Vytěžování Dat
Cvičení 5 – Textmining

Miroslav Čepek, Michael Anděl

21. 10. 2014

Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

Základní kroky pro text mining

1. Získání dokumentů a nahrání do Rapidmineru (či jiného SW)
2. Tokenizace (rozklad textu na jednotlivá slova)
3. Odfiltrování častých a nezajímavých slov
4. Převod slov na kořeny slov (**stemming**)
   - Převod na jednotná čísla
   - Převod různých časování/způsoby/vidy na infinitivy
   - Převod mezi různými variantami slov (příslovce, přídavná jména $\leftarrow$ podstatná jména).
5. Vytvoření "word vectoru". (Převod slov na čísla).
6. Tvorba modelu.

Instalace rozšíření pro Textmining

- Standardní instalace Rapidmineru neobsahuje rozšíření pro Textmining.
- Musíte nainstalovat rozšíření, ale naštěstí je to velmi jednoduché :).
- Z menu *Help* vyberte *Update RapidMiner*. Zde zaklikněte *Text Processing* a *Web Mining*.
- A klikněte na *Install*.

Získání dokumentů a nahrání do Rapidmineru

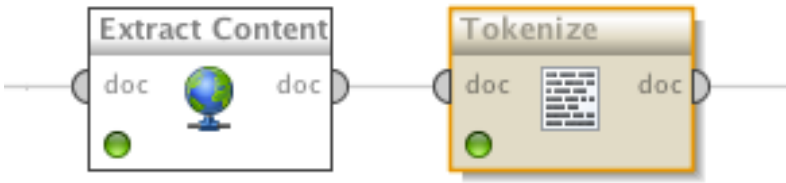
- Existuje několik uzlů, pro nahrávání dat do RapidMineru.
- Pro naše účely, kdy máme dokumenty různých typů v různých složkách, nejlépe vyhovuje uzel *Text Processing* $>$ *Process Documents from Files*.
- Jedná se o super-uzel, který bude obsahovat pod-proud transformující dokumenty na číselné vektory.

Extrakce textů z HTML

- První krok je extrakce textů z HTML (resp. odstranění HTML tagů).
- Pro to budete potřebovat uzel *Extract Content* > *HTML Processing* > *Extract Content*.

Tokenizace

- Rozklad na jednotlivá slova.
- Slova se rozdělují typicky podle "ne" písmenek. Takto získaná slova se označují jako **termy**.



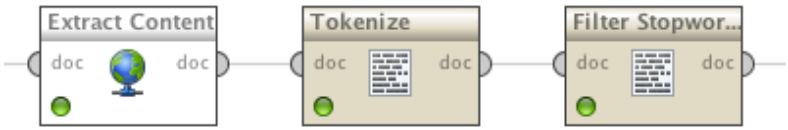
- V Rapidmineru existuje uzel "Tokenize", který najdete *Text Processing* > *Tokenization* > *Tokenize*.
- Možnosti rokladu na slova jsou: `non-letters`, `specify-characters`, `regular expression`, `linguistic tokens`, `linguistic token`.

Tokenizace (2)

- Zkuste spustit proud nyní.
- Výsledkem bude `word` objekt, který si můžete prohlédnout.
- Uvidíte počty slov podle typů dokumentů. A také celkový počet slov.
- Každé slovo nakonec bude reprezentovat vstupní proměnnou.

