

# **Natural Language Toolkit**

**prezentace do předmětu PA154**

**Nástroje pro korpusy**

**část 1 – možnosti NLTK**

# Stručná charakteristika

- NLTK je sada knihoven pro Python a programů pro symbolické a statistické zpracování přirozeného jazyka
- k dispozici jsou
  - zdrojové kódy
  - dokumentace
  - tutoriály

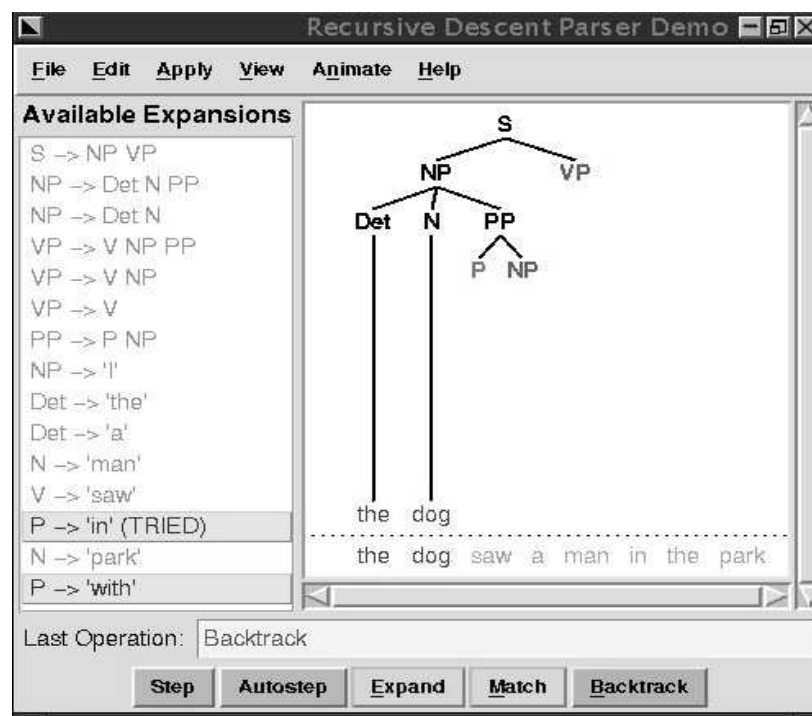
# NLTK je určeno

- pro studenty zpracování přirozeného jazyka
- pro podporu výzkumu souvisejících oblastí, například:
  - empirická lingvistika (korpora)
  - kognitivní vědy
  - umělá inteligence, strojové učení
  - vyhledávání znalostí

# Motivační příklad 1

## nltk\_lite.draw.rdparser

- Ukázka rekurzivní sestupné analýzy shora dolů



# Motivační příklad 2

## nlk\_lite.draw.srparser

- Ukázka posuvně-redukční analýzy zdola nahoru

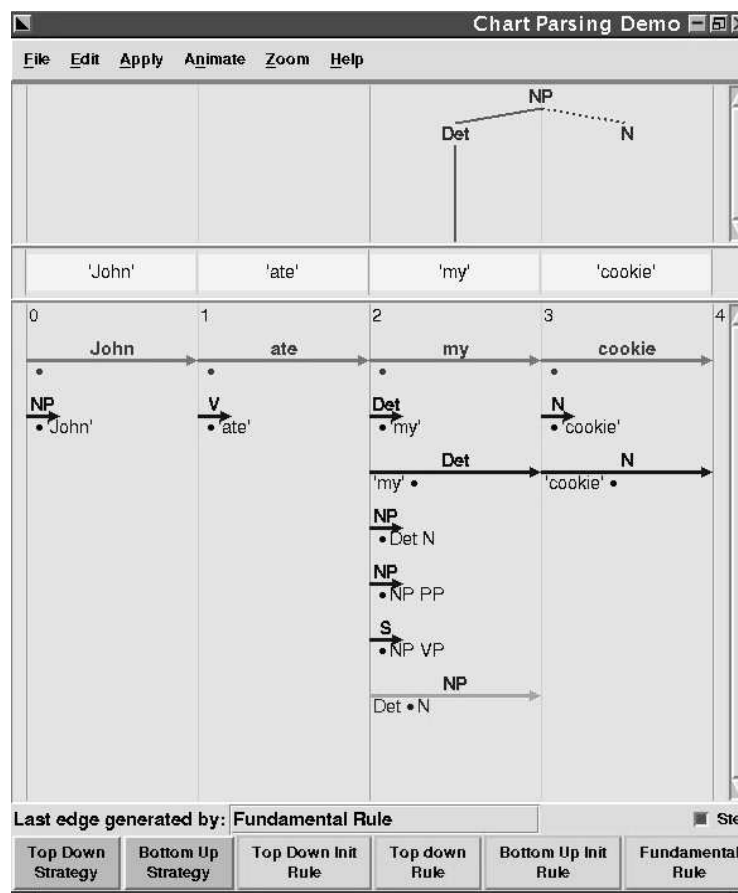
The screenshot shows the 'Shift Reduce Parser Demo' window. It features a menu bar with 'File', 'Edit', 'Apply', 'View', 'Animate', and 'Help'. The main area is divided into three sections: 'Available Reductions', 'Stack', and 'Remaining Text'. The 'Available Reductions' list includes various grammar rules, with 'N -> 'man'' selected. The 'Stack' section displays a partial parse tree for the words 'my dog saw a man', showing 'NP' as the root, with children 'Det' (my) and 'N' (dog) under 'NP', and 'V' (saw) and 'Det' (a) under 'NP'. The 'Remaining Text' section shows 'in the park with a statue'. At the bottom, a 'Last Operation' field displays 'Reduce: N -> 'man'', and there are buttons for 'Step', 'Shift', 'Reduce', and 'Undo'.

