ČESKÁ EVALUAČNÍ SPOLEČNOST

Využití text miningu při evaluacích

Tomáš Schwardy, Petr Krucký

Praha, 26. 5. 2016

institut sociologických studií
FAKULTA SOCIÁLNÍCH VĚD
UNIVERZITA KARLOVA

Lze získat informace z textu, aniž bychom ho četli?

Použitím moderních nástrojů text miningu

The better the question. The better the answer.
The better the world works.

Building a better working world

Obsah prezentace

1 Zadání evaluačního úkolu

3 Postup evaluace

2 Způsob využití text miningu

Zadání evaluačního projektu

Zadání evaluace

Cíle evaluace

- Identifikovat faktory, které mají vliv na ekonomickou udržitelnost a replikovatelnost projektů podpořených z programu LIFE
- Vytvořit scoring model, předpovídající pravděpodobnost ekonomické životaschopnosti a replikovatenosti nových projektů

Zadavatel

- NEEMO EEIG – subjekt zodpovědný za monitoring projektů podpořených z LIFE
- DG Environment – finální uživatel výstupů

Rozsah evaluace

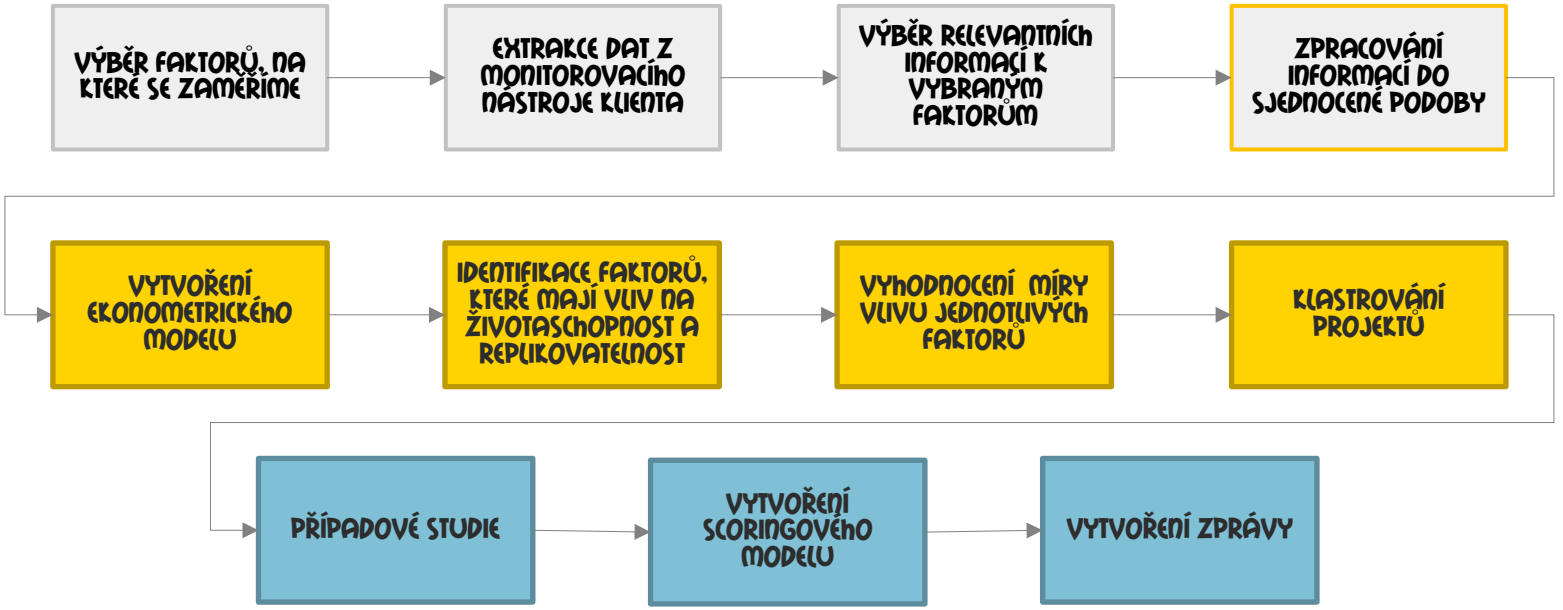
- Přes 4 500 projektů podpořených z programu LIFE
- Takřka 900 projektů, které byly detailně analyzovány

