Klasifikace dokumentů se supervizovanou redukcí témat

Ondřej Háva

Agenda

- Singulární dekompozice a MNČ
- Úloha klasifikace dokumentů
- SVD klasifikátor
- Redukce latentních dimenzí
- Experimenty
- Závěr

Zadání úlohy 1

Učební kolekce N dokumentů

- **Každý dokument reprezentován reálným vektorem o M položkách**
  - Položky vektoru nazvěme termy
    - Termy tvoří slovník
  - Termy extrahovány z textu
- **Kolekce dokumentů utvoří matici D o rozměrech NxM**
  - Dokumenty v řádcích
  - Termy ve sloupcích

$$\mathbf{D} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1M} \\ w_{21} & w_{22} & \cdots & w_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NM} \end{pmatrix}$$

Zadání úlohy 2

Zatřídění učebních dokumentů do K kategorií

- **Každému dokumentu přiřazen reálný vektor**
  - Položky vektoru nazvěme kategorie
  - Zatřídění do kategorií provedeno čtenářem
  - Kategorie nemusí být disjunktní
    - Váhy odpovídají míře příslušnosti kategoriím
- **Kolekce dokumentů utvoří matici C o rozměrech NxK**
  - Dokumenty v řádcích
  - Kategorie ve sloupcích

$$\mathbf{C} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1K} \\ v_{21} & v_{22} & \cdots & v_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ v_{N1} & v_{N2} & \cdots & v_{NK} \end{pmatrix}$$

Zadání úlohy 3

Predikce vektoru kategorií pro nový dokument

- **Extrakce termů**
  - Stejný slovník jako pro učební data
  - Dokument reprezentován reálným vektorem d o M položkách
- **Predikce příslušnosti ke kategoriím**
  - Odhad všech K položek neznámého vektoru c

$$\mathbf{c} = f(\mathbf{d})$$

$$f = ?$$

Dimensionalita

- **N…dokumenty**
- **M…termy**
- **K…kategorie**
- **Obvykle N<M, K<<M**
- **Úloha lze počítat jako hledání řešení soustavy lineárních rovnic**
  - Počet soustav je roven počtu kategorií K
  - Pokud N<M je počet neznámých větší než počet rovnic a zpravidla existuje nekonečně mnoho řešení
  - Hledáme dostatečně robustní tj. obecné řešení

Redukce dimensionality

Vstupy

- **Témata extrahovaná z termů**
  - Latentní veličiny
  - Maximální počet témat: $\max(L_D) = RANK(D)$
  - $L_D <= N$
  - Možnost omezení na důležitá témata

Výstupy

- **Témata extrahovaná z kategorií**
  - Latentní veličiny
  - Maximální počet témat: $\max(L_C) = RANK(C)$
  - $L_C <= K$
  - Možnost omezení na důležitá témata

Pozn.: Témata extrahovaná z termů (vstupní) a témata extrahovaná z kategorií (výstupní) se získávají na sobě nezávisle

Singulární dekompozice

Dokumenty x termy D

$$\mathbf{D} = \mathbf{P}\Lambda\mathbf{Q}^T$$

- P…dokumenty v prostoru vstupních témat (NxL$_D$)
- Q…termy v prostoru vstupních témat (MxL$_D$)
- Λ…důležitosti vstupních témat (L$_D$xL$_D$)
- Vstupní témata jsou ortonormální

$$\mathbf{P}^T\mathbf{P} = \mathbf{I}$$

$$\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$$

Dokumenty x kategorie C

$$\mathbf{C} = \mathbf{R}\Sigma\mathbf{S}^T$$

- R…dokumenty v prostoru výstupních témat (NxL$_C$)
- S…kategorie v prostoru výstupních témat (KxL$_C$)
- Σ…důležitosti výstupních témat (L$_C$xL$_C$)
- Výstupní témata jsou ortonormální

$$\mathbf{R}^T\mathbf{R} = \mathbf{I}$$

$$\mathbf{S}^T\mathbf{S} = \mathbf{I}$$

Projekce mezi tématy

- **Přechod od reprezentace dokumentů ve vstupních tématech do reprezentace ve výstupních tématech**
  - Namísto hledání vztahu mezi kategoriemi a termy
- **Úloha lze počítat jako hledání řešení soustavy lineárních rovnic**
  - Počet soustav je roven počtu kategorií K
  - Počet neznámých je menší než počet rovnic a zpravidla neexistuje přesné řešení
  - Hledáme optimální řešení metodou nejmenších čtverců
    - Řešení normální rovnice
  - Projekční matice M o rozměrech (L_DxL_C)

