PLIN021 SÉMANTICKÁ ANALÝZA V PRAXI

ZUZANA NEVĚŘILOVÁ

2020/21

VÍCEZNAČNOST VE SLOVNÍKU: GRANULARITA

Kolik významů má slovo **kočka** v **SSJČ**?

2. malá n. středně velká šelma s hustým kožichem; zool. rod Felis
    1. malá kočkovitá šelma, chovaná v domácnostech
    3. samice kočkovité šelmy vůbec
4. ob. kožišina na límci, kolem krku n. ramen
5. kocovina (Haš.)
6. věc připomínající někt. vlastnost kočky
7. druh důtek

POPIS VÝZNAMU VE SLOVNÍKU

Kdy významy oddělit?

- Podle syntaktických kritérií
  Lord v závěti zanechal všechen svůj majetek sirotčinci. Student zanechal studia.
- Podle sémantických kritérií
  živý/neživý, abstraktní/konkrétní, člověk/zvíře, ...
- Podle kroslinguálních kritérií
  ryba/pez,pescado, notebook/zápisník, přenosný počítač

KDY V NLP POTŘEBUJEME VÝZNAMY ROZLIŠIT?

- strojový překlad: Your eyes September.
- inteligentní vyhledávání: evropský los
- ...

LEXIKÁLNÍ DESAMBIGUACE (WORD SENSE DISAMBIGUATION)

- Vyzkoušejte sami na 10 náhodných konkordancích u libovolného slova, které má hodně významů

WSD

- **Vstup**: slovo s více významy v kontextu (podobně jako KWIC)
- **Výstup**: číslo významu ve slovníku

LEXIKÁLNÍ DESAMBIGUACE (WORD SENSE DISAMBIGUATION)

> ... the problem of computationally determining which „sense“ of a word is activated by the use of the word in a particular context.
>
> [Agirre and Edmonds, 2006]

Klasifikační úloha

Jednotlivé významy tvoří třídy.
Podle kontextu se rozhodujeme, do kterých tříd slovo na vstupu patří.
Toto rozhodnutí je konzistentní.

Předpokladem je, že:

Významy jsou diskrétní.
Je jich konečný počet.
Máme k dispozi inventář významů.
Na přiřazení konkrétního užití k významu se shodneme.

LEXIKÁLNÍ DESAMBIGUACE: ZÁKLADNÍ ALGORITMY

- „Historický“ algoritmus (Lesk, 1986)
- Založený na slovníkových definicích (případně příkladech užití)
- Nutná podmínka: strojově čitelné slovníky (Machine Readable Dictionaries)
- Naivní varianta: slovo *w*, jehož okolí sdílí nejvíc slov s definicí (nebo s příklady užití) *i*-tého významu, má význam s_i
  [Kilgarriff and Rosenzweig, 2000]

LESKŮV ALGORITMUS: DESAMBIGUACE SLOVA „KOČKA“

1. malá kočkovitá šelma, chovaná v domácnostech, na venkově zvl. pro hubení myší; kočka domácí (zool.); šedivá, černá, tříbarevná k.; hladká srst kočky; k. mňouká, přede; k. číhá na myš; k. chytá ptáky; angorská k.; být falešný, úlisný jako k.; přen. expr. je to k. falešník; to děvče je k. lichotné, úlisné; [x] jsou na sebe jako pes a k. nenávidí se..., chovají se k sobě nepřátelsky... hrát si s někým jako k. s myší zahrávat si s někým a dávat mu najevo svou převahu a jeho vlastní bezmocnost; ob. je to pro kočku k ničemu
2. malá n. středně velká šelma s hustým kožichem; zool. rod Felis: k. plavá; k. divoká; k. domácí
3. samice kočkovité šelmy vůbec; rysí k.; lví k.; expr. každá kočkovitá šelma vůbec (tygr, levhart aj.)
4. ob. kožišina na límci, kolem krku n. ramen
5. kocovina (Haš.)
6. věc připomínající někt. vlastnost u kočky: bot. velký trs ostřic vystupující z rašeliniště (na blatech); tech. pojízdný vozík jeřábu se zdvihacím ústrojím
7. druh důtek; devítiocasá k.

