Modele językowe
Lab 3

Zadanie 1. (**6p**) We will focus on word embeddings (both contextual and non-contextual). Note: the texts we will embed always consist of a single word (but not necessarily a single token).

a) Propose a method for utilizing **non-contextual** token embeddings (determined by a transformer[1] to determine word embeddings. You can use the program from Lecture 7 (embedding.ipynb). Check the quality (measured by the ABX test) of these embeddings[2].

b) Use contextual token embeddings from BERT to determine embeddings for words. Perform ABX tests again.

Also, check how robust the contextual embeddings are to distortions. This means modifying the ABX test such that, instead of comparing original words, we compare their distorted versions (e.g., 'dłuigopis' instead of 'długopis', 'krowka' instead of 'krówka', etc.). Propose two methods for distorting words and describe the results you obtain. Note: every word should be distorted in some way[3]!

Evaluation procedure: save the embeddings in a text file `word_embedings_file.txt`, where each line is formatted as:

```
[word] float_1 float_2 ... float_D
```

The embeddings are evaluated using the script `word_emb_evaluation.py`. For testing distortions, use your own modification of this file.

Zadanie 2. (**(6+X)p**) In this task, we will classify reviews using transformer models. You can use the program from the lecture (herbert.ipynb). In this task, you should use three models:

1. A generative model, such as Papuga, which determines text probabilities (similar to List 1)

2. A BERT-type encoder (e.g., Herbert), as a feature extractor

3. A traditional Machine Learning model that integrates the results of the two previous models. This model should be trained on the training review dataset and tested on the test dataset.

The bonus value is calculated as: $20 * (a - 0.85)$, where a is the accuracy on the test dataset. If you wish, you may also use results from the next task. Note: if you have difficulty running the models simultaneously, you can process all texts with one model, save the results, and then process them with the second model.

Zadanie 3. (**8+1p**) In this task, you should check whether data augmentation can improve classification results, where BERT is treated as a feature extractor. There are 3 separately scored procedures for generating new review variants:

a) Mechanical augmentation (introducing distortions in the text, such as typos, changing letter case, errors related to Polish letters, etc.). (2p)

b) Augmentation with a generative model, such as Papuga. You should generate reviews based on the original review while preserving its polarity (i.e., whether it is positive or negative). Note that the "imaginationóf the language model does not necessarily have to be a drawback – in this procedure, the generated texts do not need to be entirely correct. (3p)

c) This augmentation procedure should be based on Word2Vec and should, as much as possible, preserve the meaning of the text. For example, the review: *The hotel is generally very nice.* could be changed to *The guesthouse is especially very beautiful.*, and *I recommend this physiotherapist to everyone!* to *I suggest this orthopedist to everyone!*.

---

[1]you may use Papuga or Herbert for Polish, or GPT-2 DistilBERT for English

[2]Information: a very simple procedure can achieve a score of 0.7

[3]This means that distortions related to removing Polish diacritical marks must be combined with another procedure to allow for word distortions in the Latin alphabet