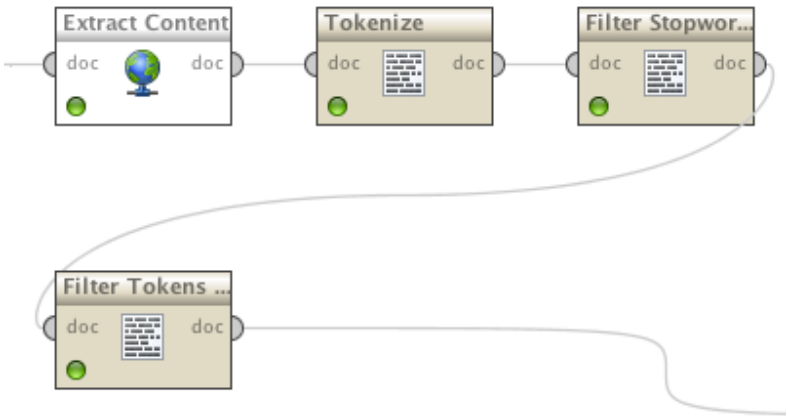
Filtrování častých a nezajímavých slov

- Protože vstupních proměnných bude i tak moc, je vhodné některé z nich eliminovat.
- První způsob je filtrování obvyklých a nezajímavých slov.
- V Rapidmineru se to děje uzlem *Text Processing* > *Filtering* > *Filter Stopwords (English)*.
- Tím z dokumentu odstraníte termy (slova), která se v angličtině vyskytují příliš často.
- Například spojky, běžná slovesa, předložky, apod...
- Uzel v Rapidmineru obsahuje seznam předdefinovaných slov.



Filtrování častých a nezajímavých slov (3)

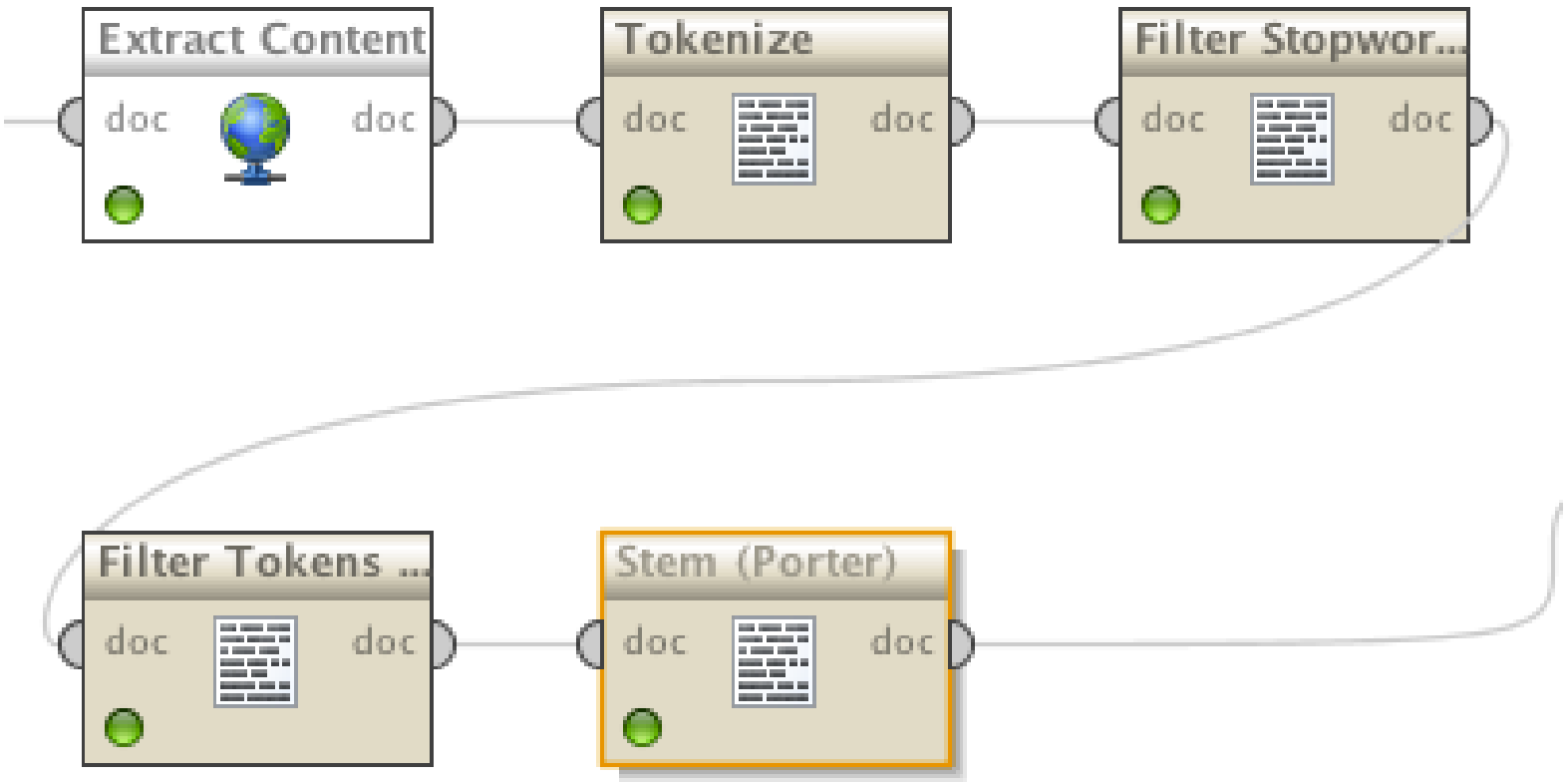
- Stejně tak může (ale nemusí) být dobrý nápad vyfiltrovat slova, která jsou příliš dlouhá nebo příliš krátká.
- K tomu slouží *Text Processing* > *Filtering* > *Filter Tokens (by Length)*.



