Vektorová reprezentace projektů VaVaI

Pavel Mikuláš

Matematicko-fyzikální fakulta
Univerzita Karlova

Ročníkový projekt, 2019/2020

Co bylo předmětem zkoumání...

- Struktura a rozdělení projektů

Co bylo předmětem zkoumání...

- Struktura a rozdělení projektů
- Minimální popis projektu

Co bylo předmětem zkoumání...

- Struktura a rozdělení projektů
- Minimální popis projektu
- Způsoby vektorové reprezentace projektů

Co bylo předmětem zkoumání...

- Struktura a rozdělení projektů
- Minimální popis projektu
- Způsoby vektorové reprezentace projektů
- Shlukování vektorových reprezentací

Co bylo předmětem zkoumání...

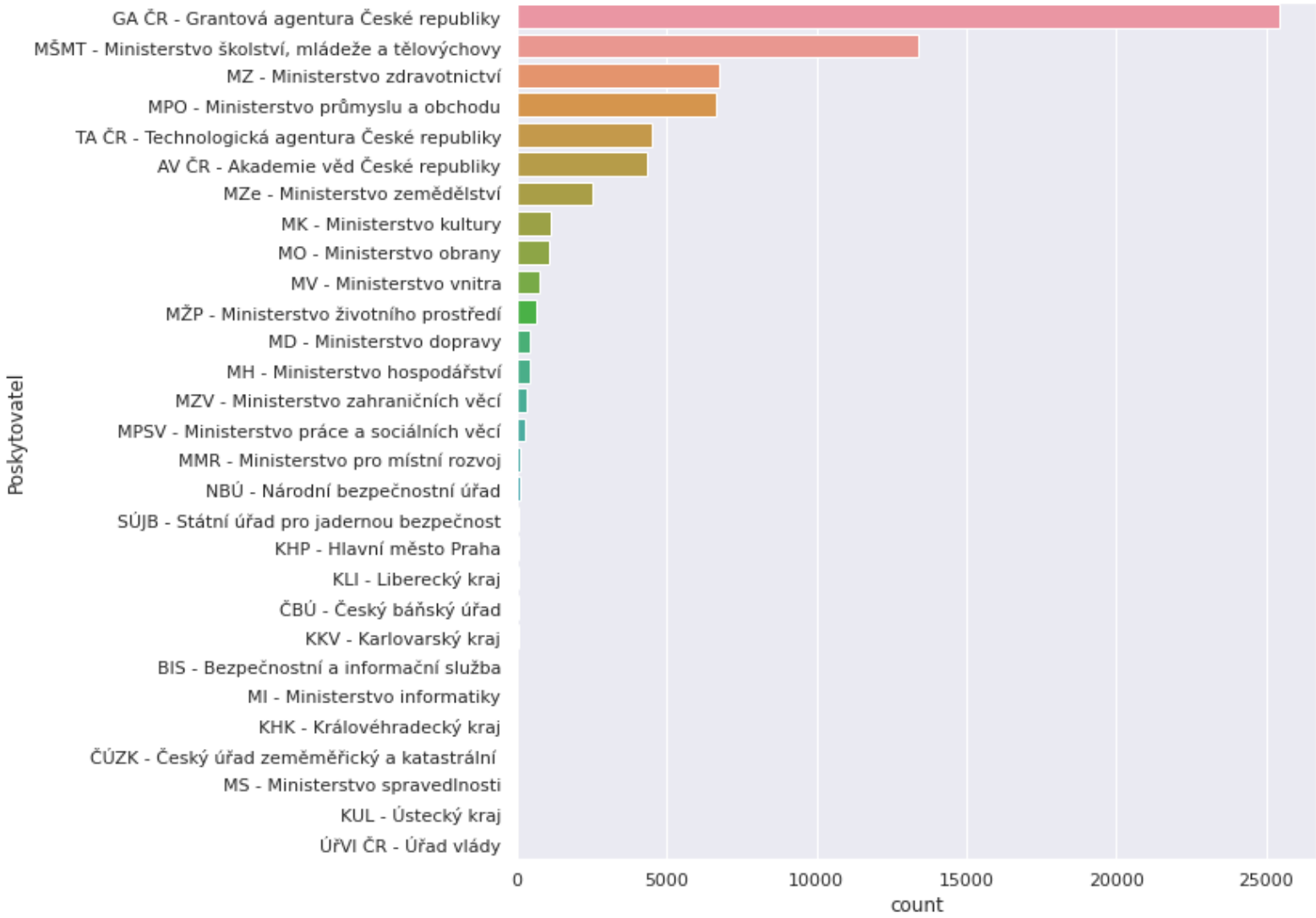
- Struktura a rozdělení projektů
- Minimální popis projektu
- Způsoby vektorové reprezentace projektů
- Shlukování vektorových reprezentací
- Interpretace úvodních pokusů o shlukování

