Natural Language Processing tasks

The student will be able to do Python assignments in addition to the exam within the scope of the course (by consulting with the course instructor). The student is expected to complete five ones of the following coding tasks and prepare a YouTube video for each of them to explain these tasks in practice.

Video presentation for each task should not exceed 10 minutes. In all of the video presentations, the student is required to make a face video recording so that the identification can be made. In the video, the student should first introduce himself, then quickly tell the title of the project, if any, the dataset details, and the libraries he/she uses. He should then run the application and comment on the results in a few sentences.

Midterm Tasks

Task 1. The results must be listed by applying at least 2 different tokenizers and 2 different lemmatizers representing the Tokenization and Lemmatization processes.

Task 2. Using N-gram analysis, list all 2-grams and 3-grams according to a certain threshold (eg, more than 5 in frequency) in a Turkish corpus. With these study, emphasize the importance of lematization for Turkish. Also, comment on what purpose this application can be used in the real world.

Task 3. By selecting 2 corpora prepared for different purposes, some basic features (unigram, bigam, trigram numbers, POS-Tag variety numbers etc.) should be compared and interpreted. The differences between the corpora and the relation between the preparation purposes of the corpora should be examined.

Task 4. A comparison will be made using different POS Taggers (at least two) on a corpus prepared for Turkish (the student will not prepare a new corpus, he/she must download it from internet, and the download link will be specified in the task). At least three words that POS Taggers labelled differently should be found and expressed.

Task 5. Basic semantic analysis of a Turkish corpus that will be downloaded from Internet will be done with statistical approaches. Accordingly, in the corpus;
- a list of all words (vocabulary) and their frequencies,
- a list of bigram and trigram tokens that have been used at least 5 times,
- the most similar two words using Latent Semantic Analysis should be determined.

Task 6. A very simple example should be prepared to show how to do word sense disambiguation using the LESK algorithm on a Turkish corpus. Here, the student can be flexible about corpus and glossary, in other words, a corpus with a few sentences and a dictionary with a few definitions can be prepared.

Task 7. First, low and high frequency words (for threshold = 5) should be determined in a Turkish collection. Words with low frequency will be assumed to have spelling errors. To solve this problem, a Lexical Similarity function will be used (student can find at internet or prepare) and it will be suggested that these erroneous assumed words can change to the most lexically similar word from high-frequency words. The code to be prepared must give a two-column list: low-frequency words in the first column and the most lexically similar high-frequency words in the second column.

Task 8. First of all, English word vectors calculated by any method (e.g. Word2Vec, GloVe, fastText or SemSpace) should be obtained from the internet. We must also have an English corpus. Using any Semantic Similarity method that the student chooses or creates, the most similar sentence in the corpus should be found for a sentence given externally.

Final Tasks

Student can prepare one of the following projects based on a LLM model (such as BERT, RoBERTa, ELMo, GPT-x, Llama-x). But the project must be able to run completely locally without internet. The student should choose a popular/known dataset for fine-tune, however confirmation should be requested from the course instructor.

Task A. Frequently Asked Questions chatbot

Task B. Machine Translation system

Task C. Keyword Extraction or Document Summarization system

Task D. A shopping assistant based on reviews on e-commerce sites

Task E. An advertisement robot based on customer reviews

Students can ask questions about homework just by attending classes. Those who cannot attend the lesson because their lessons conflict, can write an e-mail with “NLP tasks” subject.