Natural Language Processing: A practical introduction

Introduction

Text is the most popular way that people use to represent information and describe entities and relations between objects. Thanks to the vast and rapid growth of internet technologies and mobile networks, text data has become very widely available across various platforms and domains.This growth has necessitated countless applications that require not only processing and storing the huge available amount of text data but also inferring useful business insights out of it. Here comes the role of natural language processing (NLP) which is defined as the intelligent interaction between man and machine using text processing and machine learning techniques.

The main objective of this course is to familiarize the student with the important concepts of NLP and give them a head-start towards building simple NLP systems, namely document classification and a recommendation system.

Prerequisite

It is expected that the student is familiar with the following topics:

- Classification techniques such as logistic regression, support vector machines, and decision trees.
- K-means clustering

It is also recommended that the student has some programming experience using Python 3.5 or later.

Contents

Upon the completion of this course, the student should be familiar with the following topics:

- Regular Expressions
- Part-of-speech (POS) tagging
- Stemming and lemmatization
- Tokenization
- Context-free grammar and CNF grammar
- Common feature engineering methods:
  - Concept of term-document matrix and bag-of-words
  - Term frequency inverse document frequency (TFIDF)
  - Latent semantic indexing (LSI)
  - Word2Vec and Doc2Vec
  - Glove
- Clustering techniques
  - K-means clustering
  - Latent Dirichlet Allocation (LDA)
- Language modelling (LM)
  - Rule-based LM
  - Statistical n-gram

    - Recurrent neural networks
- Classification
    - Concept of word-embeddings and vocabulary
    - CNN vs RNN (and similar classifier such as)
- Very simple recommendation system
    - Collaborative filtering
- Advanced Topics
    - Transfer learning
    - Multi-task learning
    - Machine translation

In addition, the student will have hands-on experience on the following Python packages:

- NLTK
- Gensim
- Tensorflow
- PySpark

Labs

- **Lab 1:** Reuters News classification using
    - Tfidf + SVM or any other classifier
    - Doc2vec + SVM or any other classifier
    - Word embedding + LSTM
- **Lab 2:** Topic modelling of Newsgroups dataset using LDA
- **Lab 3:** Simple recommendation system for IMDB movies sample using colloborative filtering

Reports

- For every lab project, the student must submit the following for evaluation:
    a. A summary report highlighting
        - Brief description of the problem and its practical applications
        - The specific software tools used for implementation
        - Assumptions made
        - Detailed performance analysis of the developed system
    b. Github link of the developed code with proper documentation:
        - What there project is about
        - Quick start
        - Installation and instructions of use