Převod slov na kořeny slov – Stemming

- Exituje několik způsobů, jak najít kořen slova.
- Například hrubou silou – tj tabulka mapující každé slovo a každý jeho tvar na odpovídající kořen.
- Jeden z dalších používaných algoritmů (pro Angličtinu) je tzv. Porterův algoritmus.
  - Iterativně odebírá známé koncovky anglických slov.
  - Má seznam přípon a ty se pokouší postupně odebrat (pokud to lze).
  - Například `HOPEFULNESS` $\rightarrow$ `HOPEFUL` $\rightarrow$ `HOPE`.
- `http://tartarus.org/martin/PorterStemmer/def.txt`

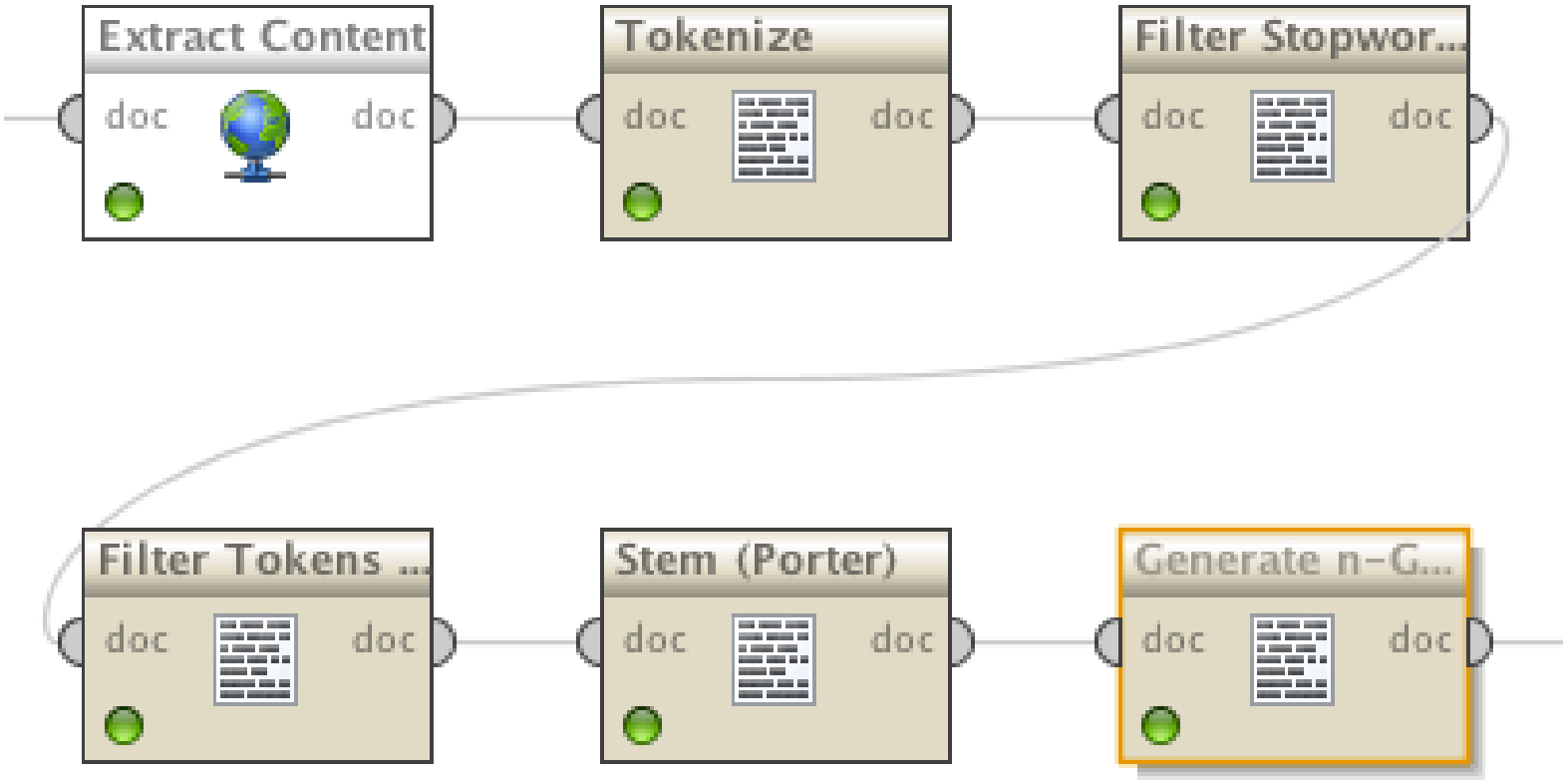
Převod slov na kořeny slov – Stemming (2)



