace

- klasifikace do dvou tříd – pozitivní (P), negativní (N)
- metriky – precision, recall, F-measure
- klasifikátory – Decission tree, KNN, SVM
- výsledky ideální metody předzpracování
  - německý jazyk – stemming
  - anglický jazyk – originál
  - španělský jazyk – odstranění stop-slov
  - francouzský jazyk – odstranění stop-slov
- pozitivní vliv využití LSI databáze

# Výsledky klasifikace DE

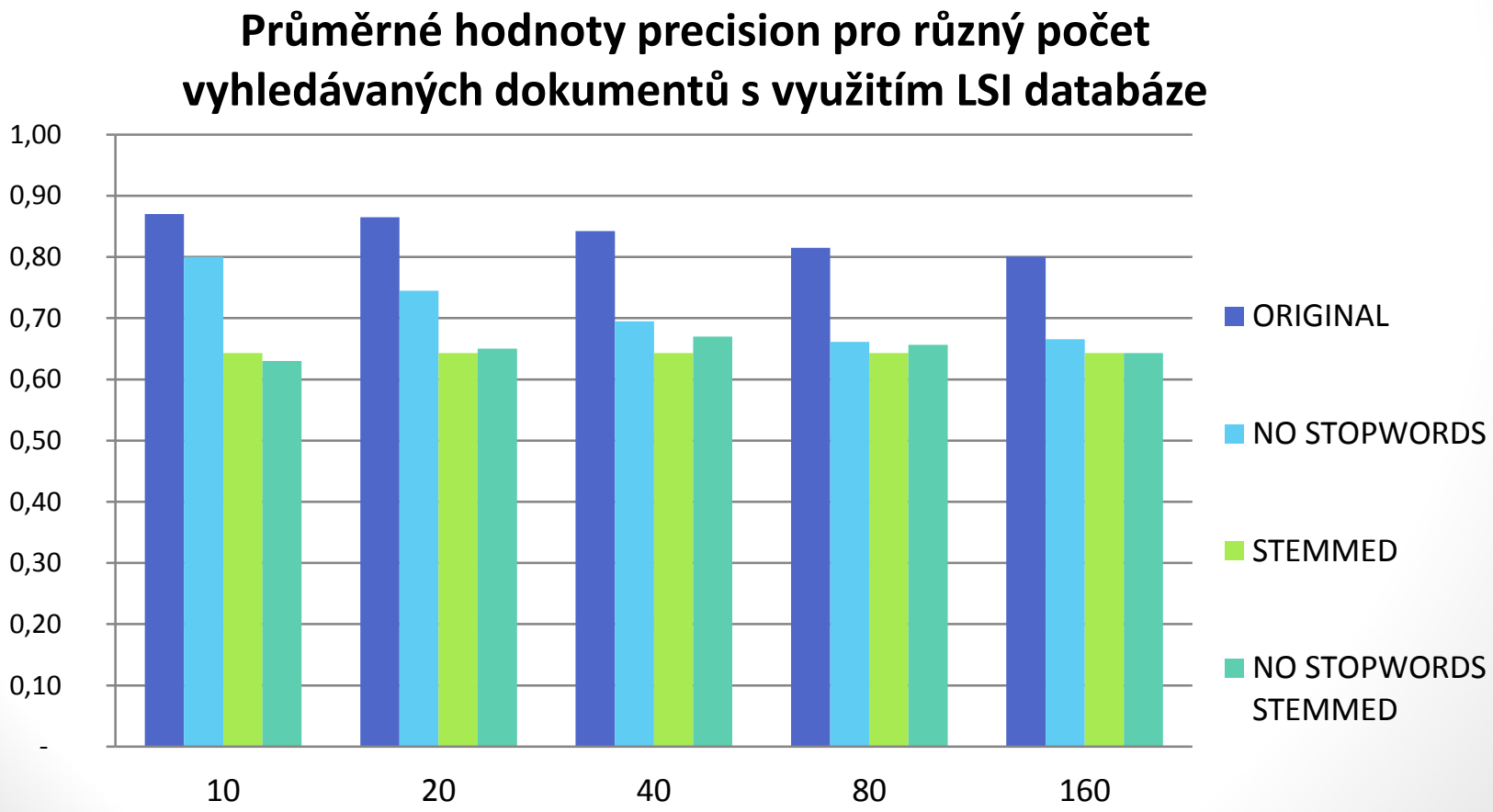


# Výsledky klasifikace FR



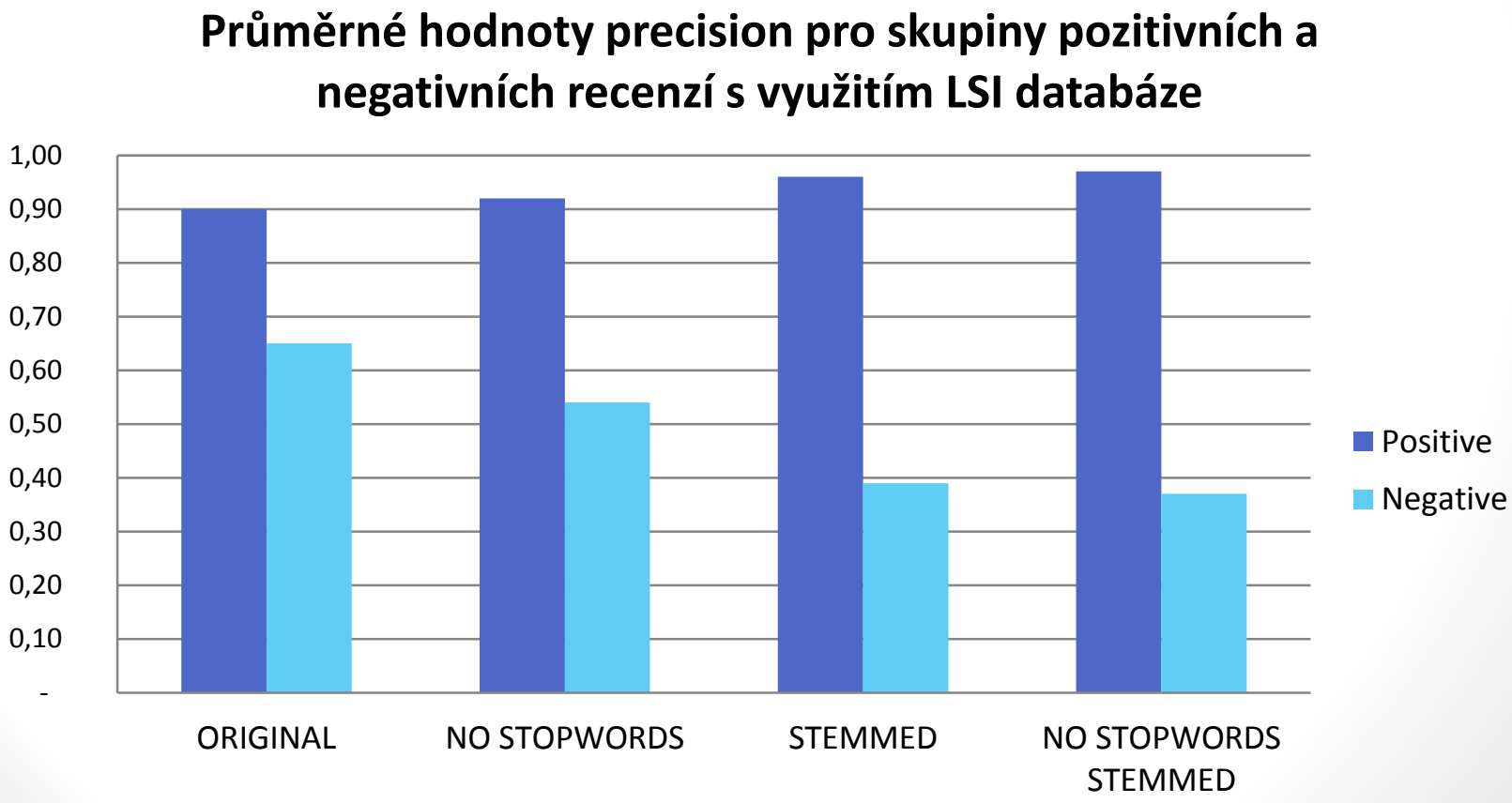
# Information retrieval

- schopnost nalézt dokumenty ze stejné třídy
  - 10 typických recenzí



# Výsledky v oblasti IR

- schopnost nalézt dokumenty ze stejného tématu
  - možnost hledat v kontextu dokumentů a termů

