

PA153 Počítačové zpracování přirozeného jazyka

10 – dolování témat, názorů, pojmenovaných entit

Karel Pala, Zuzana Nevěřilová

Centrum ZPJ, FI MU, Brno

7. prosince 2017

- 1 Analýza „bez analýzy“
- 2 Rozpoznání témat
- 3 Rozpoznávání pojmenovaných entit
- 4 Dolování názorů

Analýza textu „bez analýzy“

Z textu můžeme získat dost informací bez analýzy obsahu textu (kódování nebo jazyk, délka textu, počet odstavců, počet slov . . .).

Můžeme získat informace o obsahu bez analýzy obsahu?

Analýza textu „bez analýzy“

Z textu můžeme získat dost informací bez analýzy obsahu textu (kódování nebo jazyk, délka textu, počet odstavců, počet slov ...).

Můžeme získat informace o obsahu bez analýzy obsahu?

Ano, ale ...

Analýza textu „bez analýzy“: proč?

Při analýze obsahu textu:

větná segmentace, tokenizace, morfologická desambiguace, rozdělení na fráze, syntaktická analýza, lexikální analýza, lexikální desambiguace, sémantická analýza

Analýza textu „bez analýzy“: proč?

Při analýze obsahu textu:

větná segmentace, tokenizace, morfologická desambiguace, rozdělení na fráze, syntaktická analýza, lexikální analýza, lexikální desambiguace, sémantická analýza

- na každé úrovni vznikají chyby
- na každé úrovni zbyde část jazykových jevů, které nejsou pokryty
- některé části analýzy nejsou příliš rychlé (časová složitost syntaktické analýzy)

Analýza textu „bez analýzy“: na druhou stranu

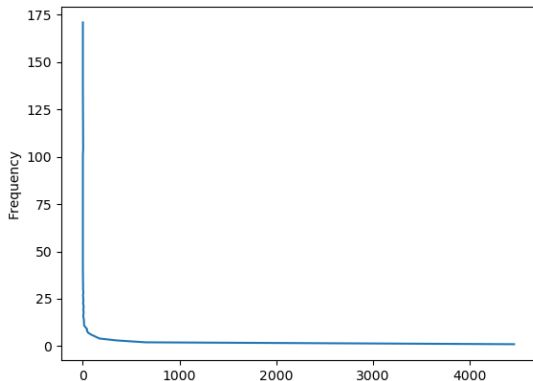
... získáme **některé** informace o obsahu textu s **určitou** přesností, většinou **rychle**.

Analýza textu „bez analýzy“: jak?

- některé části textu jsou důležitější než jiné
- pokud ty důležitější identifikujeme, můžeme dále pracovat jen s nimi

Analýza textu „bez analýzy“: jak?

- některé části textu jsou důležitější než jiné
- pokud ty důležitější identifikujeme, můžeme dále pracovat jen s nimi



Díky Zipfovu rozložení můžeme zjistit, která slova jsou důležitá na základě jejich **frekvencí**.

Rozpoznávání témat (topic recognition)

Čistý zisk energetické společnosti ČEZ za tři čtvrtletí letošního roku meziročně klesl o 4,7 procenta na 31,7 miliardy korun. Tržby se meziročně snížily o 0,3 procenta na 161,9 miliardy korun. Hlavním důvodem poklesu byly odpisy aktiv kvůli regulacím evropského energetického sektoru a související snižování velkoobchodních cen elektřiny, sdělila firma. Výsledek je tak výrazně pod očekáváním. Analytici totiž předpokládali, že čistý zisk ČEZ stoupne o víc než čtyři procenta na 34,8 miliardy korun. Společnost také oznámila, že kvůli snížení velkoobchodních cen elektřiny a regulatorním zásahům do evropského energetického sektoru snížila celoroční výhled čistého zisku na 35 miliard korun. Původně počítala s výsledkem o 2,5 miliardy vyšším. "Očekávané celoroční výsledky hospodaření ČEZ odrážejí současný stav energetiky v Evropě. Fakt, že na naše výsledky tato krize doléhá později a výrazně méně než na naše evropské konkurenty, reflektuje zejména naši úspěšnou strategii předprodejů elektřiny na roky dopředu a důraz na vnitřní úspory," uvedl k výsledkům předseda představenstva a generální ředitel Daniel Beneš.