Kombinace slov - N-Grams

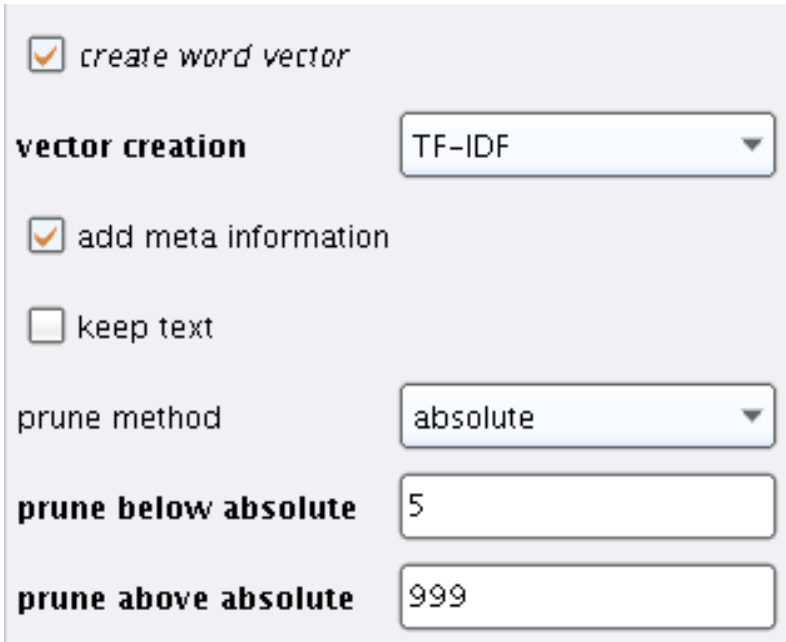
- Někdy se v dokumentech vyskytují zajímavé kombinace (po sobě jdoucích) slov.
- N-Gram je term, který obsahuje posloupnost term maximální délky N.
- Uzel *Text Processing* > *Transformation* > *Generate n-Grams (Terms)* vygeneruje vsechny kombinace termů.

Kombinace slov



Vlastnosti uzlu Process Documents from Files

- Jednak umožňuje zahodit málo (nebo moc) často se vyskytující termy (slova a n-gramy).
- Jednotlivé možnosti vybíráte combo-boxem `Prune method`.
- Další důležitá věc je zaškrtnout `Create word vector`.
- A vybrat vhodnou metodu pro `Vector creation`.