$$\mathbf{R} = \mathbf{PM}$$

$$\mathbf{M} = \left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{R}$$

$$\mathbf{M} = \mathbf{P}^T\mathbf{R}$$

Vlastnosti projekce

- **Regresní koeficienty normovaných veličin**
  - Vstupní témata jsou v prostoru učebních témat ortonormální
  - Výstupní témata jsou v prostoru učebních témat ortonormální
- **Matice M je zároveň maticí kosinových podobností vstupních a výstupních témat**

$$\mathbf{M} = \mathbf{P}^T\mathbf{R}$$

$$\mathbf{P}^T\mathbf{P} = \mathbf{I}$$

$$\mathbf{R}^T\mathbf{R} = \mathbf{I}$$

Predikce kategorií nového dokumentu

- **Reprezentace dokumentu vektorem termů d**
  - Slovník učební kolekce
- **Reprezentace vektorem vstupních témat p**
  - Využití singulární dekompozice matice D
- **Projekce do vektoru výstupních témat r**
  - Lineární regrese maticí M
- **Reprezentace vektorem kategorií c**
  - Využití singulární dekompozice matice C

$$\mathbf{p} = \mathbf{d}\mathbf{Q}\Lambda^{-1}$$

$$\mathbf{r} = \mathbf{p}\mathbf{M} = \mathbf{p}\mathbf{P}^{T}\mathbf{R}$$

$$\mathbf{c} = \mathbf{r}\Sigma\mathbf{S}^{T}$$

Výsledný klasifikátor

- **Složení předchozích tří transformací v jedinou**
- **Pomocí singulární dekompozice matic D a C lze klasifikátor zjednodušit**
  - Projekce do učebních dokumentů
    - Včetně případné redukce dimenzionality vstupních témat
  - Projekce do kategorií
    - Včetně případné redukce dimenzionality výstupních témat

$$\mathbf{c} = \mathbf{d}\mathbf{Q}\Lambda^{-1}\mathbf{P}^{T}\mathbf{R}\Sigma\mathbf{S}^{T}$$

$$\mathbf{c} = \mathbf{d}\mathbf{D}^{-}\mathbf{C}$$

$$\mathbf{D}^{-} = \mathbf{Q}\Lambda^{-1}\mathbf{P}^{T}$$

Redukce latentních dimenzí

- **Výběr vstupních a výstupních témat**
  - Výraznější je zřejmě redukce vstupních témat
  - Klasifikátor po SVD redukci nemění svůj tvar
- **Nesupervizovaná**
  - Důležitost vstupních resp. výstupních témat je dána singulárními hodnotami matice D resp. C
- **Supervizovaná**
  - Výběr řádků resp. sloupců matice M podle jejich hodnot
  - Řádky reprezentují nezávislá vstupní témata, sloupce nezávislá výstupní témata
  - **Není třeba iterovat**

Acrea
analytical creativity

Experimenty

Kolekce dokumentů

Český jazyk

- **645 dokumentů**
- **Tiskové zprávy ČTK a GP z června 2007**
- **Průměrná velikost 5kB**
- **8 disjunktních kategorií**
  - Auto (82)
  - Bydlení (73)
  - Cestování (61)
  - Kultura (89)
  - Praha (60)
  - Z domova (90)
  - Zdraví (94)
  - Ze světa (96)

Anglický jazyk

- **201 dokumentů**
- **Staženo z UCI Machine Learning Repository**
  - Reuters transcribed subset
- **Pořízeny jako strojový přepis ústních výpovědí**
  - Nižší kvalita
- **Průměrná velikost 2kB**
- **10 disjunktních kategorií**
  - Acquire (20)
  - Corn (20)
  - Crude (20)
  - Earn (20)
  - Grain (20)
  - Interest (20)
  - Money (20)
  - Ship (20)
  - Trade (20)
  - Wheat (20)

Příprava strukturovaných dat

- **Rozdělení na trénink a test 70:30**
- **Textové soubory rozděleny na slova**
- **Z tréninkových dokumentů sestaven slovník**
  - Vyřazena máločetná slova
  - Další redukce slovníku v České kolekci
    - Stop slova, slova s číslicemi, slova se speciálními znaky
- **Matice dokumentů D obsahuje tf-idf váhy**

$$tfidf = tf \log(n / df)$$

- **Matice kategorií C obsahuje 1/0 indikátory**
  - Pro disjunktní kategorie nezávislé sloupcové vektory