Rozpoznávání témat (topic recognition)

Čistý zisk energetické společnosti ČEZ za tři čtvrtletí letošního roku meziročně klesl o 4,7 procenta na 31,7 miliardy korun. Tržby se meziročně snížily o 0,3 procenta na 161,9 miliardy korun. Hlavním důvodem poklesu byly odpisy aktiv kvůli regulacím evropského energetického sektoru a související snižování velkoobchodních cen elektřiny, sdělila firma. Výsledek je tak výrazně pod očekáváním. Analytici totiž předpokládali, že čistý zisk ČEZ stoupne o víc než čtyři procenta na 34,8 miliardy korun. Společnost také oznámila, že kvůli snížení velkoobchodních cen elektřiny a regulatorním zásahům do evropského energetického sektoru snížila celoroční výhled čistého zisku na 35 miliard korun. Původně počítala s výsledkem o 2,5 miliardy vyšším. "Očekávané celoroční výsledky hospodaření ČEZ odrážejí současný stav energetiky v Evropě. Fakt, že na naše výsledky tato krize doléhá později a výrazně méně než na naše evropské konkurenty, reflektuje zejména naši úspěšnou strategii předprodejů elektřiny na roky dopředu a důraz na vnitřní úspory," uvedl k výsledkům předseda představenstva a generální ředitel Daniel Beneš.

Rozpoznávání témat (topic recognition)

- extrakce klíčových frází (key phrases)
- klasifikace textu do kategorií (sport, fotbal, finance, půjčky, ekonomie, energetika. . .)

Rozpoznávání témat (topic recognition)

- **extrakce klíčových frází (key phrases)**
- klasifikace textu do kategorií (sport, fotbal, finance, půjčky, ekonomie, energetika. . .)

Extrakce klíčových frází (key phrases) obecně

- podobný úkol jako extrakce klíčových slov, ale více jazykově závislý (potřebujeme informaci o struktuře jmenné fráze)
- klíčové n-gramy (slovo = unigram)
- zkoumaný korpus a referenční korpus
- potřebujeme (předpočítané) frekvence n-gramů
- frekvence n-gramu není srovnatelná s frekvencí m-gramu pro $n \neq m$

Extrakce klíčových frází (key phrases), projekt To|P|icks

- zkoumaný korpus je (krátký) text
- referenční korpus je (velký) korpus
- text rozdělíme na možné fráze (pomocí regulární gramatiky)
- každá fráze získá skóre: frekvence n-gramů v textu / frekvence n-gramů v korpusu
- vyhledáváme základní tvary n-gramů (např. energetický společnost ČEZ)
- skóre fráze posiluje, pokud má podfráze také nějaké skóre
- skóre fráze posiluje, pokud fráze obsahuje pojmenovanou entitu
- skóre fráze oslabuje, pokud je fráze krátká nebo pokud je číslo

Projekt To|P|icks: analýza „bez analýzy“

- pracujeme s tokeny (použili jsme tokenizaci)
 - pracujeme s n-gramy lemmat (použili jsme lemmatizaci)
 - počítáme poměr frekvencí (používáme korpus konkrétního jazyka)
 - extrahujeme kandidáty pomocí regulární gramatiky (používáme parciální syntaktickou analýzu)
 - rozpoznáváme pojmenované entity
-
- neprobíhá úplná analýza
 - nepracujeme s lexikálním významem

