NLTK

Общие сведения

Natural Language ToolKit (NLTK) – платформа для создания программ на языке Python для обработки текстов на естественном языке

www.nltk.org

Операции

- Токенизация
- Разметка по частям речи
- Анализ с помощью грамматик
- ...

Разметка по частям речи

Our dataset consists of the full 2010 run.

[('Our', 'PRP$'), ('dataset', 'NN'), ('consists', 'VBZ'), ('of', 'IN'), ('the', 'DT'), ('full', 'JJ'), ('2010', 'CD'), ('run', 'NN'), ('.', '.')]

Методы:

- Регулярные выражения
- N-gram tagging

Разметка по частям речи (2)

The W+W- leptonic decay channels are analyzed using data corresponding to 35 pb-1 of integrated luminosity collected by the ATLAS detector during 2010 at the CERN Large Hadron Collider.

[... ('integrated', 'JJ'), ('luminosity', 'NN'), ('collected', 'VBN'), ...]

Разметка по частям речи (3)

The data used for this analysis are collected over a 21-week period from 30 March 2010 to 17 August 2010, corresponding to an integrated luminosity of 1.327 pb-1 and 1.312 pb-1 in the electron and muon channels respectively.

[... ('integrated', 'VBN'), ('luminosity', 'NN'), ('of', 'IN'), ('1.327', 'CD'), ('pb-1', 'JJ'), ...]

Анализ с помощью грамматик

Our dataset consists of the full 2010 run.

```
(S
  (NP Our/PRP$ dataset/NN)
  consists/VBZ
  of/IN
  (NP the/DT full/JJ 2010/CD run/NN)
  ./.)
```

Анализ с помощью грамматик (2)

```
grammar = r"""
   NP: {<DT|JJ|CD|NN.*|PRP.*>+}
   """
```