Middle East Technical University

Department of Computer Engineering

CENG 463 - Introduction to NLP

Fall 2024 - Assignment 1

Due: 1 December 2024 23:59
Submission: **via ODTUClass**

1 Overview

In this assignment, you are asked to implement a named-entity recognition (NER) model that identifies and classifies named entities from a dataset of news articles[1]. Secondly, you need to perform topic modeling to identify common themes in the customer reviews.

2 Task1: Named Entity Recognition (50 Points)

Named entity recognition (NER) identifies predefined categories of objects in a body of text. The goal of this task is to develop a model that can accurately recognize entities such as people, organizations and locations. You will use the CoNLL 2003 NER Shared Task's English dataset. You need to perform the following steps:

1. **Feature Extraction**: The dataset consists of <token, POS tag, syntactic chunk tag, NER label> columns separated by a single space. Therefore, you need to extract features for each token. The features can be:

   - **Basic features**: Token itself, token lowercase, prefix/suffix of the token.
   - **Context features**: Neighboring tokens (previous/next token).
   - **Linguistic features**: Part-of-speech (POS) tags or word shapes (capitalization, digits, etc.).

   Note that you are expected to briefly mention which features you employ for training your model.

2. **Model Implementation**: Implement one of the following classifiers for recognizing multiple entity types (e.g., person, organization, location): Conditional Random Field (CRF), biLSTM or multinomial logistic regression. Select only one and provide a brief explanation for your choice of model.

3. **Evaluation**: Evaluate the model using metrics such as precision, recall and F1-score.

4. **Reporting**: Summarize your findings and suggest potential improvements for future iterations of the NER system. Additionally, discuss whether your model encountered class imbalance issues and how you addressed them.

[1] https://aclanthology.org/W03-0419.pdf

3 Task2: Topic Modeling (50 Points)

You are asked to analyze customer reviews from an e-commerce platform. More specifically, you will perform topic modeling to identify common themes in the reviews, apply POS tagging to extract grammatical information and utilize lemmatization to normalize the words for better analysis.

You need to perform the following steps:

1. **Preprocessing**: Clean the text data by removing special characters, numbers and stop words. Implement lemmatization to convert words to their base forms. For example, "running" becomes "run".

2. **Part-of-Speech Tagging**: Using NLTK or spaCy library, perform POS tagging on the cleaned review texts. Eliminate the tokens except nouns, noun phrases and verbs.

3. **Topic Modeling**: Implement topic modeling using Latent Dirichlet Allocation (LDA) which is a topic modeling technique for uncovering the central topics and their distributions across a set of documents. Identify and list the top 5 topics found in the reviews along with their associated keywords.

4. **Reporting**: Summarizing the findings, including:

   - How did you select the number of topics?
   - Coherence score measures how semantically related the top words are in each topic. Which coherence type did you employ (c_v, u_mass or uci)?
   - The most common topics identified in the reviews.
   - Print examples of reviews that belong to each topic.

5. **Visualization**: Plot word clouds for each topic displaying associated keywords

4 Submission

You are expected to submit a zip file (**"e1234567_hw1.zip"**) which includes the IPython notebooks: *"e1234567_phase1.ipynb"* and *"e1234567_phase2.ipynb"*

For each task, you must use the corresponding IPython file to implement the task and include your comments and discussions.

5 Tutorials

1. NLTK
2. spaCy
3. scikit-learn
4. Jupyter Notebook
5. Gensim LDA Topic Modeling

6 Regulations

- Submission will be done via ODTUClass. You are expected to submit a zip file containing your code and explanations presenting the analysis of your results.
- Late submission is not allowed.
- We have zero tolerance policy for cheating. People involved in cheating will be punished according to the university regulations.
- If you have any questions about the assignment, feel free to ask them via the Discussion Forum (on ODTUClass) or email (rfcekinel@ceng.metu.edu.tr). For more specific questions, office hours (at A206) for this assignment are on Tuesdays from 10:30 to 11:30 AM.