NLP Tutorial 1

By:
Alok Singh
alok.rawat478@gmail.com
PhD Candidate, National Institute of Technology Silchar

Content

1. What is NLP?
2. Why NLP?
3. NLP tasks
4. Basic requirements of solving any NLP task
5. Corpus
6. Types of corpus
7. Links to download corpus
8. Regular Expression
9. Tokenization

What is NLP?

- NLP = Computer Science + AI + Computational Linguistics
- To get computer to perform useful task involving human languages.
  a. Human-Machine communication.
  b. Human-Human communication: Machine Translation (MT).
  c. Extracting information from text.

Why NLP?

1. Answering questions: What are the possible symptoms of COVID19?
2. Information extraction: Extracting venue, timing and date from an Email.
3. Machine Translation

NLP tasks

1. Searching
2. Named-Entity Recognition
3. Parts-of-Speech tagging (POS)
4. Information extraction and retrieval
5. Text Classification/Clustering
6. Sentiment analysis
7. Summarization
8. Machine Translation
9. Answering queries
10. Automated speech recognition (ASR)
11. Many other……..

Basic requirements of solving any NLP task?

1. Selection of problem and identifying problem domain
2. Corpus
3. Text preprocessing tools (normalization, tokenization, stemming, lemmatization)
4. Algorithms
5. Evaluation measures

Corpus

1. What is a corpus?
   - A corpus is a collection of natural language (text, and/or transcriptions of speech or signs)
   - most available corpora are text only, but recently multimodal corpora, including sign language corpora are also getting popularity

2. Why it is required?
   a. to solve NLP problem -> train large ML/DL model
   b. help in inference some pattern

For more detail 👇👇👇
https://nordiskateckensprak.files.wordpress.com/2014/01/knb_whatisacorpus_cph-2013_outline.pdf

Different types of corpus

1. Based on language:
    a. **Monolingual corpus**
        i. It contains texts in one language only.
        ii. Used in:
            1. Language modeling
            2. parts of speech tagging
            3. word embeddings
            4. checking the correct usage of a word or
            5. looking up the most natural word combinations etc.,
    b. **Parallel corpus, multilingual corpus**
        i. Consists of two or more monolingual corpora (mostly translation)
        ii. Used in:
            1. Text to text MT, Speech to text etc.,
    c. Multimodal corpus

For more detail: https://www.sketchengine.eu/corpora-and-languages/corpus-types/

Factors to decide the quality of a corpus

1. how much clean it is?
2. size
3. different classes it includes/ diversity (Unless corpus has been collected for specific tasks)
4. how much balance it is?

Well known corpus

1. TreeBank Corpus: **a treebank is a parsed text corpus that annotates syntactic or semantic sentence structure**
   a. **Used for:** part-of-speech taggers, parsers, semantic analyzers etc,.
   b. Most commonly used treebank corpus are:
      i. Penn Treebanks: POS Tagging
      ii. Syntactic Treebanks
      iii. Semantic treebanks
2. PropBank Corpus (Proposition Bank)
3. VerbNet
4. WordNet

Penn Treebank's:
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Download links for some notable text corpora:

- **Brown Corpus:** https://www.english-corpora.org/coca/
- **Corpus of Contemporary American English (COCA)**

  https://www.english-corpora.org/coca/
- **Penn Treebank-3 (paid):** https://catalog.ldc.upenn.edu/LDC99T42
- **Data dumps of English Wikipedia** https://dumps.wikimedia.org/enwiki/latest/
- **Wikipedia Links Data:** https://code.google.com/p/wiki-links/downloads/list

Contd…

- **Amazon Customer Reviews :** https://s3.amazonaws.com/amazon-reviews-pds/readme.html
- **IMDb Reviews:** http://ai.stanford.edu/~amaas/data/sentiment/
- **Jeopardy Question-Answer Dataset:**

  http://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/
- **Enron Email Dataset:** https://www.cs.cmu.edu/~enron/
- **20 Newsgroups:** http://qwone.com/~jason/20Newsgroups/
- **Sentiment140:** http://help.sentiment140.com/for-students/
- **SMS Spam Collection:** https://archive.ics.uci.edu/ml/datasets/sms+spam+collection
- **WordNet:** https://wordnet.princeton.edu/

Regular expression

1. Regular expression: string searching, pattern matching
   a. It is a sequence of characters mainly used to find or replace patterns embedded in the text.
   b. In NLP, it play major role in text preprocessing……

Example: 1

```
path1 = "C:\Desktop\nataliya"
print("string:",path1)
```

Output: C:\Desktop\nataliya

```
path2 = r"C:\Desktop\nataliya"
print("string:",path2)
```

C:\Desktop\nataliya

2.

```
str = "4 tutorials in Feb'22."
y = re.findall("\d+", str)   print(y)
```

NOTE: re in regex library in python '+' after '\d' will continue to extract digits till encounters a space

2. Some most common commands:
   a. **/[A-Z]/** an upper case letter
   b. **/[a-z]/** a lower case letter
   c. **/[0-9]/** a single digit
   d. **/[^A-Z]/** not an upper case letter
   e. **[a^s]** look for pattern a^s
   f. **/beg.n/** any character between beg and n
   g. **Kleene *** eg a* ={ε a,aa,aaa,aaaa…. }
   h. Kleene + eg a+ = {a,aa,aaa,aaaa…..}

Tokenization

1. Process of tokenizing or splitting a string, text into a list of tokens
    a. Example:
        i. Book-> Chapters
2. Based on languages the criteria of tokenization it vary
    a. For English, HIndi we separate text based on space (but some instance it get fail)
    b. For Chinese, Arabic it will be different
3. Used for:
    a. Corpora cleaning, removing stop words etc….
    b. Analysing the occurrence of words in the text
    c. To build a vocabulary
4. Different level of tokenization: word level, sentence level etc……

Tutorial sheet 1

1. Find the number of tokens?
   a. `print("string:",path1)`
   b. `print("\d",8*9);`
2. Find the output : (in python)
   a. string = “This is NLP tutorial”
      i. string.split()
      ii. string.split(“ ”)
   b. s = 'A computer science tutorial'

      match = re.search(r'science', s)

      i. print(match.start())
      ii. print(match.end())
   c. Try below functions with above string s (in 2.b)
      i. re.findall(s), re.compile(s), re.split(s), re.sub(s), re.escape(s)

3. Write a short note on (include introduced by, size of corpus, major tasks it is used for):

   1. TreeBank Corpus
   2. PropBank Corpus
   3. VerbNet
   4. WordNet

Acknowledgements

1. https://aclanthology.org/J93-2004.pdf
2. Speech and Language Processing. Daniel Jurafsky & James H. Martin.
3. Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. Using Large Corpora, 273.