Projekt To|P|icks: hodnocení

Čistý zisk energetické společnosti ČEZ za tři čtvrtletí letošního roku meziročně klesl o 4,7 procenta na 31,7 miliardy korun. Tržby se meziročně snížily o 0,3 procenta na 161,9 miliardy korun. Hlavním důvodem poklesu byly odpisy aktiv kvůli regulacím evropského energetického sektoru a související snižování velkoobchodních cen elektřiny, sdělila firma. Výsledek je tak výrazně pod očekáváním. Analytici totiž předpokládali, že čistý zisk ČEZ stoupne o víc než čtyři procenta na 34,8 miliardy korun. Společnost také oznámila, že kvůli snížení velkoobchodních cen elektřiny a regulatorním zásahům do evropského energetického sektoru snížila celoroční výhled čistého zisku na 35 miliard korun. Původně počítala s výsledkem o 2,5 miliardy vyšším. "Očekávané celoroční výsledky hospodaření ČEZ odrážejí současný stav energetiky v Evropě. Fakt, že na naše výsledky tato krize doléhá později a výrazně méně než na naše evropské konkurenty, reflektuje zejména naši úspěšnou strategii předprodeje elektřiny na roky dopředu a důraz na vnitřní úspory," uvedl k výsledkům předseda představenstva a generální ředitel Daniel Beneš.

Extrahuje program „ty správné klíčové fráze“?

⇒ obecnější otázka: dává program správný výstup?

- je třeba stanovit přesně cíl
- je třeba stanovit vzdálenost (nejlépe metriku) mezi výstupem a cílem

Rozpoznávání pojmenovaných entit (named entity recognition)

pojmenovaná entita = jméno osoby, instituce, místa, díla, výrobku, události

- často začíná velkým písmenem
- často se skládá z více slov (multi-word expressions, MWE)
- často obsahuje slova z jiného jazyka
- často obsahuje „neslova“

Rozpoznávání pojmenovaných entit: proč?

Kdo chce vidět Idiota, nechť se dostaví do ředitelny.

(Obecná škola)

Rozpoznávání pojmenovaných entit: proč?

*Kdo chce vidět **Idiota**, nechť se dostaví do ředitelny.*

(Obecná škola)

Rozpoznávání pojmenovaných entit: proč?

*Kdo chce vidět **Idiota**, nechť se dostaví do ředitelny.*

(Obecná škola)

*Četl jsem **Obsluhoval jsem anglického krále** a pak jsem to i viděl.*

Rozpoznávání pojmenovaných entit: proč?

*Kdo chce vidět **Idiota**, nechť se dostaví do ředitelny.*

(Obecná škola)

*Četl jsem **Obsluhoval jsem anglického krále** a pak jsem to i viděl.*

Rozpoznávání pojmenovaných entit: jak?

- seznamy
 - ▶ seznamy jmen, seznamy příjmení (ČSÚ)
 - ▶ seznam obcí (PSČ)
 - ▶ seznam firem (ARES)
 - ▶ seznam uměleckých děl (ČSFD, Databáze knih)
 - ▶ seznam výrobků (Heureka.cz, Seznam zboží)
- zkratky zavedené v textu: operační systém (dále jen OS)
- vzory: (určité) klíčové fráze
 - ▶ Association for ...
 - ▶ Úřad pro ...
 - ▶ jak uvedl ředitel ...
- formátování textu: morfologický analyzátor majka
- syntaktická struktura: morfologický analyzátor/k1gMnSc1
majka/k1gFnSc1

Rozpoznávání pojmenovaných entit: jak?

- seznamy
 - ▶ seznamy jmen, seznamy příjmení (ČSÚ)
 - ▶ seznam obcí (PSČ)
 - ▶ seznam firem (ARES)
 - ▶ seznam uměleckých děl (ČSFD, Databáze knih)
 - ▶ seznam výrobků (Heureka.cz, Seznam zboží)
- zkratky zavedené v textu: operační systém (dále jen OS)
- vzory: (určité) klíčové fráze
 - ▶ Association for ...
 - ▶ Úřad pro ...
 - ▶ jak uvedl ředitel ...
- formátování textu: morfologický analyzátor majka
- syntaktická struktura: morfologický analyzátor/k1gMnSc1
majka/k1gFnSc1

Rozpoznávání pojmenovaných entit: jak?