Matice

Česká kolekce

- **Matice dokumentů D**
  - Trénink
    - Řádky: 448
    - Sloupce: 5108
  - Test
    - Řádky: 197
    - Sloupce: 5108
- **Matice kategorií C**
  - Trénink
    - Řádky: 448
    - Sloupce: 8
  - Test
    - Řádky: 197
    - Sloupce: 8

Anglická kolekce

- **Matice dokumentů D**
  - Trénink
    - Řádky: 133
    - Sloupce: 1487
  - Test
    - Řádky: 68
    - Sloupce: 1487
- **Matice kategorií C**
  - Trénink
    - Řádky: 133
    - Sloupce: 10
  - Test
    - Řádky: 68
    - Sloupce: 10

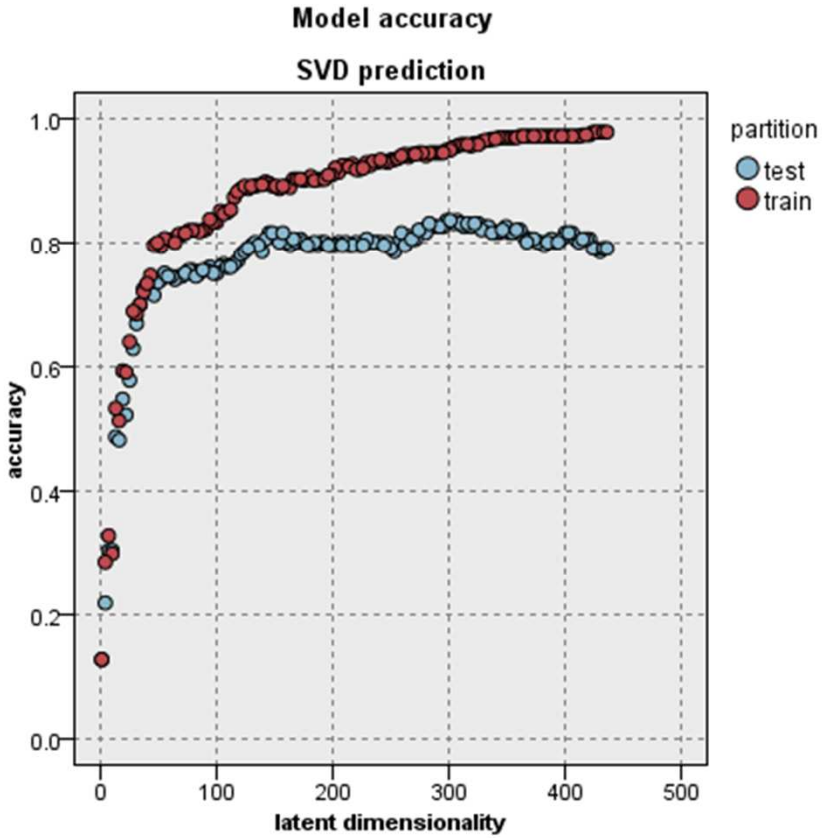
Redukce témat

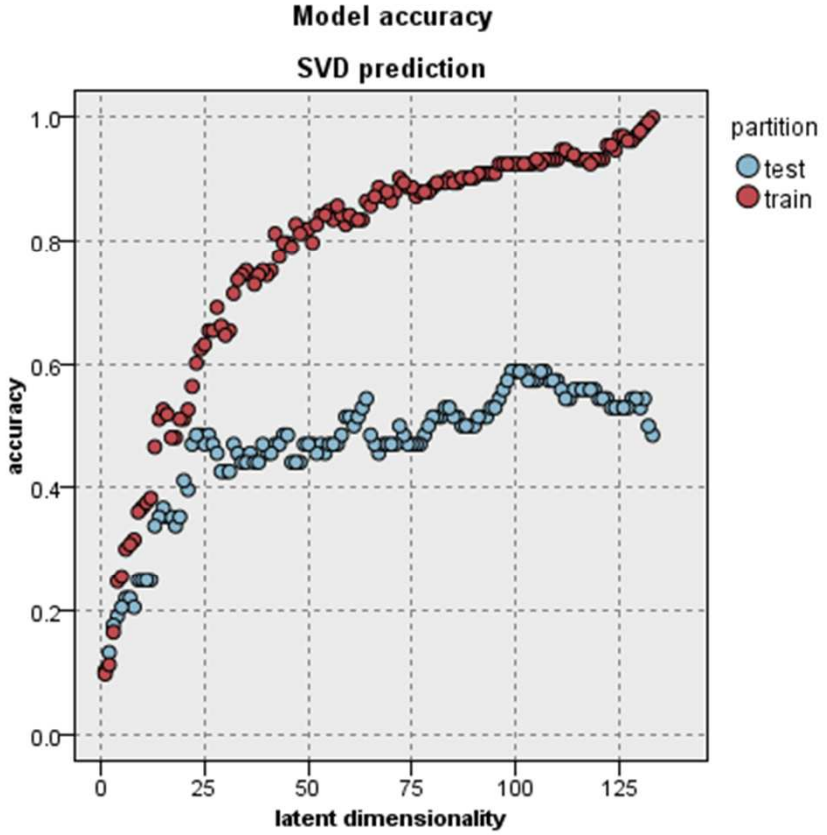
- **Redukce dimensionality matice dokumentů D**
  - Nejvýznamnější témata v dokumentech
    - Nesupervizovaná
    - Témata řazena podle singulárních hodnot sestupně
    - Ponechán vybraný počet prvních témat
