The Natives are Revolting!

This and other interesting sentences deconstructed and analysed using the Python Natural Language Toolkit

Safe Hammad
Python Sheffield
30th October 2012
John Rylands Library

At a Special Meeting of The Council of the City of Manchester, held in the Town Hall on Wednesday the 6th day of September, 1899.

It was moved by The Right Honourable The Lord Mayor, Councillor William Henry Audrey, and seconded by Alderman James Wilson Southern, and

Resolved Unanimously that the Members of this Council desire to express their opinion that the powers accorded to them by law for the recognition of eminent services would be fittingly exercised by conferring upon Mrs. Enriqueta Augustina Rylands the Freedom of the City, the highest distinction which it is their privilege to bestow; and that Mrs. Rylands is distinguished and honoured by the community.
“At a special meeting of the council of the City of Manchester held in the town hall on Wednesday the 6th day of September, 1899, it was moved by the Right Honourable The Lord Mayor (Councillor William Henry Vaudrey), seconded by Alderman James Wilson Southern, and resolved unanimously that the members of this council desire to express their opinion that the powers accorded to them by law for the recognition of eminent services would be fittingly exercised by conferring upon Mrs Enriqueta Rylands the Freedom of the City, the highest distinction which it is their privilege to bestow.”

How do we make sense of a sentence like this?
Basic Sentence Analysis

...in pure Python.
Basic Sentence Analysis

...using NLTK.
NLTK

• A platform for building Python applications to work with human language data.

• A library of text processing libraries for classification, tokenisation, stemming, tagging, parsing and semantic analysis. Flat library structure.

• Interfaces to many corpora and lexical resources installable on demand:

  \[
  \texttt{nltk.download()}\]

What can you do with Natural Language Processing?

Automatic summarisation
Information extraction
Information retrieval (search)
Machine translation
Named entity recognition
Natural language generation
Natural language understanding
Optical character recognition
Question answering (think Watson)
Speech recognition
Spoken dialogue systems
Text simplification
Text-to-speech systems
Advanced Sentence Analysis

Techniques which complement statistical analysis to dig into the structure, and ultimately, meaning of a sentence.

1. **Tokenisation.**
   Split sentence into constituent entities (mainly words).

2. **Part of speech tagging.**
   Assign a part of speech to each word.

3. **Parsing.**
   Transform the tagged text into a syntax tree and context free grammar.
1. Tokenisation

- Sentence tokenisation is not as simple as:
  ```python
  text.split()
  ```
- NLTK uses *Penn Treebank Project* rules to handle punctuation and contractions consistently, e.g.:
  ```python
  children's → [children, 's]
  Don't → [Do, n't]
  I'm → [I, 'm]
  ```

```python
nltk.word_tokenize()
```
2. Part of Speech Tagging

- A *part of speech* is a category of words defined by the behaviour of that word in a sentence. Traditionally we're taught about a handful:
  - **Noun**: a person, thing place or idea e.g. *dog, cat, love*.
  - **Verb**: conveys action, occurrence or state e.g. *run, be*.
  - **Adjective**: describes a noun e.g. *green, large, hungry*.
  - **Adverb**: modifies a verb e.g. *quickly, masterfully*.
  - **Pronoun**: reference to a person or thing e.g. *he, she*.
  - **Conjunction**: joins two words or phrases e.g. *and, but*.
  - **Preposition**: adds direction to a noun e.g. *in, to*.
  - **Determiner**: an article e.g. *the, a*. 
2. Part of Speech Tagging (continued)

NLTK uses the Penn Treebank tag set with many more tags encompassing singular v. plural, present v. past tense, noun v. verb phrases:


```
nltk.pos_tag()
```
Effectiveness of Automated POS Tagging

- Our tagger achieved about 85% effectiveness. We should be aiming for over 90% with a bit more work, but hand correction is almost always required (but still difficult – see 34 page Penn Treebank tagging guide!)

- Typical errors are those where words can be used in two senses e.g. both a noun and a verb. Most commonly, the noun is chosen by the tagger as a “best guess” as there are far more nouns than verbs in the English language.

  - ('desire', 'NN') → ('desire', 'VB')
  - ('express', 'NN') → ('express', 'VB')
  - ('that', 'WDT') → ('that', 'IN')
  - ('accorded', 'VBG') → ('accorded', 'VBN')
  - ('eminent', 'NN') → ('eminent', 'JJ')
  - ('conferring', 'NN') → ('conferring', 'VBG')
3. Parsing

- Sentence structure can be represented with a **syntax tree**.
- The syntax tree can then be generalised into a **context free grammar**.
- Analysing sentence structure tends to be a manually intensive process.
- One technique is to iteratively split a sentence into its constituent parts, then split each of those parts in turn.
- A common notation for sentence splitting is **lisp format**.
3. Parsing – syntax tree

Iteration 1

(S the members of this council desire to express their opinion that the powers accorded to them by law for the recognition of eminent services would be fittingly exercised by conferring upon Mrs Enriqueta Rylands the Freedom of the City)

Iteration 2

(S (NP the members of this council) (VP desire to express their opinion that the powers accorded to them by law for the recognition of eminent services would be fittingly exercised by conferring upon Mrs Enriqueta Rylands the Freedom of the City))

Iteration 3

(S (NP the members of this council) (VP (VP desire to express their opinion) (SBAR (IN that) (S the powers accorded to them by law for the recognition of eminent services would be fittingly exercised by conferring upon Mrs Enriqueta Rylands the Freedom of the City))))
3. Parsing - context free grammar

S → NP VP
NP → NP PP | DT NNS | DT NN | PRPS NN | NP IN NP | NP VBN PP | JJ
NNS | DT NNP
PP → IN NP | TO VP | TO PRP IN NN | IN VP
SBAR → IN S
VP → VP SBAR | VB PP | VB NP | VP NP | VP PP | MD VB RB VBN | VBG
RP NNP NNP NNP NP

DT → 'the' | 'this'
NNS → 'members' | 'powers' | 'services'
IN → 'of' | 'that' | 'by' | 'for'
NN → 'council' | 'opinion' | 'law' | 'recognition'
VB → 'desire' | 'express' | 'be'
TO → 'to'
PRPS → 'their'
VBN → 'accorded' | 'exercised'
PRP → 'them'
JJ → 'eminent'
MD → 'would'
RB → 'fittingly'
VBG → 'conferring'
RP → 'upon'
NNP → 'Mrs' | 'Enriqueta' | 'Rylands' | 'Freedom' | 'City'
NLTK and the Context Free Grammar

...time for some more code!
NLTK and the Context Free Grammar

...can we perform the same process in reverse?
Crash Blossom

- Syntactic ambiguity, especially common in the English language. Examples:

  The natives are revolting.
  Japan Violinist linked to JAL crash blossoms.
  Bridge Held Up By Red Tape.
  Milk Drinkers Turn to Powder.
  McDonald’s Fries the Holy Grail for Potato Farmers.
  English teacher given long sentence.
Thank you!

Safe Hammad
Python Sheffield
30th October 2012

http://safehammad.com
@safehammad