LESKŮV ALGORITMUS: DESAMBIGUACE SLOVA „KOČKA“

1. malá kočkovitá šelma, chovaná v domácnostech, na venkově zvl. pro hubení myší; kočka domácí (zool.); šedivá, černá, tříbarevná k.; hladká srst kočky; k. mňouká, přede; k. číhá na myš; k. chytá ptáky; angorská k.; být falešný úlisný jako k.; přen. expr. je to k. falešník; chovají se k sobě nepřátelsky... hrát si s n a jeho vlastní bezmocnost; ob. je to pro kc
2. malá n. středně velká šelma s hustým kož
3. samice kočkovité šelmy vůbec; rysí k.; lví k.; expr. každá kočkovitá šelma vůbec (tygr, levhart aj.)
4. ob. kožišina na límci, kolem krku n. ramen
5. kocovina (Haš.)
6. věc připomínající někt. vlastnost u kočky: bot. velký trs ostřic vystupující z rašeliniště (na blatech); tech. pojízdný vozík jeřábu se zdvihacím ústrojím
7. druh důtek; devítiocasá k.

Vstup: Aminokyselina DL-methionin okyseluje moč, čímž chrání močové ústrojí psů i **koček** (důležitá vlastnost zvláště u kastrovaných jedinců).

LESKŮV ALGORITMUS: DESAMBIGUACE SLOVA „KOČKA“

aminokyselina, což, DL-methionin, důležitý, chránit, i, jedinec, kastrovaný, moč, močový, okyselovat, pes, u, ústrojí, vlastnost, zvláště

1. a, angorský, být, černý, číhat, děvče, domácí, domácnost, expresivně, falešník, falešný, hladký, hubení, chovaný, chytat, jako, kočkovitý, lichotný, malý, mňoukat, myš, na, nenávidět, pes, pro, přeneseně, přísť, pták, se, srst, šedivý, šelma, to, tříbarevný, úlisný, v, venkov, zoologicky, zvláště
2. divoký, domácí, Felis, hustý, kožich, malý, nebo, plavý, rod, s, středně, šelma, velký, zoologicky
3. a, expresivně, jiný, každý, kočkovitý, levhart, lví, samice, šelma, rysí, tygr, vůbec
4. kolem, kožišina, krk, límec, na, nebo, obecně, rameno
5. Hašek, kocovina
6. bláto, botanicky, jeřáb, na, některý, ostřice, pojízdný, připomínající, rašeliniště, s, technicky, trs, u, ústrojí, věc, velký, vozík, vlastnost, vystupující, z, zdvihací
7. devítiocasá, druh, důtky

LESKŮV ALGORITMUS: DESAMBIGUACE SLOVA „KOČKA“

aminokyselina, což, DL-methionin, důležitý, chránit, i, jedinec, kastrovaný, moč, močový, okyselovat, pes, u, ústrojí, vlastnost, zvláště

- D_1 = {pes, zvláště}
- D_2 = { }
- D_3 = { }
- D_4 = { }
- D_5 = { }
- D_6 = {u, ústrojí, vlastnost}
- D_7 = { }

LEXIKÁLNÍ DESAMBIGUACE: ZÁKLADNÍ ALGORITMY

- „Historický“ algoritmus (Lesk, 1986)
- Založený na slovníkových definicích (případně příkladech užití)
- Nutná podmínka: strojově čitelné slovníky (Machine Readable Dictionaries)
- Naivní varianta: slovo w, jehož okolí sdílí nejvíc slov s definicí (nebo s příklady užití) i-tého významu, má význam s_i
  [Kilgarriff and Rosenzweig, 2000]
- Jednoduchá varianta: slovo w, jehož okolí sdílí slova s definicí (nebo s příklady užití) i-tého významu a součet vah těchto slov je nejvyšší, má význam s_i

INVERZNÍ ČETNOST V DOKUMENTU [MANNING ET AL., 2008]

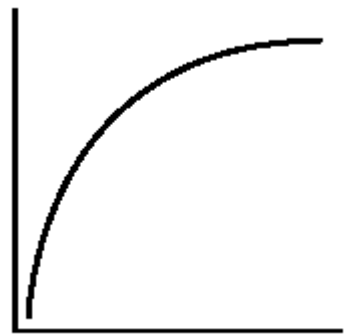
Term frequency tf – četnost znaku t v určitém dokumentu

Počet dokumentů N

Document frequency df_t – počet dokumentů, ve kterých se vyskytuje t

Inverse document frequency

$$idf_t = log \frac{N}{df_t}$$



-

Příklad: mějme dokumenty:

- D1=Máma mele maso
- D2=Ema maso solí, z masa bude oběd
- D3=Máma má Emu
- D4=Ema má mámu i oběd

N=4

tf (maso,D2)=$\frac{2}{7}$

df_t(maso)=2

idf_t(maso)= $log \frac{4}{2}$

tf (Ema,D2)=$\frac{1}{7}$

df_t(Ema)=3

idf_t(Ema)= $log \frac{4}{3}$

LESKŮV ALGORITMUS: DESAMBIGUACE SLOVA „KOČKA“

pro každý význam si slova w:

nastav váhu v na 0: $v(s_i) := 0$

najdi množinu slov O v okolí slova w

pro každé slovo o_j z okolí O

- pro každý význam s_i

pokud se o_j nachází v definici nebo užití D_i

přičti jeho váhu: $v(s_i) := v(s_i) + v(o)$

Vyber s_i s nejvyšší váhou: return $\max(v(s_i))$

(váha slova $v(o) = idf_o$)

