Natural Language Processing

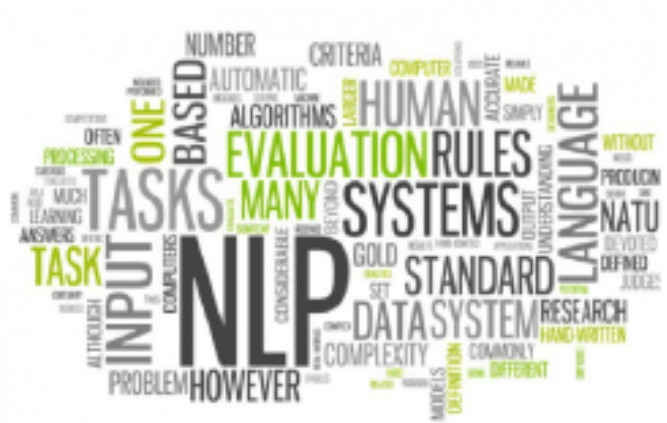
Unit 1: Preprocessing

August 2019

What is Preprocessing?

Preprocessing

Text Preprocessing is a natural language task which process input text in order to make it predictable and analyzable for a natural language system performing one or many tasks.



Why is important?

- Text is often unestructured and is hard to process for a system

Why is important?

- Text is often unestructured and is hard to process for a system
- Natural text produced by humans has ortographic errors

Why is important?

- Text is often unestructured and is hard to process for a system
- Natural text produced by humans has ortographic errors
- Many words are not useful for some sets of specific NLP tasks

Why is important?

- Text is often unestructured and is hard to process for a system
- Natural text produced by humans has ortographic errors
- Many words are not useful for some sets of specific NLP tasks
- Not every verb conjugation and adjective form is useful for every task

Case normalization

Case normalization

Normalizing ALL your text data, although commonly overlooked, is one of the simplest and most effective form of text preprocessing. It is applicable to most text mining and NLP problems and can help in cases where your dataset is not very large and significantly helps with consistency of expected output. Most of the time all data is represented as lower case text.

Stop words removing

Stop words removal

Stop words are a set of commonly used words in a language. Examples of stop words in English are “a”, “the”, “is”, “are” and etc. The intuition behind using stop words is that, by removing low information words from text, we can focus on the important words instead.

Stemming

Stemming

Stemming is the process of reducing inflection in words (e.g. troubled, troubles) to their root form (e.g. trouble). The "root" in this case may not be a real root word, but just a canonical form of the original word. Stemming uses a crude heuristic process that chops off the ends of words in the hope of correctly transforming words into its root form. So the words "trouble", "troubled" and "troubles" might actually be converted to troubl instead of trouble because the ends were just chopped off (ughh, how crude!). There are different algorithms for stemming. The most common algorithm, which is also known to be empirically effective for English, is Porters Algorithm.

Lemmatization

Lemmatization

Lemmatization on the surface is very similar to stemming, where the goal is to remove inflections and map a word to its root form. The only difference is that, lemmatization tries to do it the proper way. It doesn't just chop things off, it actually transforms words to the actual root. For example, the word "better" would map to "good". It may use a dictionary such as WordNet for mappings or some special rule-based approaches.

Text normalization

Text normalization

A highly overlooked preprocessing step is text normalization. Text normalization is the process of transforming text into a canonical (standard) form. For example, the word “gooood” and “gud” can be transformed to “good”, its canonical form. Another example is mapping of near identical words such as “stopwords”, “stop-words” and “stop words” to just “stopwords”. Text normalization is important for noisy texts such as social media comments, text messages and comments to blog posts where abbreviations, misspellings and use of out-of-vocabulary words (oov) are prevalent. This paper showed that by using a text normalization strategy for Tweets, they were able to improve sentiment classification accuracy by 4

Noise removal

Noise removal

Noise removal is about removing characters digits and pieces of text that can interfere with your text analysis. Noise removal is one of the most essential text preprocessing steps. It is also highly domain dependent.

Text enrichment

Text enrichment

Text enrichment involves augmenting your original text data with information that you did not previously have. Text enrichment provides more semantics to your original text, thereby improving its predictive power and the depth of analysis you can perform on your data.

Which tasks you should do?

- Must Do:
  - Noise removal
  - Case normalization
- Should Do:
  - Simple normalization
- Task dependant:
  - Advanced normalization (e.g. addressing out-of-vocabulary words)
  - Stop-word removal
  - Stemming / lemmatization
  - Text enrichment / augmentation

Tokenization

Tokenization

Tokenization is the task to extract functional units or tokens from an information string in order to perform sintactic and lexical based tasks. The process of segmenting running text into words and sentences. Electronic text is a linear sequence of symbols (characters or words or phrases). Naturally, before any real text processing is to be done, text needs to be segmented into linguistic units such as words, punctuation, numbers, alpha-numerics, etc. This process is called tokenization.

Tokenization

English and spanish words are often separated from each other by whitespace, but whitespace is not always sufficient:

- New York is treated as a large word despite it contain spaces
- We need to separate I'm into the words I and am
- To process tweets we need to tokenize emoticons like :)
- Chinese don't have white spaces. How we tokenize? (jieba https://github.com/fxsjy/jieba)

Regular expressions

Regular expressions are expressions written in an special format in order to find specific patterns in text corpus:

- Emails (mario.campos@upy.edu.mx)
- Numbers (1245.32)
- Company names (SoldAI SAPI S.A de C.V)
- Phone numbers (+529994263828)

Let's code

Execute in your machine

git clone https://github.com/MaxSob/nlp-introduction

Assignments

Assignment 2: Write a jupyter notebook to perform the following NLP tasks defined as functions:

- Stemming
- Lemmatizing
- Cleansing
- Word Tokenize
- Phrase Tokenize
- Text normalization
- Regex entity extractor

Make a jupyter notebook explaining the code and loading the tweey database in order to pre process the text extracting hashtags and twitter user names as tokens

References

[1] Jurafsky, D., Martin, J.: Speech and Language Processing 2nd. ed. (2009).

[2] Mikolov, T., Corrado, G., Chen, K. y Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: CoRR, (2013). https://dblp.org/rec/bib/journals/corr/abs-1301-3781

[3] Pinker, S.: The Stuff of Thought - Language as a window into human nature. https://www.youtube.com/watch?v=5S1d3cNge24

[4] https://kavita-ganesan.com/text-preprocessing-tutorial