Popis projektu

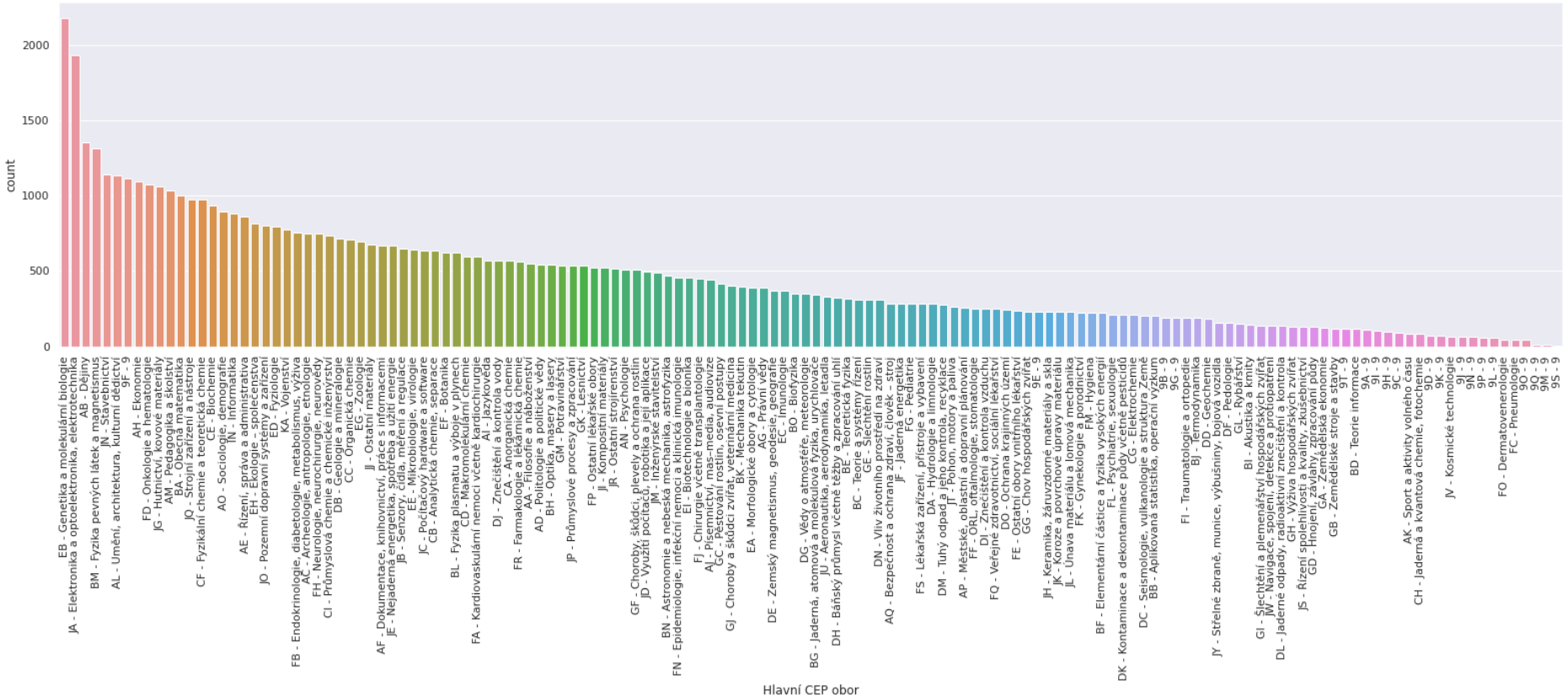
Kód projektu	AACBUAP
Název česky	Dynamika výstupu důlních plynů z podzemních prostor uzavřených dolů
Název anglicky	Dynamics of mine gases emissions from underground spaces of closed down mines
Anotace česky	Cílem projektu je vypracovat metodiku měření, provedení její verifikace praktickým měřením v katastrálním území města Orlová a porovnání s výsledk...
Anotace anglicky	The objective of the project is to develop the methodology of measurement, performing its verification on the basic of practical measurements in c...
Hlavní CEP obor	DH - Báňský průmysl včetně těžby a zpracování uhlí
Hlavní účastníci	Vysoká škola báňská - Technická univerzita Ostrava (IČO: 61989100)
Výčet právních forem účastníků	VVS - Veřejná nebo státní vysoká škola
Výčet krajů účastníků	Moravskoslezský kraj
Výčet zemí účastníků	CZ - Česká republika
Podrobné informace o účastnících	Vysoká škola báňská - Technická univerzita Ostrava (ico: 61989100; forma: VVS - Veřejná nebo státní vysoká škola; adresa: 17. listopadu 2172/15, 7...
Hlavní řešitelé	prof. Ing. Pavel Prokop CSc. (vedidk=8016534)
Klíčová slova	Mines
Výčet druhů dosažených výsledků	1; 48; 3; 49
Poskytovatel	ČBÚ - Český báňský úřad
Program	AA - Zvýšení úrovně bezpečnosti práce v dolech a eliminace nebezpečí od unikajícího metanu z uzavřených důlních prostor
Uznané náklady	2000
Podpora ze SR	2000
Ostatní veřejné zdroje fin.	0
Neveřejné zdroje fin.	0
Začátek řešení	2002
Konec řešení	2003
Relevance	1