LIFE program

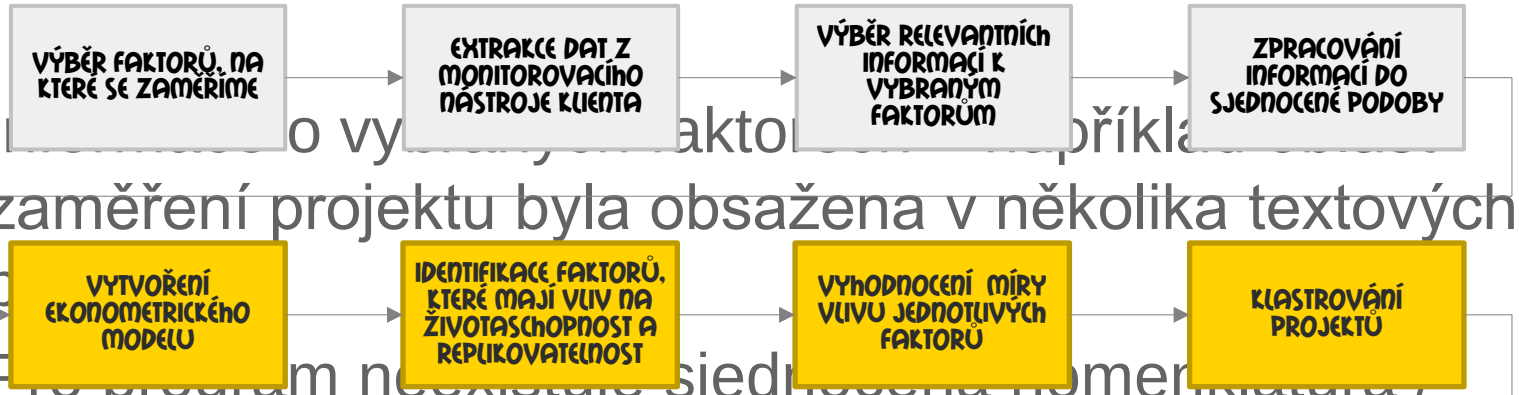
- Finanční instrument (dotační program) spravovaný Evropskou komisí podporující environmentální, konzervační a klimatické aktivity v EU
- Počátek 1992
- Hlavní cíle:
  - Fungovat jako katalyzátor
  - Prosazování a integrace cílů v oblasti životního prostředí
  - Přispět k lepší správě
  - Přispět k naplnění priorit EU: hospodárné využívání zdrojů, ztráta biodiverzity a potlačení změny klimatu

Postup evaluace

Zvolený postup evaluace



Zvolený postup evaluace



- Informace o vybraných faktorech (například oblast zaměření projektu byla obsažena v několika textových p
- Pro program neexistuje sjednocená nomenklatura / slovník
- V progra stuje obla y / prioritních os
- Popis cíle projektu / jeho zaměření je subjektivizováno

K čemu jsme došli

- Identifikace sektorů pro všechny projekty by byla extrémně časové náročná

- Větší část projektů není pro evaluaci relevantní

- Potřeba sjednoceného zařazení projektů
- Vysoká pravděpodobnost chybovosti při ručním zatřídění

Způsob využití text miningu

Využili jsme text mining k identifikaci oblasti zaměření jednotlivých projektů

- Zadání pro naše data science specialisty:
  - Pro každý z 4 500 projektů potřebujeme zjistit jeho téma/oblast, na kterou se zaměřuje
  - Zdrojem jsou textová pole obsažená v monitorovací databázi

Zdroj informací

ANNEX 4: PROJECT INFORMATION

PROJECT BACKGROUND

De Liereman is a complex area of habitats laid out in a mosaic pattern which includes wet depressions and steep continental dunes. The ecological gradients between sandy. loamy and peaty soils and the associated vegetation communities made this an outstanding landscape. Here. after the deforestation of the Middle Ages. heathland developed in combination with small-scale grasslands and arable fields within a context of low-key farming.