  - Maximální důležitost témat přes všechny kategorie
    - Prahová hodnota kosinové podobnosti v převodní matici M
    - Témata řazena podle maximální hodnoty v řádcích M
    - Ponechána témata s nadprahovou hodnotou
  - Individuální důležitost témat v jednotlivých kategoriích
    - Prahová hodnota kosinové podobnosti v převodní matici M
    - Témata vybírána pro predikci každé kategorie zvlášť
    - Podprahové hodnoty v matici M nahrazeny nulami
- **Redukce dimensionality matice kategorií C**
  - Není, matice má ortogonální sloupce

Výsledky evaluace

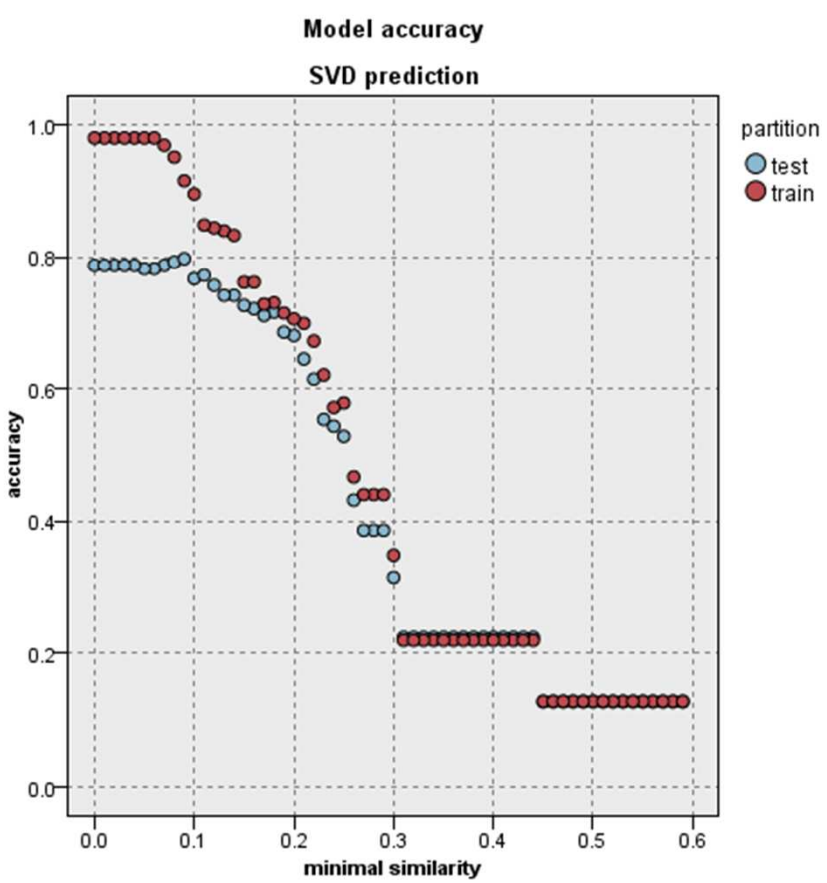
- **Absolutní přesnost klasifikátoru**
  - V závislosti na počtu latentních témat
  - V závislosti na prahové hodnotě kosinové podobnosti
- **Počet latentních témat**
  - V závislosti na prahové hodnotě kosinové podobnosti