Celkový počet projektů: 69357

Počty projektů podle poskytovatelů



Počty projektů podle oborů CEP



Nevyužité atributy

Kód projektu	AACBUAP
Název česky	Dynamika výstupu důlních plynů z podzemních prostor uzavřených dolů
Anotace česky	Cílem projektu je vypracovat metodiku měření, provedení její verifikace praktickým měřením v katastrálním území města Orlová a porovnání s výsledk...
Hlavní účastníci	Vysoká škola báňská - Technická univerzita Ostrava (IČO: 61989100)
Výčet právních forem účastníků	VVS - Veřejná nebo státní vysoká škola
Výčet krajů účastníků	Moravskoslezský kraj
Výčet zemí účastníků	CZ - Česká republika
Podrobné informace o účastnících	Vysoká škola báňská - Technická univerzita Ostrava (ico: 61989100; forma: VVS - Veřejná nebo státní vysoká škola; adresa: 17. listopadu 2172/15, 7...
Hlavní řešitelé	prof. Ing. Pavel Prokop CSc. (vedidk=8016534)
Poskytovatel	ČBÚ - Český báňský úřad
Program	AA - Zvýšení úrovně bezpečnosti práce v dolech a eliminace nebezpečí od unikajícího metanu z uzavřených důlních prostor

Atributy použité k reprezentaci projektů

Název anglicky	Dynamics of mine gases emissions from underground spaces of closed down mines
Anotace anglicky	The objective of the project is to develop the methodology of measurement, performing its verification on the basic of practical measurements in cadastral territory of Orlova Town and comparing the results with measurements of concentration of gases in sThere will be updated the present legal regulations and the methodology for building assessment in areas of mine gases emissions to the surface. The aim is to get scientific based input materials for drafting laws for safety at work and safety at trafficin mines and also for carrying out of supervision of the state mining authority
Klíčová slova	Mines

Reprezentace projektu jedním řetězcem

Zřetězením vybraných textových atributů získáme minimální popis projektu pomocí jednoho retězce.

Reprezentace projektu jedním řetězcem

Zřetězením vybraných textových atributů získáme minimální popis projektu pomocí jednoho retězce.

Řetězec převedeme na malá písmena a odstraníme diakritiku a "stop-words".

Reprezentace projektu jedním řetězcem

Zřetězením vybraných textových atributů získáme minimální popis projektu pomocí jednoho retězce.

Řetězec převedeme na malá písmena a odstraníme diakritiku a "stop-words".

Pro redukci velikosti slovníku tento řetězec následně zlemmatizujeme.

Příklad textu před lemmatizací

text
this sentence is about whales
this sentence is about kangaroos
this sentence is about rhinos
this is an another sentence about whales
yet another sentence about kangaroos

Příklad textu před lemmatizací

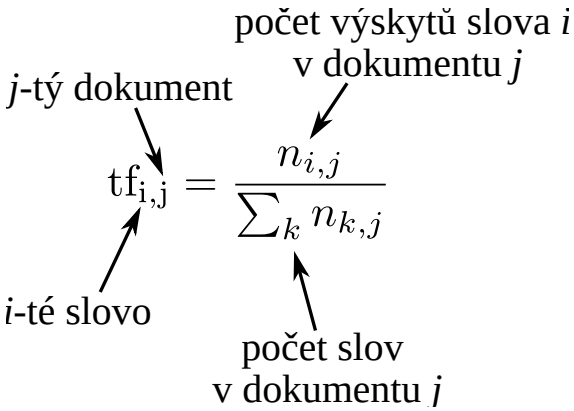
text	text_lemmatized
this sentence is about whales	this sentence is about whale
this sentence is about kangaroos	this sentence is about kangaroo
this sentence is about rhinos	this sentence is about rhino
this is an another sentence about whales	this is an another sentence about whale
yet another sentence about kangaroos	yet another sentence about kangaroo