However. in order to better exploit the land. the depressions were drained and the dunes were afforested with pine (in the 19th century. timber was urgently needed for the shafts in the coal mines). In the 20th century. agricultural intensification further affected the heathlands and semi-natural grassland. which only survived as relicts inside a nature reserve. Most of the oligotrophic lakes disappeared or were turned into fishponds; small wet heathlands. mire vegetations (small raised bogs and *Cladium mariscus* vegetations) and brook forests remained but suffered from desiccation and increased fertilization. The former important *Nardetalia* grassland only survived as tiny relicts along road verges.

Thus by the end of the 20th century the Liereman had lost most of its conservation value to habitat fragmentation. conifer plantations. changes in the natural hydrology and lack of appropriate land use.

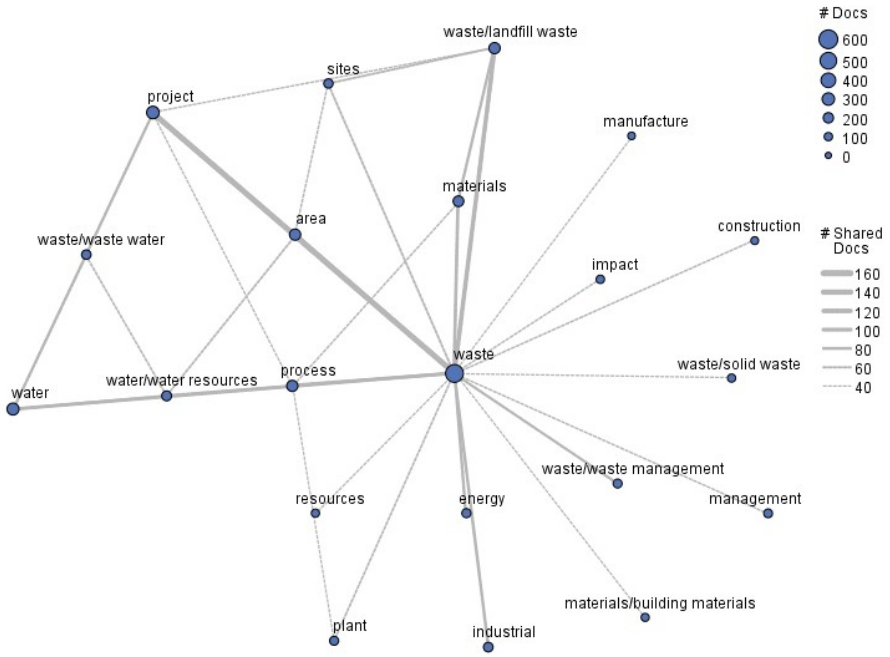
Identifikované témata

- Zemědělství <= “farming”
- Vodohospodářství = „meadow area“, „wetland area“
- Lesnictví <= „deforestation“
- Hornictví / Těžební průmysl <= „coal“
- Ochrana přírody <= „conservation“
- Odpady <= „waste“

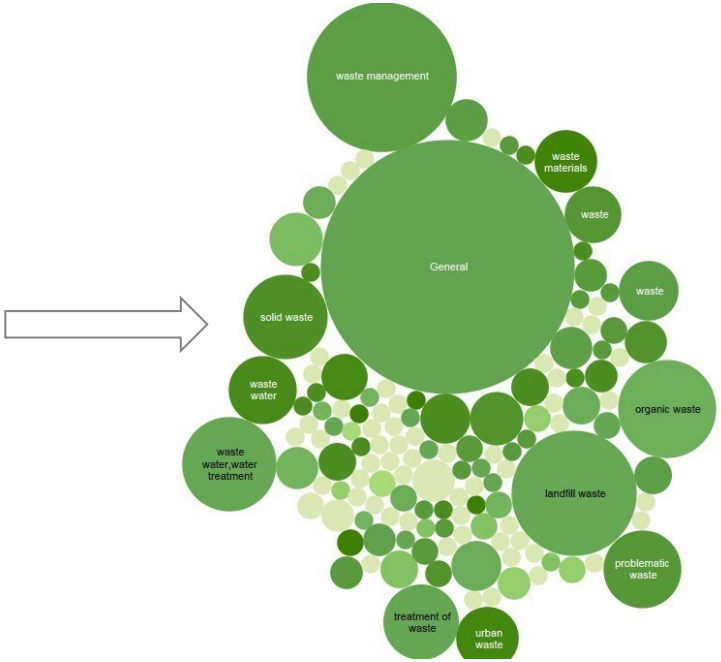
Metoda využitím odborné kategorizace

- Vyvinuli jsme nástroj, který přiřazuje kategorie k projektům LIFE na základě významných slov a frází, které se objevují v textech, a následně je shlukuje do větších celků dle témat, kterých se týkají.

