NLP and Text Classification

sreejata@leadsift.com

Agenda

- What is NLP
- Difference between the science and the art
- Wordnet
- Concepts:
    - Chunking, Part of speech tagging, Lemming…
- Simplest classifier: Bag of words
- NLTK

What is NLP

- **NLP: Natural Language Processing**
- **The study of how human languages are understood by and interact with computers… It's a branch that crosses artificial intelligence and linguistics.**
- NLP is the most important building block of Information Retrieval/Big Data/Data Mining
- Best language - Python! NLTK (Natural Language Tool Kit) is the most advanced library out there, dedicated to making NLP more accessible
- If any of your data analysis depends on the understanding of the language or any rules of language, you're using NLP.
- How is it different from what linguists do?

Parts and Pieces of using NLP

- Cleaning of data: depends on type of data (Reuters? Twitter? SMS? Reviews? LinkedIn? Logs made by engineers who test a plane engine before takeoff?)
- All the different kinds of taggers and parsers
- Being able to classify properly: the various learning algorithms like Naive Bayes, SVMs, Neural Networks etc...
- To be good at NLP, you need to enjoy programming/coding
- Always be ready for your experiments and classifiers to fail miserably and reiterate
- Sometimes the simplest (and most manual) solutions work best! (Why?)

Common NLP problems

- Summarization and Excerpt creation
- Translation
- Question-Answering
- Auto-Correct/Auto-Suggest
- Text to Speech
- OCR - Why is this an NLP problem?
- Named Entity Recognition - What is this?

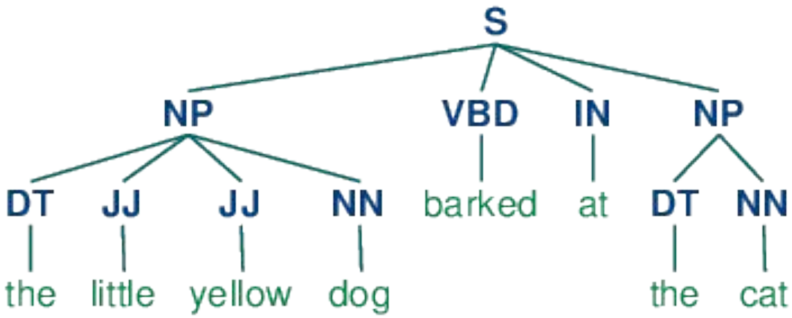
Can you name some uses of NLP where it performs well and others where it fails?

Concepts of NLP: POS tagging

- Part of Speech Tagging
- Where every word is assigned a part of speech (Noun, Verb, Adjective)
- Very fast
- Accurate if you use good libraries and cleaned data
- Limited usefulness
- Typical POS Tagger output:
  - “And now for something totally different”
  - [('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something', 'NN’), ('completely', 'RB'), ('different', 'JJ')]
- https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html ← What the POS Tags stand for - they are pretty standard over the various POS taggers

Concepts of NLP: Parsing

- Analyzing a sentence into not only it’s part of speech, but also phrases and identifying their syntactic roles (subject, predicate etc).
- It breaks a sentence into parts of speeches, noun phrases, verb phrases etc.
- Can be visualized as part of the tree.
- Very slow but useful when doing text analytics



Concepts of NLP: Chunking

- Chunking/Shallow Parsing/Semantic Role Labeling
    - In between POS tagging and full known Parsing
    - Quicker and yet provides some syntactical value
    - Q - Who jumped over the fence? A - The “quick brown fox”
    - He reckons the current account deficit will narrow to only # 1.8 billion in September .
    - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

- Stemming and Lemmatizing
    - Getting words recognizable by our algorithms, an attempt to finding their “root” word
    - Bit, bite, bitten, biting: bit.

Classifiers:

- Bag of Words
- Rule based
- Naive Bayes
- Neural Networks
- Clustering
- SVM
- … many others (name some?)

WordNet

- Princeton University
- https://wordnet.princeton.edu
- A huge corpus of data
- Manually tagged and curated
- Available for free download
- Has things like: part of speech tagging, relation between words, positive and negative sentiments, etc

NLTK

- Developed in the University of Pennsylvania
- Natural Language Toolkit
- Super library in Python for all kinds of natural language processing
- In-built classifiers, lots of tagged corpuses
- (I have personally only scratched the surface)

Try it out

```
from nltk import pos_tag, word_tokenize
text = word_tokenize("And now for something completely different")
print text
nltk.pos_tag(text)

from nltk.stem import PorterStemmer
porter = PorterStemmer()
print(porter.stem("planning"))


from nltk.corpus import wordnet as wn
wn.synsets('dog')
print(wn.synset('dog.n.01').definition())
wn.synsets('happy')

wrd = wn.synset('dog.n.01')
wrd.hypernyms()
```

The Ultimate Prestige: Loebner Prize

- Started in 1990, and is done every year!
- Format is a standard Turing Test (a test designed by Alan Turing)
- A human judge would “chat” simultaneously, via a computer terminal, to two entities, a computer program and an actual human, for 5 mins (now increased to 25 mins).
- After that time, they have to judge which is the computer and which the human
- The prize is $100,000 for anyone who can fool a judge, and the competition will stop when someone successfully fools one!
- Famous ones: ELIZA, Elbot, CleverBot… (Please Try it out!)
- Winner of 2016 - 2018: http://www.mitsuku.com
- *What are some of the “cheat” strategies they are using?*
- Most Notable commercial examples: SIRI, Google Assistant, Phone Customer Service, Watson that won Jeopardy, T1000...

Tutorial

Use the bag of words strategy to find out the sentiment of sentences

Things to consider:

How will you test it?

How will you find the dataset which tags words as positive or negative?

(https://sentiwordnet.isti.cnr.it/)