- seznamy
 - ▶ seznamy jmen, seznamy příjmení (ČSÚ)
 - ▶ seznam obcí (PSČ)
 - ▶ seznam firem (ARES)
 - ▶ seznam uměleckých děl (ČSFD, Databáze knih)
 - ▶ seznam výrobků (Heureka.cz, Seznam zboží)
- zkratky zavedené v textu: operační systém (dále jen OS)
- vzory: (určité) klíčové fráze
 - ▶ Association for ...
 - ▶ Úřad pro ...
 - ▶ jak uvedl ředitel ...
- formátování textu: morfologický analyzátor majka
- syntaktická struktura: morfologický analyzátor/k1gMnSc1
majka/k1gFnSc1

Rozpoznávání pojmenovaných entit: jak?

- seznamy
 - ▶ seznamy jmen, seznamy příjmení (ČSÚ)
 - ▶ seznam obcí (PSČ)
 - ▶ seznam firem (ARES)
 - ▶ seznam uměleckých děl (ČSFD, Databáze knih)
 - ▶ seznam výrobků (Heureka.cz, Seznam zboží)
- zkratky zavedené v textu: operační systém (dále jen OS)
- vzory: (určité) klíčové fráze
 - ▶ Association for ...
 - ▶ Úřad pro ...
 - ▶ jak uvedl ředitel ...
- **formátování textu: morfologický analyzátor majka**
- syntaktická struktura: morfologický analyzátor/k1gMnSc1
majka/k1gFnSc1

Rozpoznávání pojmenovaných entit: jak?

- seznamy
 - ▶ seznamy jmen, seznamy příjmení (ČSÚ)
 - ▶ seznam obcí (PSČ)
 - ▶ seznam firem (ARES)
 - ▶ seznam uměleckých děl (ČSFD, Databáze knih)
 - ▶ seznam výrobků (Heureka.cz, Seznam zboží)
- zkratky zavedené v textu: operační systém (dále jen OS)
- vzory: (určité) klíčové fráze
 - ▶ Association for ...
 - ▶ Úřad pro ...
 - ▶ jak uvedl ředitel ...
- formátování textu: morfologický analyzátor majka
- syntaktická struktura: morfologický analyzátor/k1gMnSc1
majka/k1gFnSc1

Rozpoznávání pojmenovaných entit: projekt CNER

Czech NER:

- seznam jmen a příjmení (ve všech pádech jednotného čísla)
- seznam NE z (české) Wikipedie (někdy i v jiných pádech než nominativu)
- seznam zboží z Heureka.cz
- seznam knih a filmů
- vzory pomocí regulárních výrazů (datum, číslo a jednotky, měna a číslo ...)
- čísla zákonů a paragrafů

Rozpoznávání pojmenovaných entit: problémy

- najít hranice NE (Opera Vladimíra Franze *Válka s mlouky* vzbudila zasloužený ohlas.

Franz Válek



Rozpoznávání pojmenovaných entit: problémy

- najít hranice NE (Opera Vladimíra Franze Válka s mloky vzbudila zasloužený ohlas.
- interpunkce uvnitř NE (Čtyři vraždy stačí, drahoušku)

Rozpoznávání pojmenovaných entit: problémy

- najít hranice NE (Opera Vladimíra Franze Válka s mloky vzbudila zasloužený ohlas.
- interpunkce uvnitř NE (Čtyři vraždy stačí, drahoušku)
- skloňování NE (Mnohé mám dodnes před očima: Erži z Kočičí hry, Runu z Radúze a Mahuleny, Čapkovu Matku, Bontovou z Přísných milenců, Isabelu z Cesty Karla IV. do Francie a zpět, Hejtmanku z Revizora, Matku z Kočky na rozpálené plechové střeše. . .)

Rozpoznávání pojmenovaných entit: problémy

- najít hranice NE (Opera Vladimíra Franze Válka s mloky vzbudila zasloužený ohlas.
- interpunkce uvnitř NE (Čtyři vraždy stačí, drahoušku)
- skloňování NE (Mnohé mám dodnes před očima: Erži z Kočičí hry, Runu z Radúze a Mahuleny, Čapkovu Matku, Bontovou z Přísných milenců, Isabelu z Cesty Karla IV. do Francie a zpět, Hejtmanku z Revizora, Matku z Kočky na rozpálené plechové střeše. . .)
- To je strašidelný román Stephena Kinga. To vím taky.