TF-IDF - definice

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

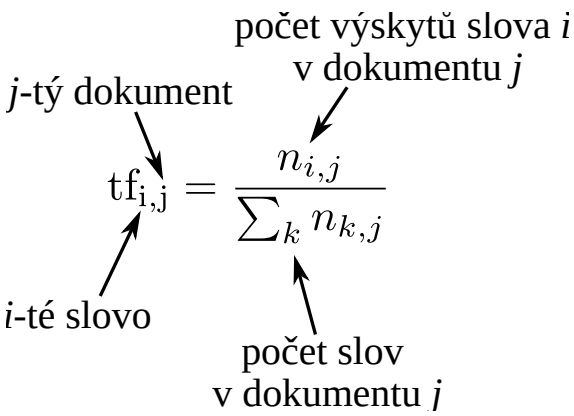
$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

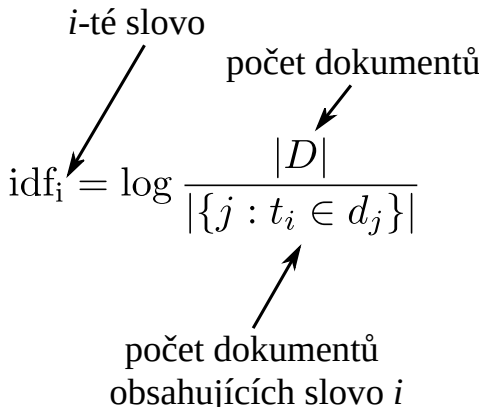
TF-IDF - definice



$$\mathrm{idf}_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

TF-IDF - definice





TF-IDF - shluky před a po odstranění stop-words

	about	an	another	is	kangaroo	rhino	sentence	this	whale	yet
this sentence is about whale	0.361255	0.000000	0.000000	0.427120	0.000000	0.000000	0.361255	0.427120	0.611659	0.000000
this sentence is about kangaroo	0.361255	0.000000	0.000000	0.427120	0.611659	0.000000	0.361255	0.427120	0.000000	0.000000
this sentence is about rhino	0.329691	0.000000	0.000000	0.389801	0.000000	0.691894	0.329691	0.389801	0.000000	0.000000
this is an another sentence about whale	0.258774	0.543066	0.438142	0.305954	0.000000	0.000000	0.258774	0.305954	0.438142	0.000000
yet another sentence about kangaroo	0.287033	0.000000	0.485990	0.000000	0.485990	0.000000	0.287033	0.000000	0.000000	0.602372

TF-IDF - shluky před a po odstranění stop-words

	about	an	another	is	kangaroo	rhino	sentence	this	whale	yet
this sentence is about whale	0.361255	0.000000	0.000000	0.427120	0.000000	0.000000	0.361255	0.427120	0.611659	0.000000
this sentence is about kangaroo	0.361255	0.000000	0.000000	0.427120	0.611659	0.000000	0.361255	0.427120	0.000000	0.000000
this sentence is about rhino	0.329691	0.000000	0.000000	0.389801	0.000000	0.691894	0.329691	0.389801	0.000000	0.000000
this is an another sentence about whale	0.258774	0.543066	0.438142	0.305954	0.000000	0.000000	0.258774	0.305954	0.438142	0.000000
yet another sentence about kangaroo	0.287033	0.000000	0.485990	0.000000	0.485990	0.000000	0.287033	0.000000	0.000000	0.602372