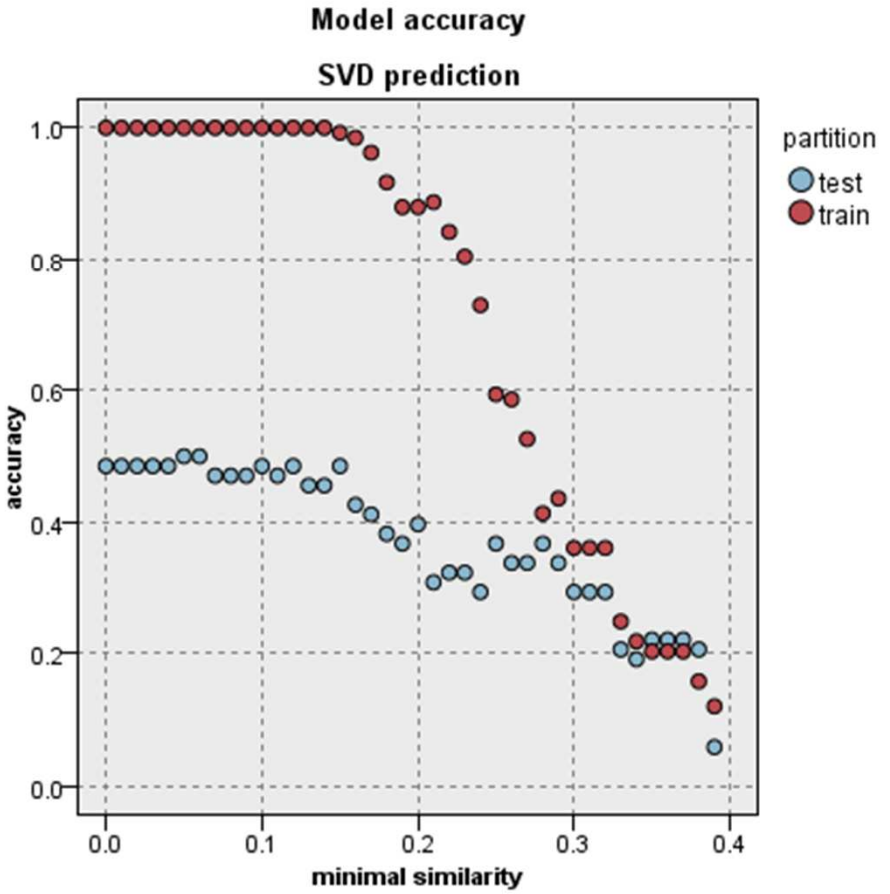
Nejvýznamnější témata



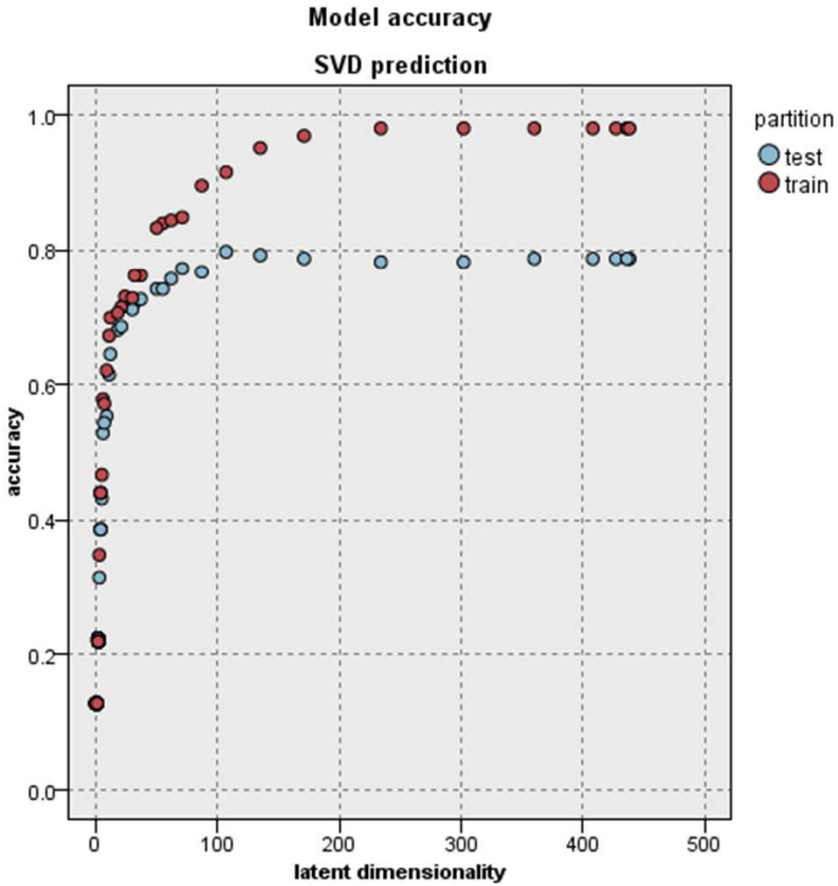


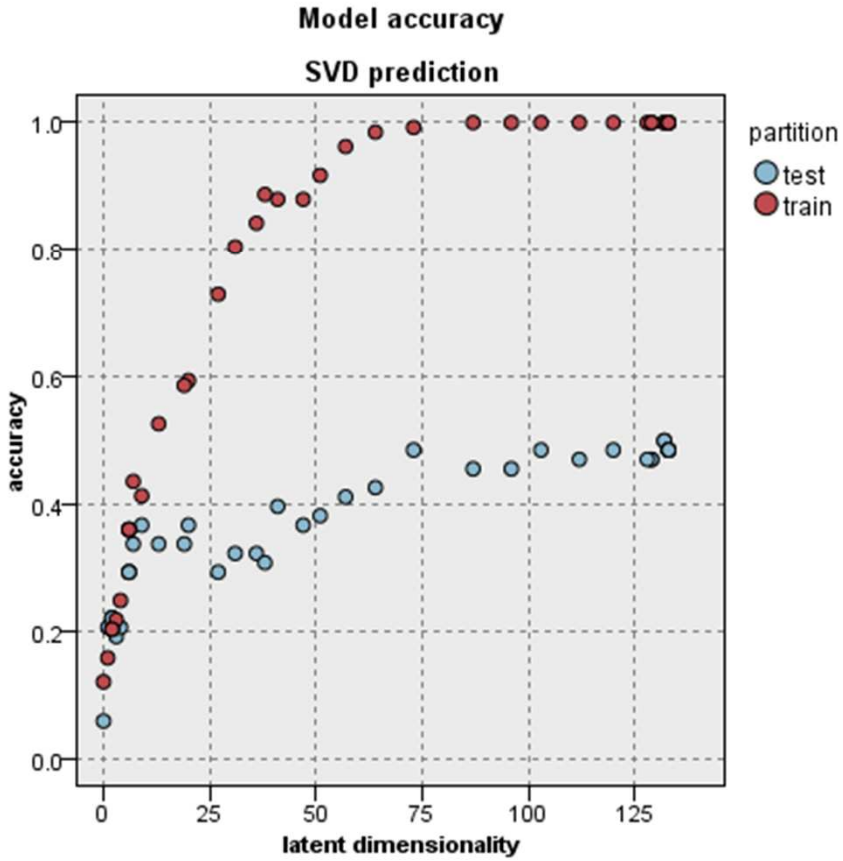
Maximální podobnost 1



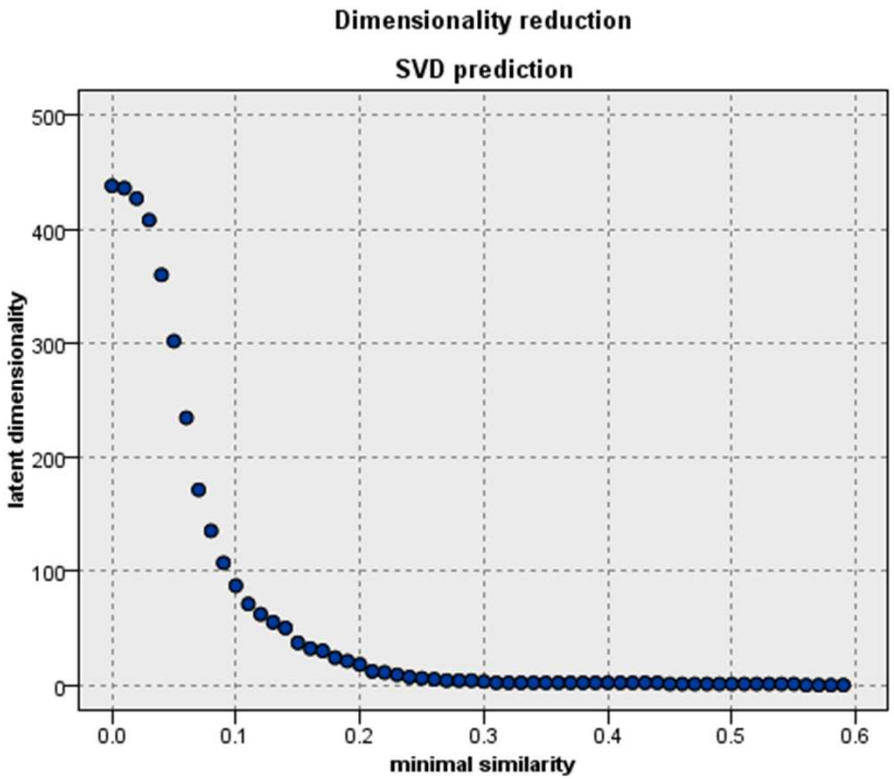


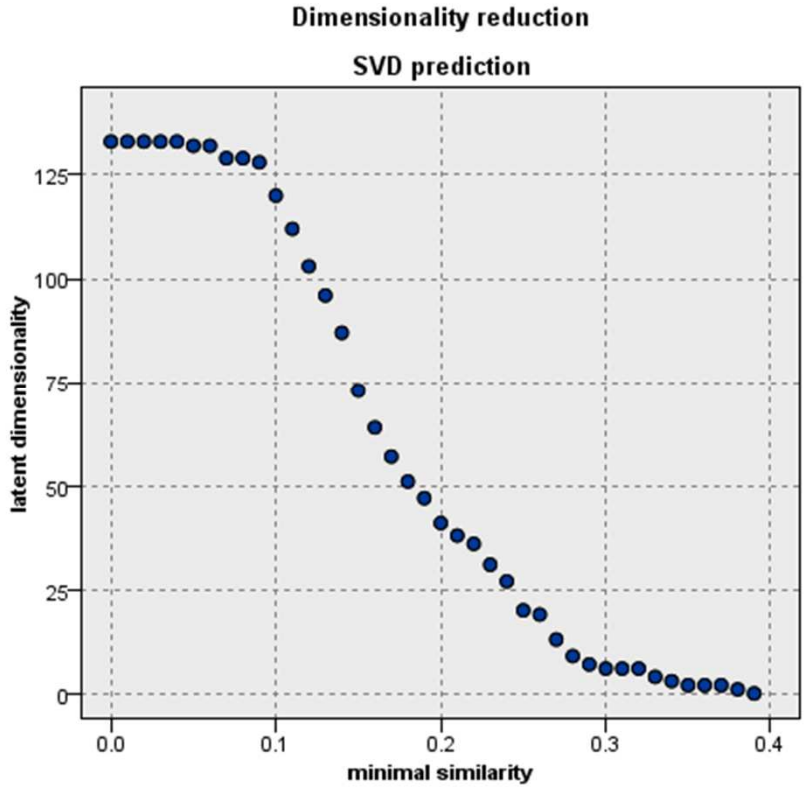
