Jazykové modely a textové korpusy

Pavel Rychlý, Aleš Horák

E-mail: hales@fi.muni.cz
http://nlp.fi.muni.cz/nlp_intro/

Obsah:

- Jazykové modely
- Co to je korpus?
- Anglické a národní korpusy
- Formáty korpusů
- Korpusové manažery

n-gramy

Úkol:

Je zadáno n slov textu, jaké slovo následuje s největší pravděpodobností?

např. diktování:

Nově označené { *láhve* / *láhvové* } *se dostanou na trh ...*

n-gramy – pokrač.

Obecně – máme text jako řetězec slov $W = w_1 w_2 w_3 \dots w_n$
Na vstupu zatím $w_1 w_2 \dots w_{i-1}$, chceme určit nejpravděpodobnější w_i

Možnosti:

- použijeme pravděpodobnost $P(w_i)$ – vypočítáme unigramy
  ty ale neberou v úvahu předchozí kontext
- nejlepší – pravděpodobnost podle celého předchozího vstupu

$$P(w_i|w_1 w_2 \dots w_{i-1})$$

n-gramy:

$$P(w_i|w_1 w_2 \dots w_{i-1}) = \frac{P(w_1 \dots w_i)}{P(w_1 \dots w_{i-1})}$$

$$P(w_1 \dots w_i) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1 w_2) \cdot \dots \cdot P(w_i|w_1 \dots w_{i-1})$$

Markovovy modely

problém – potřebujeme *n*-gramy pro velké *n*
řešení – Markovův předpoklad o lokálním kontextu (řádu *n*)

Nejbližší kontext (n slov) nejvíce ovlivňuje pravděpodobnost slova w_i

Pro $n = 1$:

$$P(w_1 \dots w_i) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_2) \cdot \dots \cdot P(w_i|w_{i-1})$$
$$P(w_i|w_1 w_2 \dots w_{i-1}) = P(w_i|w_{i-1})$$
$$P(w_i|w_{i-1}) = \frac{\text{počet}(w_{i-1} w_i)}{\text{počet}(w_{i-1})} \dots \text{bigramy!}$$

(skrytý) Markovův model (*hidden Markov model, HMM*) –
pravděpodobnostní konečný automat pro všechna slova a sekvence

Markovovy modely – využití

Využití jazykových modelů:

- rozpoznávání řeči
- určování morfologických a syntaktických kategorií
- strojový překlad
- určování vztahů mezi slovy
- filtrování generovaných textů

Tvorba jazykových modelů – z textových korpusů

kvalitní model potřebuje (velmi) velké korpusy

Proč velmi velké korpusy

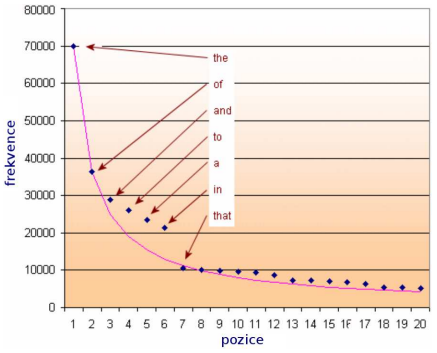
Zipfův zákon (zákon mocniny) distribuce jazyka

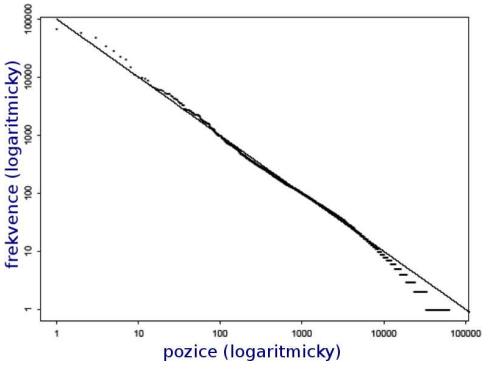
$$\text{frekvence} \cdot \text{pozice} = \textit{konstanta}$$

tedy

$$\text{pozice} = \textit{konst}/\text{frekv}$$

$$\log(\text{pozice}) = \log(\textit{konst}) - \log(\text{frekv})$$





Proč velmi velké korpusy

např. British National Corpus (BNC) – cca 100 mil.slov, 774 tis. různých slov

různá slova podle frekvence:

400,000 ×	freq = 1
374,000 ×	freq ≥ 2
273,000 ×	freq ≥ 3
130,000 ×	freq ≥ 10
88,000 ×	freq ≥ 20
53,000 ×	freq ≥ 50
35,000 ×	freq ≥ 100
12,400 ×	freq ≥ 500
7,600 ×	freq ≥ 1,000
1,000 ×	freq ≥ 10,000

podstatné jméno "test":