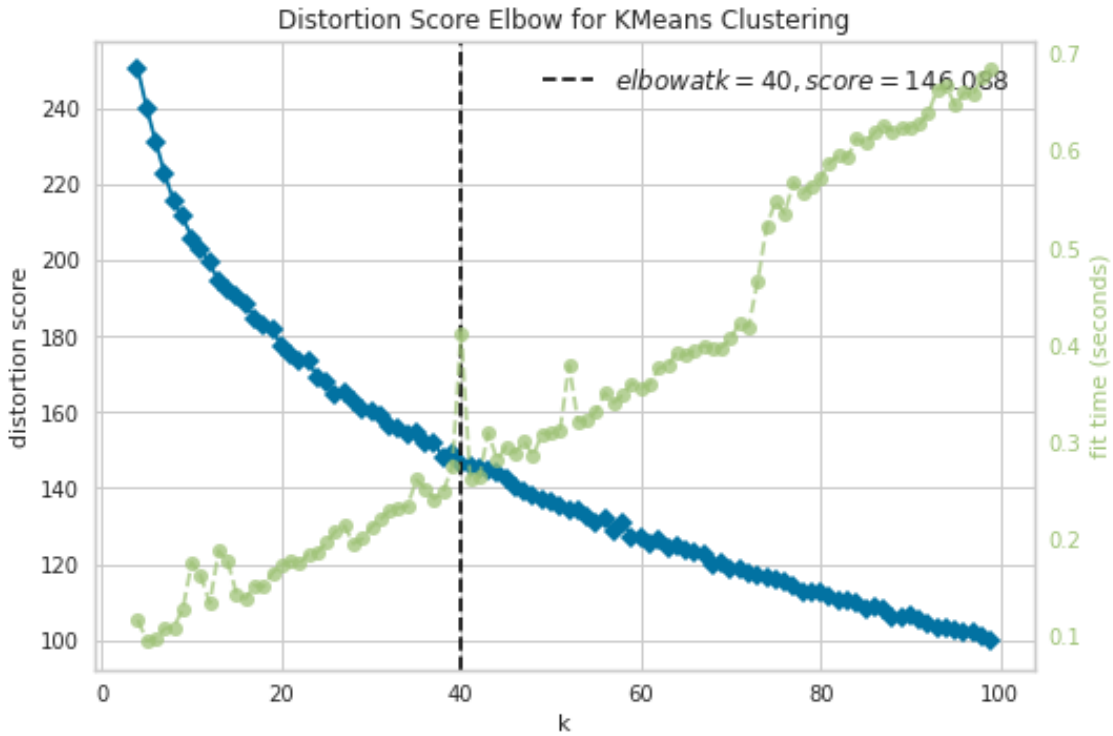
	kangaroo	rhino	sentence	whale
sentence whale	0.000000	0.00000	0.508542	0.861037
sentence kangaroo	0.861037	0.00000	0.508542	0.000000
sentence rhino	0.000000	0.90275	0.430165	0.000000
sentence whale	0.000000	0.00000	0.508542	0.861037
sentence kangaroo	0.861037	0.00000	0.508542	0.000000

TF-IDF - shluky před a po odstranění stop-words

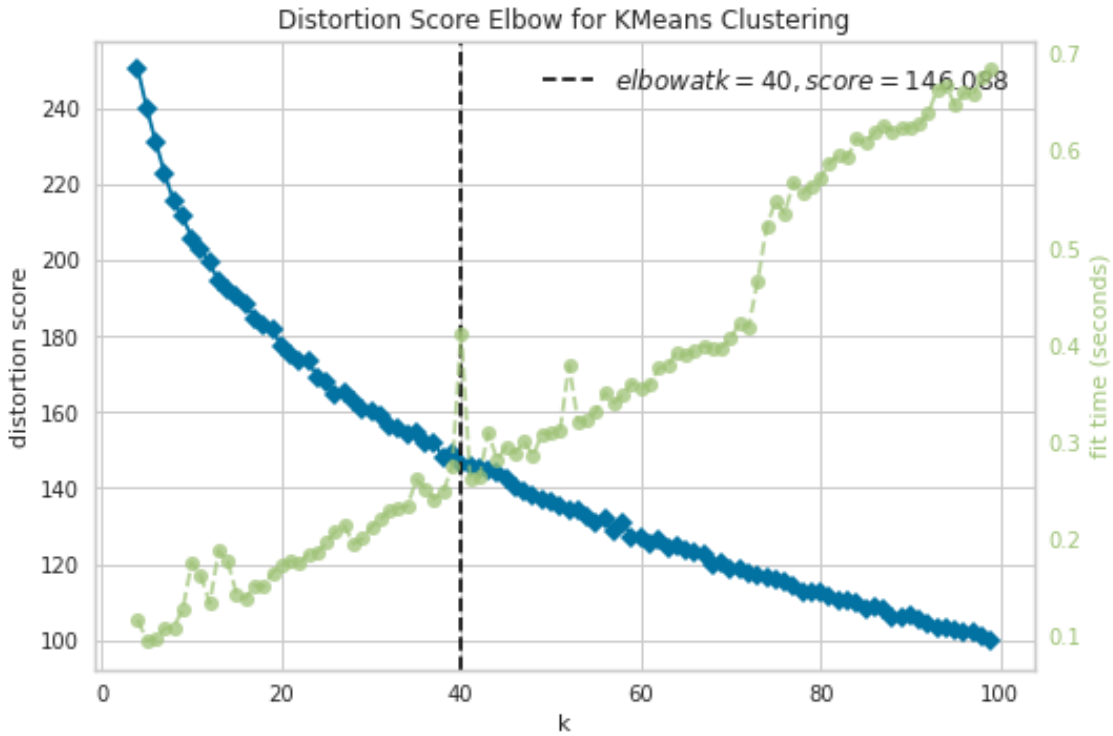
	about	an	another	is	kangaroo	rhino	sentence	this	whale	yet	cluster
this sentence is about whale	0.361255	0.000000	0.000000	0.427120	0.000000	0.000000	0.361255	0.427120	0.611659	0.000000	2
this sentence is about kangaroo	0.361255	0.000000	0.000000	0.427120	0.611659	0.000000	0.361255	0.427120	0.000000	0.000000	1
this sentence is about rhino	0.329691	0.000000	0.000000	0.389801	0.000000	0.691894	0.329691	0.389801	0.000000	0.000000	1
this is an another sentence about whale	0.258774	0.543066	0.438142	0.305954	0.000000	0.000000	0.258774	0.305954	0.438142	0.000000	2
yet another sentence about kangaroo	0.287033	0.000000	0.485990	0.000000	0.485990	0.000000	0.287033	0.000000	0.000000	0.602372	0