Maximální podobnost 2



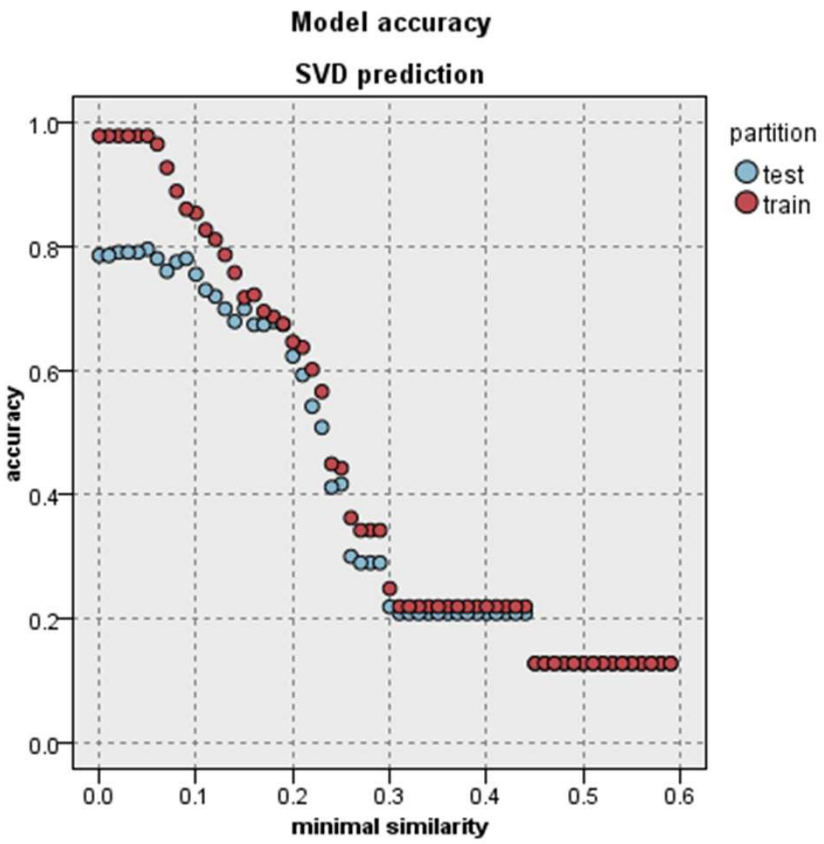


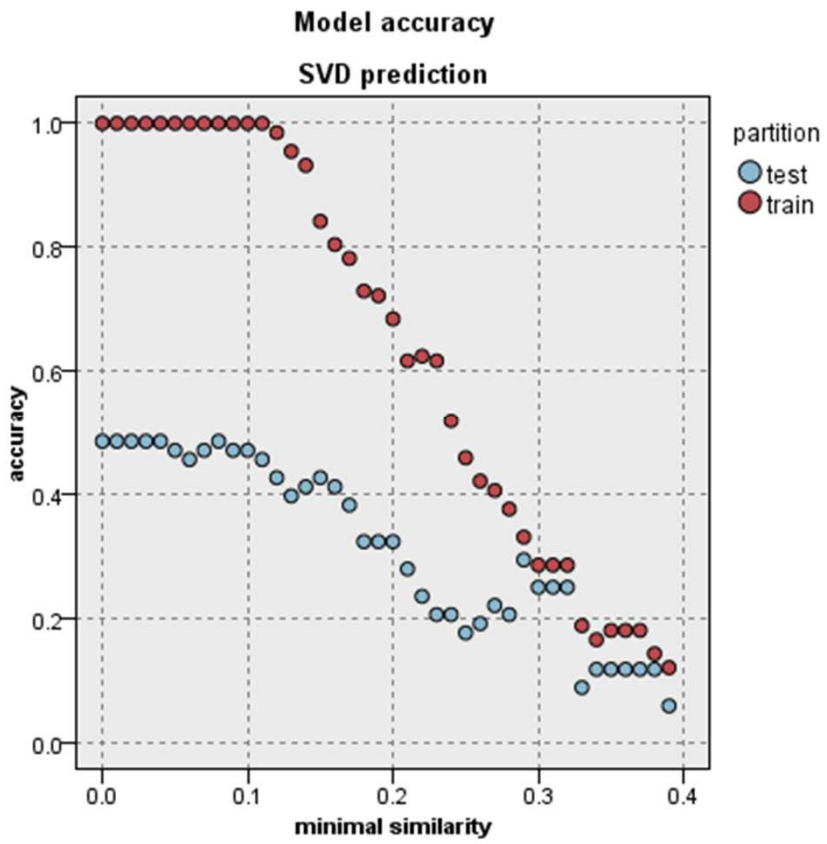
Maximální podobnost 3





Individuální podobnost





Závěry experimentů

- Redukce dimensionality významně přispívá ke zvýšení obecnosti klasifikátorů
- Optimálního kompromisu mezi přesností a obecností lze dosáhnout supervizovaným i nesupervizovaným výběrem
- Při supervizovaném výběru témat dosáhneme zvolené přesnosti při menším počtu latentních témat

Děkuji za pozornost

Přesnost predikce bez redukce dimesionality

Česká kolekce

- **Strom C5.0**
  - Trénink: 89,7%
  - Test: 52,8%
- **Podpůrné vektory**
  - Trénink: 98,0%
  - Test: 50,3%

Anglická kolekce

- **Strom C5.0**
  - Trénink: 84,2%
  - Test: 32,4%
- **Podpůrné vektory**
  - Trénink: 100%
  - Test: 8,8%