Skládání koncových kategorií

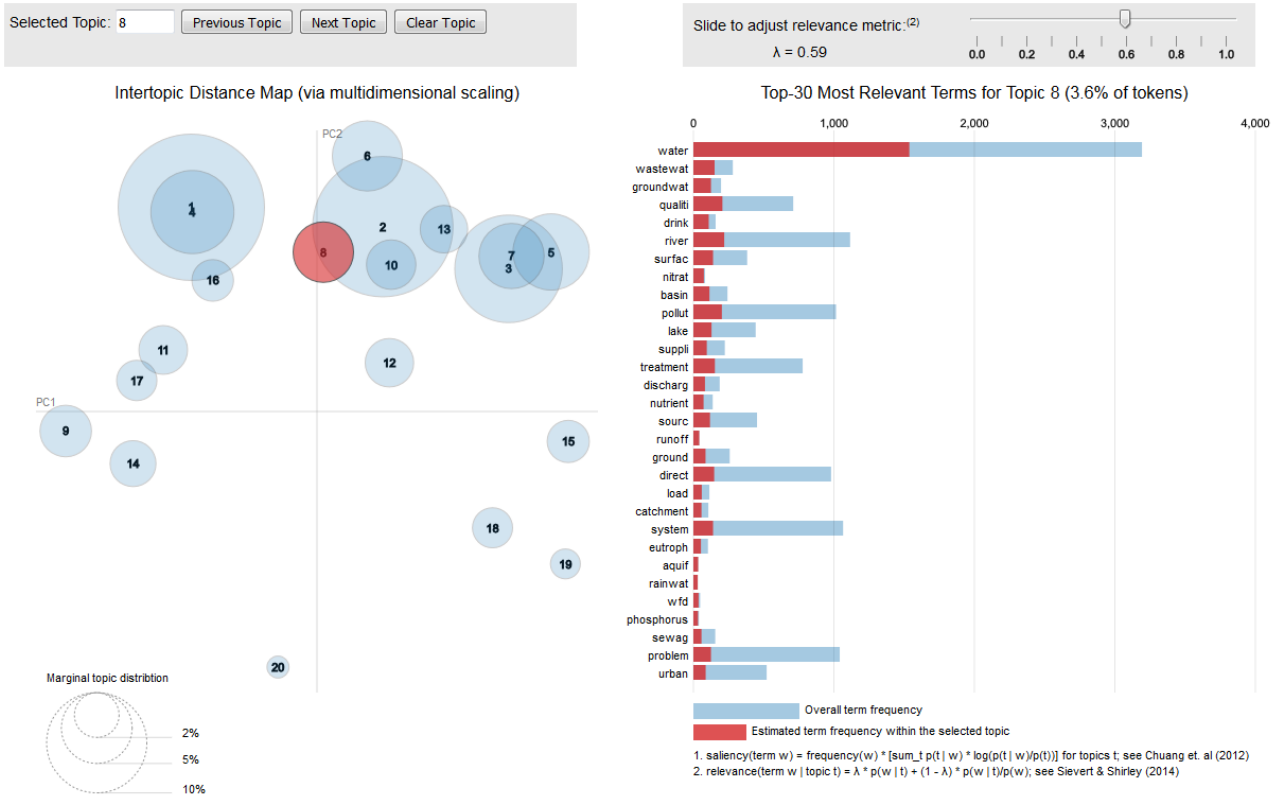


Přiřazení projektů dle četnosti



Metoda pomocí „Topic mining“

- Použili jsme model, který na základě statistické analýzy četnosti výskytu slov hledá v souboru dokumentů témata, o kterých se píše.



ČESKÁ EVALUAČNÍ SPOLEČNOST

Děkujeme za pozornost

Tomáš Schwardy a Petr Krucký

Heřmanova 22, 170 00 Praha 7, e-mail: ces@czecheval.cz, web: www.czecheval.cz