	kangaroo	rhino	sentence	whale	cluster
sentence whale	0.000000	0.00000	0.508542	0.861037	2
sentence kangaroo	0.861037	0.00000	0.508542	0.000000	1
sentence rhino	0.000000	0.90275	0.430165	0.000000	0
sentence whale	0.000000	0.00000	0.508542	0.861037	2
sentence kangaroo	0.861037	0.00000	0.508542	0.000000	1

Elbow method

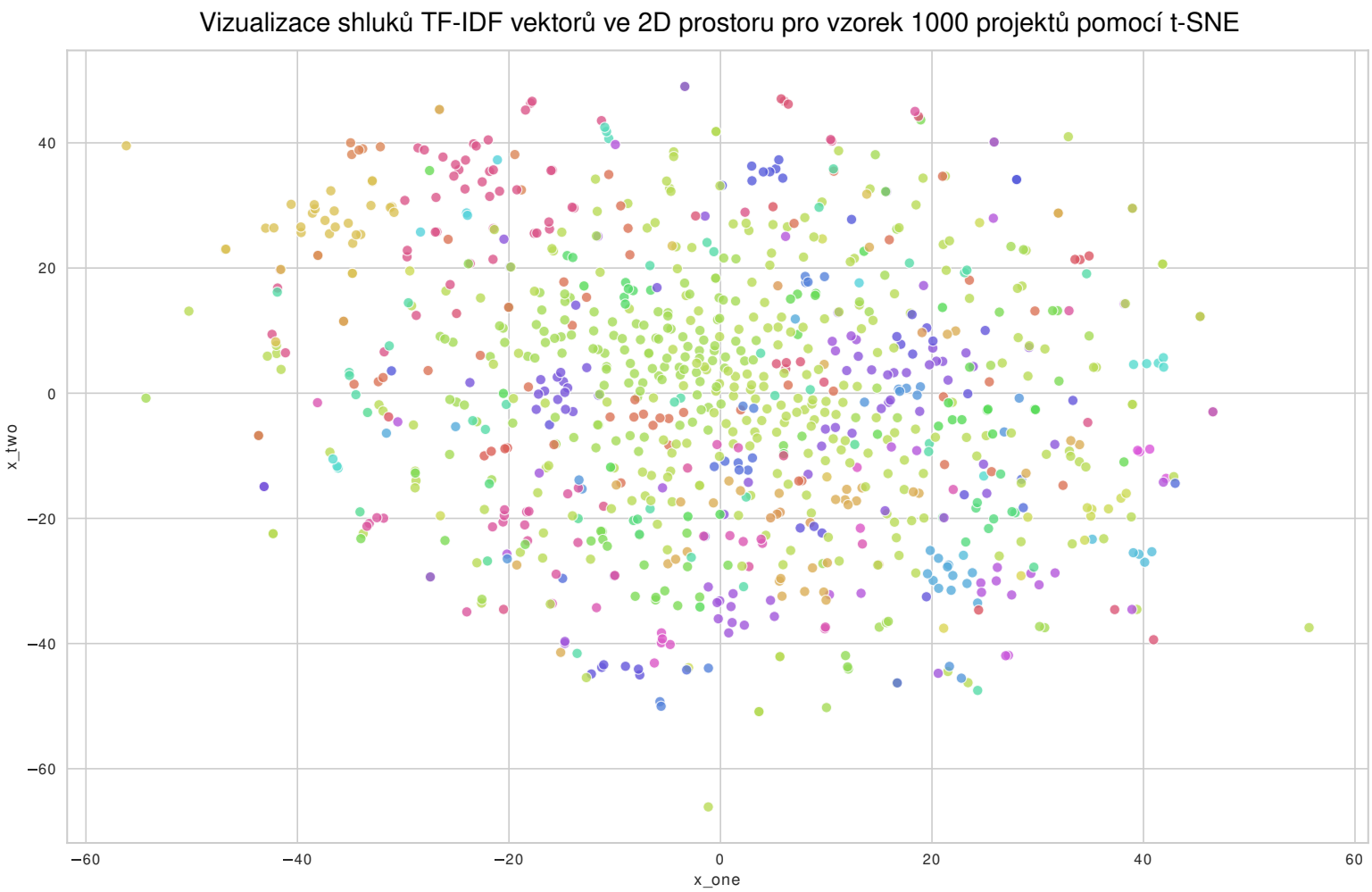


Elbow method



Optimální k je v tomto případě 40, to využijeme pro shlukování TF-IDF vektorů.

t-SNE


Vizualizace shluků TF-IDF vektorů ve 2D prostoru pro vzorek 1000 projektů pomocí t-SNE
x_two
x_one
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60