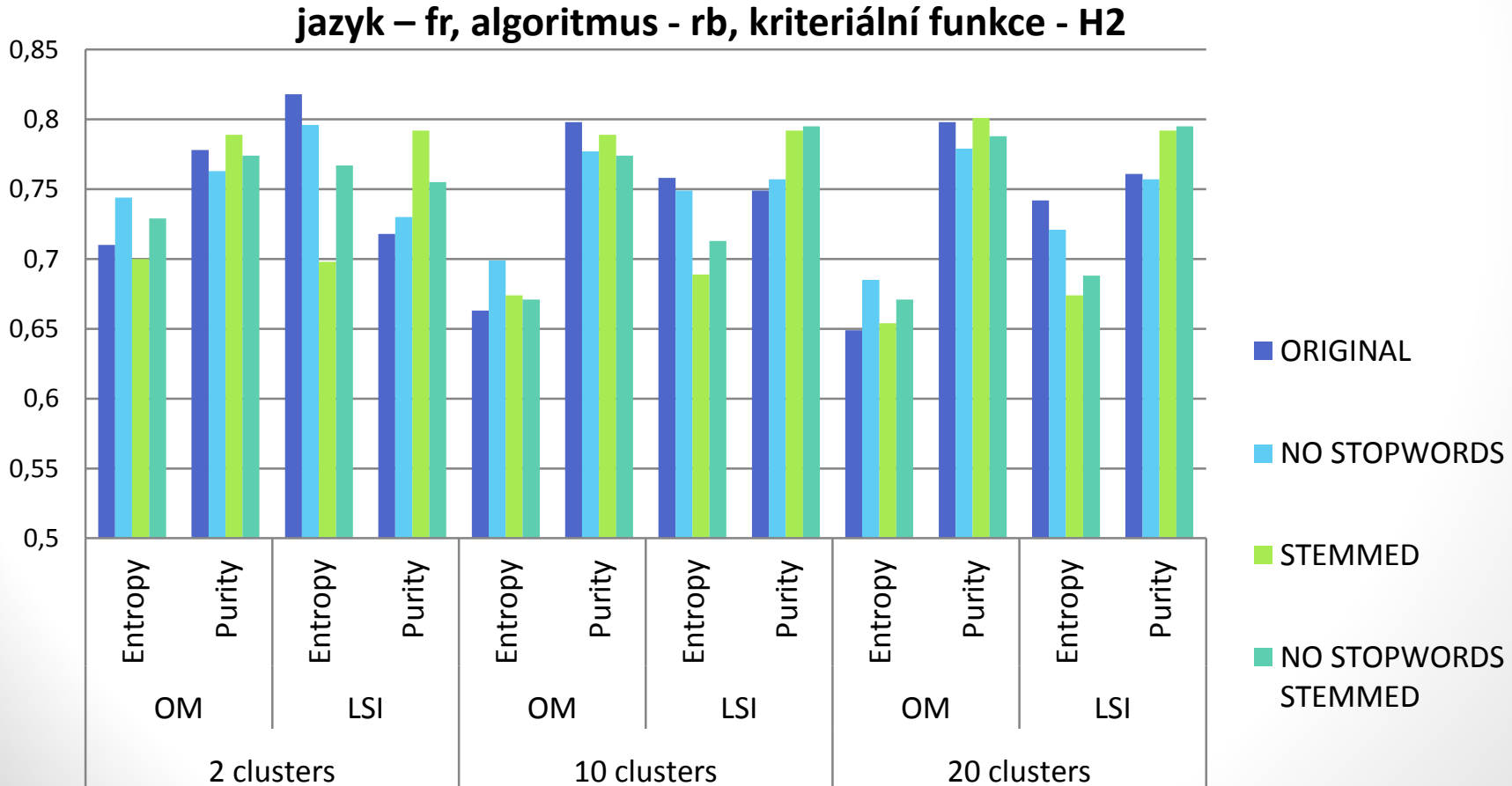


# Shlukování

- shluky ani jejich počet nebyly předem známy
- shlukování do 2, 10 a 20 shluků
- hodnocení na základě metrik entropy a purity
- vliv různého typu předzpracování

# Výsledky shlukování

- LSI databáze – lepší výsledky při shlukování do 2 shluků
- stemming – ideální metoda předzpracováná



# Závěr

- vhodné metody předzpracování dat pro vybrané jazyky
- efektivita klasifikace pro Term-Document matice a LSI databáze
- vylepšený vyhledávací model s využitím LSI databáze
- důsledky metod předzpracování dat na proces shlukování
- dílčí programy implementované v jazyce Python využitelné pro další analýzy dat