Vytvoření "word vectoru"

- Nyní máme slova (termy) a jejich počty v jednotlivých dokumentech.
- Před předložením shlukovací (či jakékoliv jiné) metodě je potřeba tyto počty nějak přetransformovat.
- V Rapidmineru jsou na výběr následující možnosti:
  - Term Frequency – normalizovaný počet výskytů termu $(\frac{\text{počet výskytu termu}}{\text{celkový počet termů}})$
  - Term Occurences
  - Binary Term Occurences
  - **TF-IDF**

Term Frequency - Inverse Document Frequency

- Míra ukazující, jak moc je term specifický pro daný dokument.
- Zahrnuje v sobě dvě části Term Frequency a Inverse Document Frequency.
- Term Frequency je definován takto:

$$tf(t) = \frac{\text{počet výskytu termu}}{\text{celkový počet termů}}$$

Term Frequency - Inverse Document Frequency (2)

- Inverse Document Frequency ukazuje, jak často se vyskytuje term v ostatních dokumentech.

$$idf(t) = \log \frac{\|D\|}{\|\{d : t \in d\}\|}$$

- $\|D\|$ – Celkový počet dokumentů.
- $\|\{d : t \in d\}\|$ – Počet dokumentů, ve kterých se term t vysktytuje.

Term Frequency - Inverse Document Frequency (3)

- Term Frequency - Inverse Document Frequency nakonec získáme, když tyto dvě míry vynásobíme.

