

Vytěžování dat, cvičení 7: Textmining

Miroslav Čepek



Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

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 ← podstatná jména).
5. Vytvoření "word vectoru". (Převod slov na čísla).
6. Tvorba modelu.

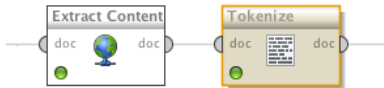
- ▶ Standardní instalace Rapidmineru neobsahuje rozšíření pro Textmining.
- ▶ Musíte nainstalovat rozšíření, ale naštěstí je to velmi jednoduché :).
- ▶ Z menu *Help* vyberte *Update RapidMiner*. Zde zaklikněte *Text Processing* a *Web Mining*.
- ▶ A klikněte na *Install*.

- ▶ 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.

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



- ▶ 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 rozkladu na slova jsou: non-letters, specify-characters, regular expression, linguistic tokens, linguistic token.

- ▶ 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.

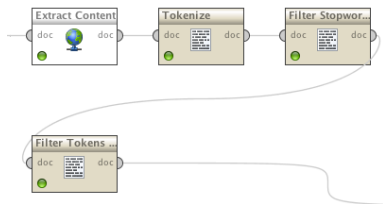
Filtrování častých a nezajímavých slov

- ▶ 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.



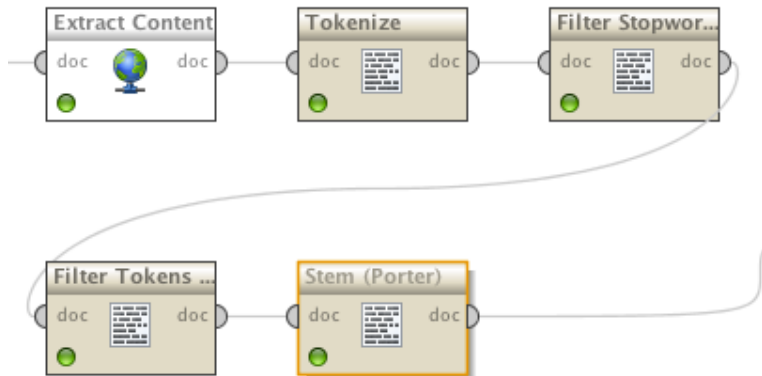
Filtrování častých a nezajímavých slov (3)

- ▶ 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)*.

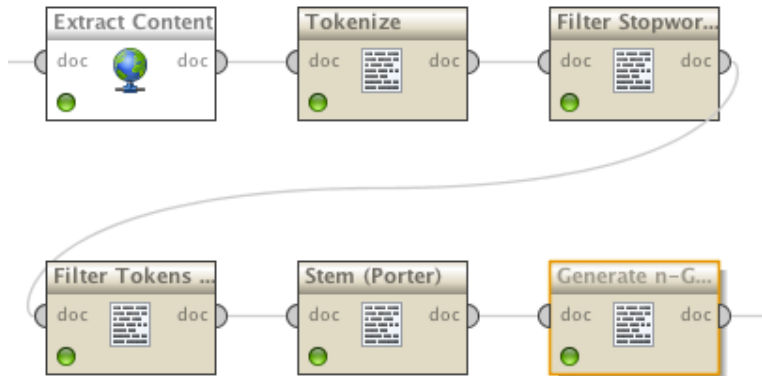


- ▶ Existuje několik způsobů, jak najít kořen slova.
- ▶ Například hrubou silou – tj tabulka mapující každé slovo a každý jeho tvar na odpovídající kořen.
- ▶ Jeden z dalších používaných algoritmů (pro Angličtinu) je tzv. Porterův algoritmus.
 - ▶ Iterativně odebírá známé koncovky anglických slov.
 -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

Převod slov na kořeny slov – Stemming (2)



- ▶ 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 všechny kombinace termů.



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.

create word vector

vector creation

add meta information

keep text

prune method

prune below absolute

prune above absolute

- ▶ 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**

- ▶ 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:

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

- ▶ 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 vyskytuje.

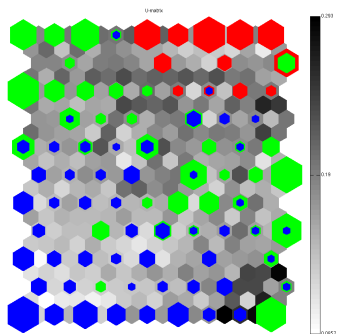
- ▶ Term Frequency - Inverse Document Frequency nakonec získáme, když tyto dvě míry vynásobíme.

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

- ▶ 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`.

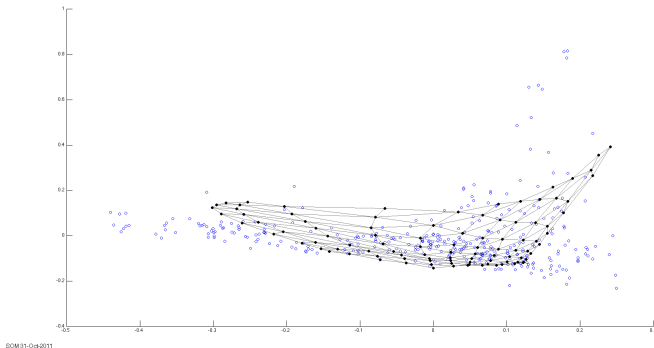
- ▶ Vytvoření a naučení SOM mapy:
- ▶ `map = som_make(x.data);`
- ▶ Zobrazení UMatice:
- ▶ `som_show(map, 'umat', 'all')`
- ▶ Jak to dopadlo?

UMatice se zobrazenými třídami dokumentů



SOM 91-Oct2011

Zobrazení mapy a dat pomocí PCA projekce



- ▶ 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://videlectures.net/ess07_grobelnik_twdmI/

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