LESKŮV ALGORITMUS: DESAMBIGUACE SLOVA „KOČKA“

1. malá kočkovitá šelma, chovaná v domácnostech, na venkově zvl. pro hubení myší; kočka domácí (zool.); šedivá, černá, tříbarevná k.; hladká srst kočky; k. mňouká, přede; k. číhá na myš; k. chytá ptáky; angorská k.; být falešný úlisný jako k.; přen. expr. je to k. falešník; chovají se k sobě nepřátelsky... hrát si s n a jeho vlastní bezmocnost; ob. je to pro kc
2. malá n. středně velká šelma s hustým kož
3. samice kočkovité šelmy vůbec; rysí k.; lví k.; expr. každá kočkovitá šelma vůbec (tygr, levhart aj.)
4. ob. kožišina na límci, kolem krku n. ramen
5. kocovina (Haš.)
6. věc připomínající někt. vlastnost u kočky: bot. velký trs ostřic vystupující z rašeliniště (na blatech); tech. pojízdný vozík jeřábu se zdvihacím ústrojím
7. druh důtek; devítiocasá k.

Vstup: Aminokyselina DL-methionin okyseluje moč, čímž chrání močové ústrojí psů i **koček** (důležitá vlastnost zvláště u kastrovaných jedinců).

LESKŮV ALGORITMUS: DESAMBIGUACE SLOVA „KOČKA“

aminokyselina, což, DL-methionin, důležitý, chránit, i, jedinec, kastrovaný, moč, močový, okyselovat, pes, u, ústrojí, vlastnost, zvláště

1. a, angorský, být, černý, číhat, děvče, domácí, domácnost, expresivně, falešník, falešný, hladký, hubení, chovaný, chytat, jako, kočkovitý, lichotný, malý, mňoukat, myš, na, nenávidět, pes, pro, přeneseně, přist, pták, se, srst, šedivý, šelma, to, tříbarevný, úlisný, v, venkov, zoologicky, zvláště
2. divoký, domácí, Felis, hustý, kožich, malý, nebo, plavý, rod, s, středně, šelma, velký, zoologicky
3. a, expresivně, jiný, každý, kočkovitý, levhart, lví, samice, šelma, rysí, tygr, vůbec
4. kolem, kožišina, krk, límec, na, nebo, obecně, rameno
5. Hašek, kocovina
6. bláto, botanicky, jeřáb, na, některý, ostřice, pojízdný, připomínající, rašeliniště, s, technicky, trs, u, ústrojí, věc, velký, vozík, vlastnost, vystupující, z, zdvihací
7. devítiocasá, druh, důtky

LESKŮV ALGORITMUS: DESAMBIGUACE SLOVA „KOČKA“

aminokyselina, což, DL-methionin, důležitý, chránit, i, jedinec, kastrovaný, moč, močový, okyselovat, pes, u, ústrojí, vlastnost, zvláště

- **D_1 = {2,53 (pes), 1,83 (zvláště)}**
- **D_2 = { }**
- **D_3 = { }**
- **D_4 = { }**
- **D_5 = { }**
- **D_6 = {0,36 (u), 2,32 (ústrojí), 1,29 (vlastnost)}**
- **D_7 = { }**

- $v(s_1) = 4{,}36$
- $v(s_6) = 3{,}97$

LESKŮV ALGORITMUS

SHRNUTÍ

- V definici nutně nemusejí být slova, která se vůbec kdy s hledaným slovem vyskytují.
- Úspěch algorimu silně závisí na použitém slovníku.
- Slovníky nebyly napsány s cílem být zdrojem pro algoritmus WSD.
- Naštěstí na scénu vstoupilo **strojové učení**.

Paper

Paper, matted or felted sheet, usually made of cellulose fibres, formed on a wire screen from water suspension. A brief treatment of paper follows. For full treatment, see papermaking. Paper has been traced to China in about ad 105. It reached Central Asia by 751 and Baghdad by 793, and by the 14th...

LITERATURA

- Agirre, E. and Edmonds, P. (2006). ***Word sense disambiguation: algorithms and applications.*** Text, speech, and language technology. Springer.
- Kilgarriff A. and Rosenzweig, J. (2000). ***English senseval: Report and results.*** In Proceedings of the 2nd International Conference on Language Resources and Evaluation, pages 1239-1244.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). ***Introduction to Information Retrieval.*** Cambridge University Press, New York, NY, USA.