Rozpoznávání pojmenovaných entit: problémy

- najít hranice NE (Opera Vladimíra Franze Válka s mloky vzbudila zasloužený ohlas.
- interpunkce uvnitř NE (Čtyři vraždy stačí, drahoušku)
- skloňování NE (Mnohé mám dodnes před očima: Erži z Kočičí hry, Runu z Radúze a Mahuleny, Čapkovu Matku, Bontovou z Přísných milenců, Isabelu z Cesty Karla IV. do Francie a zpět, Hejtmanku z Revizora, Matku z Kočky na rozpálené plechové střeše. . .)
- To je strašidelný román Stephena Kinga. To vím taky.
- NE uvnitř NE (Obraz Doriana Graye)

Rozpoznávání pojmenovaných entit: problémy

- najít hranice NE (Opera Vladimíra Franze Válka s mloky vzbudila zasloužený ohlas.
- interpunkce uvnitř NE (Čtyři vraždy stačí, drahoušku)
- skloňování NE (Mnohé mám dodnes před očima: Erži z Kočičí hry, Runu z Radúze a Mahuleny, Čapkovu Matku, Bontovou z Přísných milenců, Isabelu z Cesty Karla IV. do Francie a zpět, Hejtmanku z Revizora, Matku z Kočky na rozpálené plechové střeše. . .)
- To je strašidelný román Stephena Kinga. To vím taky.
- NE uvnitř NE (Obraz Doriana Graye)
- synonyma (Karel Schwarzenberg–Karel Jan Nepomuk Josef Norbert Bedřich Antonín Vratislav Menas kníže ze Schwarzenbergu–Karl Johannes Nepomuk Josef Norbert Friedrich Antonius Wratislaw Mena Fürst zu Schwarzenberg–kníže–Šláfenberg)

Rozpoznávání pojmenovaných entit: problémy

- najít hranice NE (Opera Vladimíra Franze Válka s mloky vzbudila zasloužený ohlas.
- interpunkce uvnitř NE (Čtyři vraždy stačí, drahoušku)
- skloňování NE (Mnohé mám dodnes před očima: Erži z Kočičí hry, Runu z Radúze a Mahuleny, Čapkovu Matku, Bontovou z Přísných milenců, Isabelu z Cesty Karla IV. do Francie a zpět, Hejtmanku z Revizora, Matku z Kočky na rozpálené plechové střeše. . .)
- To je strašidelný román Stephena Kinga. To vím taky.
- NE uvnitř NE (Obraz Doriana Graye)
- synonyma (Karel Schwarzenberg–Karel Jan Nepomuk Josef Norbert Bedřich Antonín Vratislav Menas kníže ze Schwarzenbergu–Karl Johannes Nepomuk Josef Norbert Friedrich Antonius Wratislaw Mena Fürst zu Schwarzenberg–kníže–Šláfenberg)
- homonyma (Queen Elisabeth: osoba, jiná osoba, loď, prezidentská limuzína, hudební skupina)

Rozpoznávání pojmenovaných entit: vyhodnocení

- systém rozpozná NE a skutečně se jedná o NE (Můj oblíbenec Stephen King)
- systém rozpozná NE, ale nejedná se o NE (To vím taky.)
- systém nerozpozná NE a skutečně se nejedná o NE (To jsem celý já.)
- systém nerozpozná NE, ale jedná se o NE (morfologický analyzátor majka)

Rozpoznávání pojmenovaných entit: vyhodnocení

- systém rozpozná NE a skutečně se jedná o NE (Můj oblíbenec Stephen King)
- systém rozpozná NE, ale nejedná se o NE (To vím taky.)
- systém nerozpozná NE a skutečně se nejedná o NE (To jsem celej já.)
- systém nerozpozná NE, ale jedná se o NE (morfologický analyzátor majka)

matice záměn (confusion matrix):

	co určil systém	
správná klasifikace	+	-
+	true positive	false negative
-	false positive	true negative