$$td - idf(t, d) = tf(t, d) * idf(t)$$

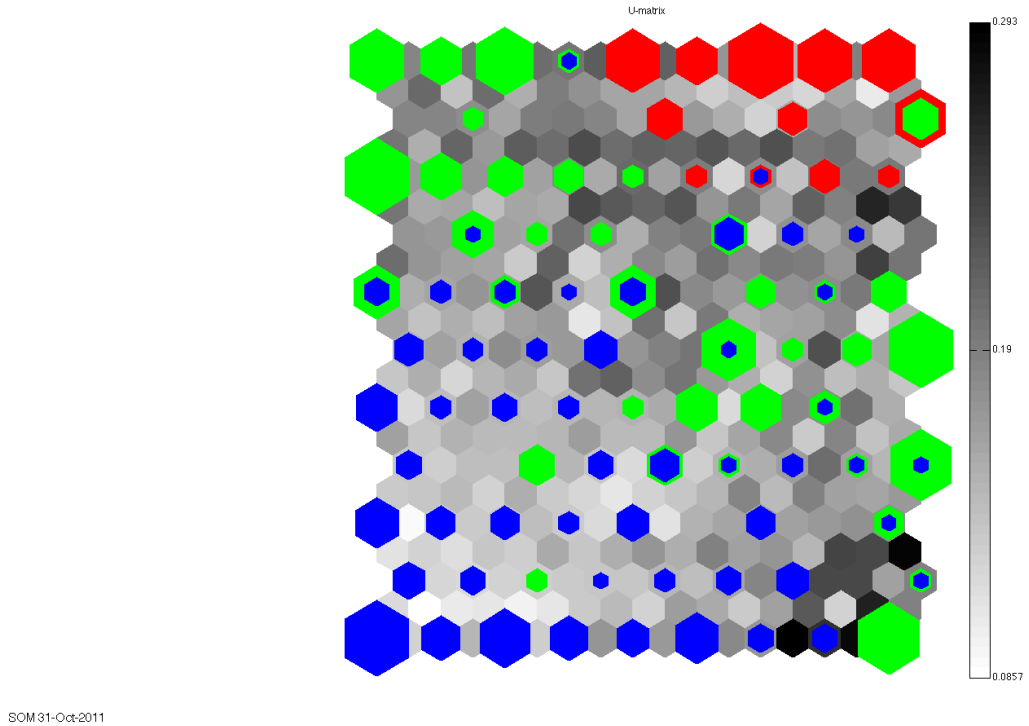
Export dat do CSV a import do MATLABu

- V RapidMineru bohužel nejsou žádné vhodné shlukovací metody. Čili použijeme Matlab a SOM toolbox.
- K exportu z RapidMineru lze použít uzel *Export* > *Data* > *Write CSV*
- Abychom se nemuseli trápit v Matlabu s načítáním ošklivých hodnot, můžeme využít uzlu *Export* > *Data* > *Write CSV* k odstranění sloupců, které obhashují nečíselná a pomocná data.
- V mém případě jde o sloupce: Description, Keywords, Language, Robots, Title, label, metadata_date, metadata_file, metadata_path.
- Pro import použijeme v MATLABu funkci `importdata`.

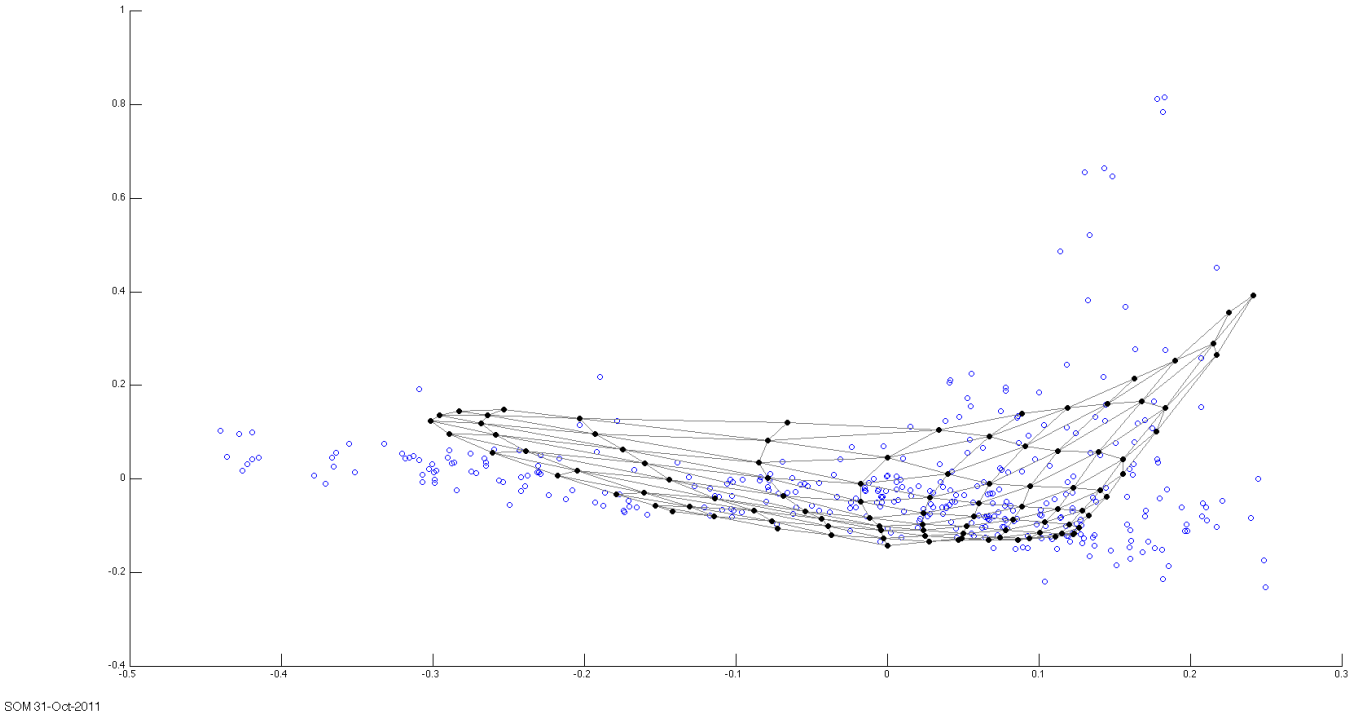
Shlukování v SOM toolboxu

- Vytvoření a naučení SOM mapy:
- `map = som_make(x.data);`
- Zobrazení UMatice:
- `som_show(map, 'umat','all')`
- Jak to dopadlo?