- frekvence 15789, pozice 918
- relace object-of: *pass*, *undergo*, *satisfy*, *fail*, *devise*, *conduct*, *administer*, *perform*, *apply*, *boycott*
- relace modifier: *blood*, *driving*, *fitness*, *beta*, *nuclear*, *pregnancy*

Proč velmi velké korpusy

slovní spojení podstatného jména "test":

- "blood test"
  - v BNC, 204 výskytů, relace object-of: *order* (3), *take* (12)
  - v enClueWeb (70 mld.slov), 205220 výskytů, relace object-of: *order* (2323), *undergo* (808), *administer* (456), *perform* (2783), *screen* (129), *request* (442), *conduct* (860), *refuse* (195), *repeat* (254), *scan* (203), *require* (2345), *recommend* (502), *schedule* (192), *run* (1721), *take* (5673), *interpret* (102), *arrange* (162)
- "pregnancy test"
  - v BNC, 26 výskytů, žádná významná slovní spojení
  - v enClueWeb, 54103 výskytů, relace object-of: *take* (7953), *administer* (134), *buy* (1094), *undergo* (145), *perform* (560)

Co to je korpus?

Korpus – skupina dokumentů
Různé typy korpusů:

- textové
- mluvené

Textový korpus:

- soubor textů
- charakteristiky
  - rozsáhlý (stovky milionů až desítky miliard pozic/slov)
  - v jednotném formátu
  - stukturovaný
  - v elektronické podobě

Typy korpusů

- vždy záleží na účelu a způsobu použití
- možnosti dělení korpusů podle
  - jazyk
  - typy textů
  - zdroj dat
  - značkování
  - ...

První korpus

Brown

- americká angličtina (1961)
- Brown University, 1964
- gramatické značkování, 1979
- 500 textů (à ≈2000 slov), 1 mil. slov
- W. N. Francis & H. Kučera
  - první statistické charakteristiky anglických slov
  - relativní četnosti slov a slovních druhů

BNC

British National Corpus

- britská angličtina, 10 % mluva
- první velký korpus pro lexikografy
- vydavatelé slovníků (OUP) + univerzity
- 1. verze: 1991–1994, 2. verze: World Edition 2000
- ≈3000 dokumentů, 100 mil. slov
- gramatické značkování automatickým nástrojem

BoE

Bank of English

- britská angličtina
- COBUILD (HarperCollins), University of Birmingham
- 1991, dále rozšiřován
- 2002, ≈450 mil. slov

Další národní korpusy

- Český národní korpus
  - ÚČNK, FF UK
  - SYN2000, SYN2005, ..., SYN2020 à 100 mil. slov
  - SYN – 5 mld. slov
  - autorské korpusy – Čapek, Havlíček, ...
  - mluvené korpusy – ORAL, BMK, DIALEKT, ...
  - diachronní (historický) DIAKORP
  - paralelní a porovnatelné korpusy – InterCorp, Aranea
- Slovenský, Maďarský, Chorvatský, ...
- Americký

Korpusy na FI

vytvořené na FI, příklady:

- Desam
  - 1996, ručně značkovaný (desambiguovaný)
  - ≈1 mil. slov
- Czes
  - periodika z webu, z let 1996–1998, další el. zdroje, webové zdroje (crawl)
  - ≈465 mil.
- *TenTen
  - různé jazyky, ve spolupráci s LCL, UK
  - 1–20 mld. pozic
- Chyby
  - práce studentů předmětu Základy odb. stylu s vyznačenými chybami
  - ≈400 tis.

Korpusy na FI

spolupráce

- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- Francouzský, Slovinský, Britská angličtina, ...

Formáty korpusů

1. archiv/kolekce
   - různé formáty, podle zdroje/typu
2. textové banky
   - jednotný formát a základní struktura
   - dokumenty/texty, základní metainformace
3. vertikální text
4. binární data v aplikaci
   - pomocná data pro rychlejší zpracování
     - indexy
     - statistiky

Kódování metainformací

- escape-sekvence
  - speciální znak mění význam následujících znaků
  - `\n, \t, &, <tag>`
- SGML
  - Standard Generalised Markup Language
  - ISO 8879:1986(E)
- XML
  - Extensible Markup Language
  - W3C, 1998

XML

- struktura popsána v DTD/XML Schema
- elementy
  - počáteční, koncová značka
  - `<doc>, <head>, </head>, <g/>`
- atributy elementů/značek
  - `<doc title="Jak pejsek ..." author="Čapek">`
  - `<head type="main">`
- entity
  - `&gt;, &lt;, &amp;, &eacute;`

Standardy pro ukládání textů

- SGML/XML
- TEI
  - Text Encoding Initiative (1994)
  - TEI Guidelines for Electronic Text Encoding and Interchange
- CES, XCES
  - Corpus Encoding Standard

Obsah korpusu

Co je v korpusu uloženo?