word2vec

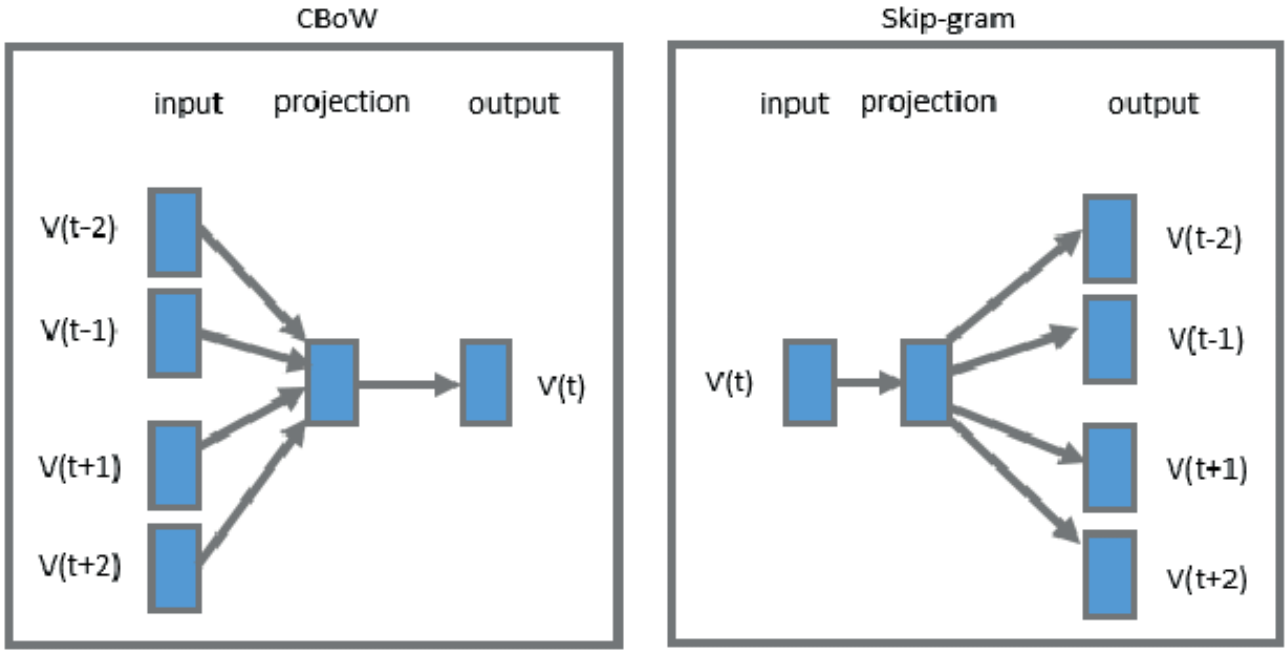
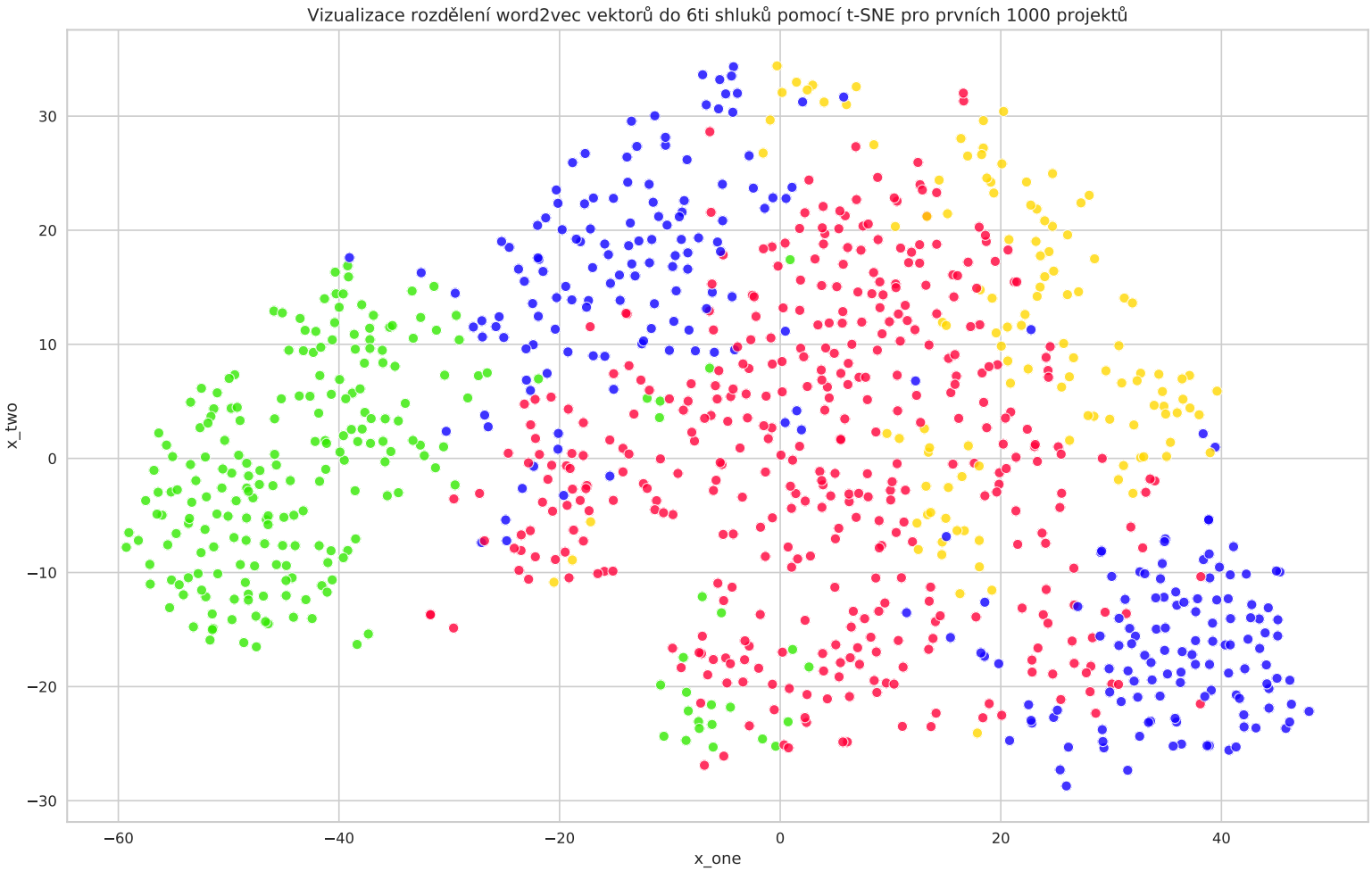


Schéma word2vec modelu (10.1109/UBMK.2017.8093492.)

word2vec


Vizualizace rozdělení word2vec vektorů do 6ti shluků pomocí t-SNE pro prvních 1000 projektů
x_two
x_one
30
20
10
0
−10
−20
−30
−60
−40
−20
0
20
40