Application of AI on IT Service Management

Arun singh Sivaprakash
Pramod kumar Nagaraj
Van Tien NGUYEN
Alexander POPPE

EPITA Graduate School of Computer Science

Advisor
Olivier Berthet

Master of Science in Artificial Intelligence Systems

07-July-2021

Introduction

Here is the introduction of our Thesis

State of Art

Natural Language Processing includes three main stages:

1. Text Processing

    a. Data Cleaning
    Here we remove special characters (ie. "@", "!",...), html tags (ie. "/h1", "/span",...) from the raw text as they do not contain any information for the model to learn and are irrelevant or noisy data.

    b. Data Normalization
    Data normalization involes steps such as case normalization, punctuation removal,...so that the text is in a single format for the machine to learn.

    c. Punctuaion Removal
    Replace punctuation with space.

    d. Tokenization (NLTK)
    Tokenization is the process of breaking up text documents into individual words called tokens.

    e. Stop Word Removal
    In this step, we remove all non-important words like "a", "is", "the", "and",...
    There is an in-built stopword list in NLTK which we can use to remove them from the text documents. However, this is not the standard stopwords list for all problem. In our project, we need to define our own set of stopwords.

    f. Parts of Speech Tagging (POS)
    Determine POS tags for each word (ie. noun, verb, adverb,...). We can also use the in-built part of speech tag provided by NLTK.

    g. Named Entity Recognition
    In information extraction, a named entity is a real-world object,

such as persons, locations, organizations, products,..., that can be denoted with a proper name.

h. Stemming
Stemming is a process of reducing a word to its root form.

i. Lemmatization
Lemmatization is a technique for reducing words to its normalized form. But in this case, the transformation actually uses a dictionary to map words to their actual form.

2. Feature Extraction
This is a way of extracting feature vectors from the text after processing step so that it can be used in the machine learning model as input. This extracted feature from the text documents can be a wordnet of a grapth of nodes, a vector reprenting words.

a. Bag of Words (BoW)
BoW is a way of extracting features from the text for use in modeling such as machine learning alogrithms. It treats each document as a collection/ bag of words.
A BoW is a representation of text taht describes a occurrence of words within a document. It involves two things:

- A vocabulary of known words in the corpus/ set of documents.
- A measure of the presence of known words.

b. Term Frequency-Inverse Document Frequency (TF-IDF)

- Term Frequency is a numerical statistic taht is intended to reflect how important a word is to a document in a collection or corpus. It is a measure of how frequently a term appears in a document.
- IDF is a measure of how important a term is. We need the IDF because computing just the TF alone is not suffient to understand the importance of words.

3. Modeling
The final stage of the NLP pipeline is modeling, which includes designing a statistical or machine learning model, fitting its parameters to training data, using an optimization procedure, and then using it

to make predictions about unseen data. Some of the machine learning algorithms used here are:

a. Clustering

b. Neural network

c. Random Forest Classifier

System's Artchitecture

Methodology

Results

Conclusion

References

References

[1] Wikipeida: Natural language processing

[2] Article: iClass Gyansetu