- text
- metainformace (většinout atributy <doc>)
- struktura dokumentu
  - odstavce, nadpisy, verše, věty
- značkování
  - informace o slovech/pozicích
  - morfologie, základní tvary, syntaktické vazby, ...

Tokenizace

Rozdělení textu do pozic

- může silně ovlivnit výsledky dotazování, četnosti i značkování
- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
  - bude-li, don't – 4 možnosti:
    1. |don't|
    2. |don| |'t|
    3. |don| |'| |t|
    4. |do| |n't| – v BNC
  - zkratky (s tečkami?)
  - datumy
  - desetinná čísla, ...

Vertikální text

- jednoduchý formát i jeho zpracování
  - každý token na samostatném řádku (⇒ udává tokenizaci)
  - struktury formou XML značek
  - značkování odděleno tabulátorem (různé atributy k dané pozici)

```
<doc n=2 id="CMP/94/10">
<head p="80%">
Úpadku          úpadek      k1gInSc3
zabránili       zabránit    k5mAgMnPaP
výkonem         výkon       k1gInSc7
</head>
<p>
<s p="90%">
Po              po          k7c6
několika        několik     k4gFnPc6
akcích          akce        k1gFnPc6
```

- podrobnosti na `nlp.fi.muni.cz/cs/PopisVertikalu`

Zpracování textů na UNIXu

- coreutils
  - cat, head, tail, wc, sort, uniq, comm
  - cut, paste, join, tr
- grep
- awk
- sed / perl

Příklady použití coreutils

- slovník z vertikálního textu

```
cut -f 1 -s desam.vert |sort |uniq -c \
|sort -rn >desam.dict
```

- jednoduchá tokenizace

```
tr -cs 'a-zA-Z0-9' '\n' <GPL >GPL.vert
cat GPL.vert |sort |uniq -c |sort -rn >GPL.dict
```

- všechny bigramy

```
tail -n +2 GPL.vert |paste GPL.vert - |sort |uniq -c
|sort -rn
```


Nástroje pro tvorbu velkých korpusů

projekt `corpus.tools`

samostatné nástroje pro dávkové úkoly zpracování textů, např.:

- JusText – inteligentní extrakce textu z webové stránky
- Spiderling – procházení a stahování textů z webu pro daný jazyk
- Unitok – konfigurovatelný tokenizátor pro více jazyků
- Onion – odstraňuje duplicitní texty
- Chared – detekce kódování textu

Korpusové manažery

nástroje na zpracování korpusů

- uložení textu
- editace/příprava textu
- značkování
- rozdělení do pozic (tokenizace)
- vyhledávání (konkordance)
- statistiky

Systém Manatee

- korpusový manažer
- pro Masarykovu univerzitu dostupný na `ske.fi.muni.cz`
- přímo podporuje
  - uložení textu
  - vyhledávání (konkordance)
  - statistiky
- externí nástroje
  - značkování
  - rozdělení do pozic

Systém Manatee

hlavní zaměření

- velké korpusy
- rozsáhlé značkování
  - morfologické, syntaktické, metainformace
- návaznost na další aplikace/nástroje
  - korpusový editor (CED), tvorba slovníků
- univerzálnost
  - různé jazyky, kódování, systémy značek

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
  - grafické uživatelské rozhraní (Bonito)
  - aplikační programové rozhraní (API)
  - příkazový řádek
- rozsáhlá data
  - stovky mld. pozic
  - neomezeně atributů a metainformací
- rychlost
  - vyhledávání, statistiky

Klíčové vlastnosti

- multihodnoty
  - zpracování víceznačných značkování
- dynamické atributy
  - vyhledávání a statistiky na počítaných datech
- subkorpusy, paralelní korpusy
- silný dotazovací jazyk
  - dotazy na všechny atributy, metainformace
  - pozitivní/negativní filtry
  - regulární výrazy + booleovské operátory

Klíčové vlastnosti

- frekvenční distribuce
  - víceúrovňová
  - všechny atributy a metainformace
- kolokace
  - různé statistické funkce