UMatice se zobrazenými třídami dokumentů



Zobrazení mapy a dat pomocí PCA projekce



Zadání domácího úkolu

- Je potřeba zjistit, která fóra se odlišují a naopak která jsou těžko rozlišitelná.
- Pomocí SOM shlukování (praktické povídání bude příště) vytvořte shluky dodaných dokumentů.
- Dokumenty obsahují zprávy z několika diskusních fór. Každé fórum má jeden adresář a každá zpráva v něm je jeden soubor.
- Ze stránek předmětu (cvičení) stáhněte tato data.
- Z dokumentů extrahujte důležitá slova a příznakové vektory pomocí rozšíření rapidmineru pro textmining.
- Takto extrahovaná data uložte do CSV souboru.

Zadání domácího úkolu (2)

- Tento CSV soubor načtěte do MATLABu pomocí funkce dlmread (nebo podobné).
- Pomocí SOM Toolboxu shlukněte načtená data a pomocí různých vizualizací zobrazte výsledky shlukování.
- Za pomoci vizualizací učiňte závěry, která fóra jsou si podobná z pohledu použitých slov (strojově obtížně odlišitelná) za pomoci text-miningu a která naopak zle odlišit jednoduše.
- Pomocí funkce kmeans_clusters zjistěte, kolik segmentů v datech je. A pomocí funkce som_cplane si zobrazte, které neurony patří do stejného clusteru. Pomocí som_hits zobrazte, které typy dokumentů jsou ve kterých neuronech a pak učiňte závěr, které dokumenty jsou si podobné.

Nastavení textmeningu

- Tokeny (slova) jsou odděleny znaky, která nejsou písmena.
- Doporučuji, abyste vyfiltrovali příliš krátká slova (řekněme kratší než 5 znaků) a často se vyskytující slova (stopwords) – předložky, spojky, ...
- Pro hledání kořenů slov použijte Porterův algoritmus.
- Volitelně můžete zkusit zkusit zkontruovat n-gramy (tokeny sestávající se z více slov) – doporučuji maximálně 3 slova.
- Také doporučuji odstranit slova, která se vyskytují příliš řídce (příliš málo -krát).

Obsah zprávy

- Zpráva bude obsahovat:
- Popis proudu v Rapidmineru, kterým jste vyextrahovali příznaky z dokumentů a jeho screenshot (alespoň důležité části).
- Popis postupu, jakým jste vytvořili SOM síť a její vizualizace.
- Vytvořené vizualizace a jejich popis.
- Závěr o tom, zda se diskusní fóra podobají nebo ne (případně která).
- **DEADLINE**: 11. 11. 2014

Užitečné zdroje o Textminingu

- Pokud se chcete podívat, jak se textmining provádí v Rapidmineru, doporučuji následující sérii videí:
  - `http://www.youtube.com/watch?v=hpvda_Rfg3s`
  - `http://www.youtube.com/watch?v=EjD2M4r4mBM`
  - `http://www.youtube.com/watch?v=vhMzUi-FMy0`
  - `http://www.youtube.com/watch?v=ToxzfYECxOU`
  - `http://www.youtube.com/watch?v=BRvjWLwSScQ`
  - `http://www.youtube.com/watch?v=9I0BcMuhPe8`
- Video přednáška o Textminingu
  `http://videolectures.net/ess07_grobelnik_twdmI/`

Užitečné zdroje o Textminingu (2)

- http://eprints.pascal-network.org/archive/00000017/01/Tutorial_Marko.pdf
- http://www.cs.sunysb.edu/~cse634/presentations/TextMining.pdf

Užitečné příkazy SOM Toolboxu

- som_demo1, som_demo2, som_demo3, som_demo4
- som_randinit
- som_make
- som_quality
- som_show
- Kompletní dokumentaci všech funkcí naleznete na `http://www.cis.hut.fi/projects/somtoolbox/package/docs2/somtoolbox.html`