# Motivační příklad 3

## nlk\_lite.draw.chart

- Ukázka

tabulková  
analýza



# Motivační příklad 3

## `nltk_lite.draw.chart`

- Analýza zdola nahoru může najít jen jedno vyhodnocení, někdy nenalezne existující řešení
- Analýza shora dolů může být značně neefektivní (pro LR gramatiky může cyklit)
- Řešíme znovuužitím výpočtů (dynamické programování) -> chart parsing

# NLTK - lite

- vývoj: červen – prosinec 2005
- od prosince 2005 je to jediná podporovaná verze, proto se budeme zabývat právě jí
- stejná funkčnost jako klasické NLTK, avšak s nižšími nároky na programátora (používá standardní objekty Pythonu, atd.)



# Autoři a licence

- autoři: Steven Bird, Edward Loper
- mnoho přispěvatelů
- licence:
  - projekt je open source bez záruky
    - GNU General Public License
  - dokumentace
    - Creative Commons Attribution-ShareAlike 2.5 License



# OpenNLP



- organizační centrum projektů zabývajících se zpracování přirozeného jazyka
- soustřed'uje
  - projekty
  - užitečné odkazy
  - diskuzní fórum
- <http://opennlp.sourceforge.net/>

# Instalace (1)

- Instalace vyžaduje Python 2.4 a vyšší
- Platformy
  - Linux
  - Mac
  - Windows

# Instalace (2)

## 1) Python

*<http://www.python.org/download/>*

## 2) Numerical Python (Numarray)

*[http://sourceforge.net/project/showfiles.php?group\\_id=1369&package\\_id=32367](http://sourceforge.net/project/showfiles.php?group_id=1369&package_id=32367)*

## 3) NLTK lite

*[http://sourceforge.net/project/showfiles.php?group\\_id=30982&package\\_id=156043](http://sourceforge.net/project/showfiles.php?group_id=30982&package_id=156043)*

## 4) NLTK lite corpora

*[http://prdownloads.sourceforge.net/nltk/nltk\\_lite-corpora-0.6.3.zip](http://prdownloads.sourceforge.net/nltk/nltk_lite-corpora-0.6.3.zip)*

# Python a NLP

- Python je vhodný nástroj pro NLP
  - jednoduchý
  - snadno “debugovatelný”
    - výjimky
    - interpretovaný jazyk
  - strukturovatelný
    - moduly, OOP
  - výkonná práce nad (znakovými) řetězci

# Moduly a balíky

- moduly *modules* umožňují znovu použít kód
- balíky *packages* jsou hierarchické *moduly*
- *příkazy pro práci*
  - *import*
  - *from ... import*
  - *reload*

# Moduly a balíky

## import

.Příkaz `import` načítá modul:

```
# Load the regular expression module
```

```
>>> import re
```

.Použití přístupu k metodám (pomocí tečkové notace)

```
# Use the search method from the re module
```

```
>>> re.search('\w+', str)
```

.Zobrazení obsahu modulu pomocí `dir`:

```
>>> dir(re)
```

```
['DOTALL', 'I', 'IGNORECASE', ...]
```

# Moduly a balíky

## from .. import

•Příkaz *from...import* načítá jednotlivé funkce:

```
# Load the search function from the re module
>>> from re import search
>>> nltk_lite.draw.rdparser import *
```

•Poté již může být příkaz použit přímo:

```
# Use the search method from the re module
>>> search('\w+', str)
>>> demo()
```



# Moduly NLTK-lite

- nltk\_lite
- nltk\_lite.chat
- nltk\_lite.contrib
- nltk\_lite.corpora
- nltk\_lite.draw
- nltk\_lite.misc
- nltk\_lite.model
- nltk\_lite.parse
- nltk\_lite.tag
- nltk\_lite.tokenize

# **Natural Language Toolkit**

**prezentace do předmětu PA154**

**Nástroje pro korpusy**

**část 2 – nástroje NLTK**

# **Tokenizace**

## **úvod**

# Tokenizace

## text = sekvence tokenů

```
>>> from nltk_lite.corpora import brown, extract
>>> print extract(0, brown.raw('a'))
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', 'Friday',
'an', 'investigation', 'of', "Atlanta's", 'recent', 'primary',
'election', 'produced', '"', 'no', 'evidence', ""', 'that',
'any', 'irregularities', 'took', 'place', '.']
```

# Tokenizace

## statistika pomocí tokenů

```
>>> def length_dist(text):
...     fd = FreqDist()                # initialize an empty frequency
distribution
...     for token in genesis.raw(text): # for each token
...         fd.inc(len(token))         # found another word with this length
...     for i in range(15):           # for each length from 0 to 14
...         print "%2d" % int(100*fd.freq(i)), # print the percentage of words
with this length
...     print

>>> length_dist('english-kjv')
0 2 14 28 21 13 7 5 2 2 0 0 0 0 0 >>> length_dist('finnish')
0 0 9 6 10 16 16 12 9 6 3 2 2 1 0
```

# **Tagování nástroje NLTK**

# **Analýza parsing v NLTK**

# **Analýza částí informací chunk parsing v NLTK**



# Shrnutí

- NLTK je vhodný nástroj pro NLP
- ..

# Literatura

- **NLTK-Lite Tutorials**  
Steven Bird, Ewan Klein, Edward Loper, 2001-2006
  - <http://nltk.sourceforge.net/lite/doc/en/>
- **Getting Started with NLTK**
  - [http://nltk.sourceforge.net/getting\\_started.html](http://nltk.sourceforge.net/getting_started.html)
- **nltk\_lite API**
  - <http://nltk.sourceforge.net/lite/doc/api/>