Rozpoznávání pojmenovaných entit: vyhodnocení

- systém rozpozná NE a skutečně se jedná o NE (Můj oblíbenec Stephen King)
- systém rozpozná NE, ale nejedná se o NE (To vím taky.)
- systém nerozpozná NE a skutečně se nejedná o NE (To jsem celý já.)
- systém nerozpozná NE, ale jedná se o NE (morfologický analyzátor majka)

matice záměn (confusion matrix):

	co určil systém	
správná klasifikace	+	-
+	true positive	false negative
-	false positive	true negative

celková správnost (overall accuracy): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

celková chyba (overall error): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

Rozpoznávání pojmenovaných entit: vyhodnocení

- systém rozpozná NE a skutečně se jedná o NE (Můj oblíbenec Stephen King)
- systém rozpozná NE, ale nejedná se o NE (To vím taky.)
- systém nerozpozná NE a skutečně se nejedná o NE (To jsem celý já.)
- systém nerozpozná NE, ale jedná se o NE (morfologický analyzátor majka)

matice záměn (confusion matrix):

	co určil systém	
správná klasifikace	+	-
+	true positive	false negative
-	false positive	true negative

celková správnost (overall accuracy): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

celková chyba (overall error): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

přesnost (precision): $\frac{TP}{TP+FP}$

pokrytí/úplnost (recall): $\frac{TP}{TP+FN}$

Dolování názorů (opinion mining, sentiment analysis): proč?

Klidně se nazvou Věci veřejné, slíbí vám transparentnost, antikorupci, žádné dinosaury a již zítra si sednou do vlády s největšími dinosaury, sami iniciují zachování akcí na doručitele a uzavřou „veřejnou“ tajnou hradní dohodu. Klidně se nazvou TOP – v překladu tradice – odpovědnost – prosperita a do čela si postaví provařeného politického turistu, nejneodpovědnější osobu v oblasti financí v politice a sedřou z vás zaživa kůži.

- rozlišit fakta a názory
- sledovat mediální obraz (lukrativní téma)

Dolování názorů: jak?

- rozpoznání klíčových frází: politika, odpovědnost, dinosaurus, dohoda
- rozpoznání pojmenovaných entit: Věci veřejné, TOP
- hodnoticí fráze: provařený, nejneodpovědnější, tajný, dinosaurus, persóna, sedřít kůži zaživa

Dolování názorů: hodnoticí fráze

kladná: prima, super, kvalitní, ocenit, vážit si, pomoci, užitečný
záporná: k ničemu, prolhaný, pod'obanec, bastard, Arabáč, sgarb, vlezdobruselista

neutrální, ale v kontextu hodnoticí: (politický) turista, (o člověku) dinosaurus, (o elektronice) šumítka, (o člověku) plevel, (o politickém názoru) rudý, (o Václavu Klausovi) klimatolog

- jak bez analýzy poznat, k čemu se hodnoticí slovo vztahuje?
Ani se nedivím, že tam dali Nokii C3. Vedle bliká reklama a tam se jasně píše, že má dotykový display:-DHolt naše milá redakce:-D
- jak najít názor na některou část objektu (optika je výborná, ale firmware nestojí za nic)
- jak objevit nová hodnoticí slova? (eurohujer)
- jak detekovat sarkasmus (to se vám tedy povedlo)

Závěr: dolování čehokoliv

- většinou docela rychlé
- často poměrně nepřesné
- využívá informací z korpusu
- používá vždy aspoň základní analýzu (tokenizace, slovní druhy, stemming)
- pro jazyky s bohatou flexí je výhodnější použít více analytických nástrojů (extrakce frází, lemmatizace . . .)

Odkazy I



Diatelová, I. (2013 [cit. 2013-11-24]).

Urážlivé, vulgární a rasistické projevy na internetových diskusních fórech [online].

Bakalářská práce, Masarykova univerzita, Filozofická fakulta.



Liu, B. (2004-2012).

Opinion mining, sentiment analysis, and opinion spam detection.

<http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>.