# NLP course Assignment 2: Contextualized Vectors, Parts of Speech, and Named Entities

Daniel Bazar 314708181

Lior Krengel 315850594

## 0) Warmup

1. We encoded the sentence “I am so <mask>” and:
   1. We extracted the vectors for “am” and for “<mask>”. Both vectors are partially represented because of their shape (768):

am:

[ 2.9096e-01, 9.2609e-02, 1.4434e-01, -1.8008e-01, 5.1247e-01,

…

1.3153e-01, -8.0886e-02, 3.9851e-02]

<mask>:

[ 3.4503e-01, -1.1836e-01, -1.9594e-02, -8.2120e-02, 7.9033e-01,

…

2.3184e-01, -3.3112e-02, 2.8167e-02]

* 1. We extracted the top-5-word predictions for “am” and for “<mask>” and their probabilities:

am: <mask>:

|  |  |
| --- | --- |
| am | 0.9999 |
| is | 3.9379e-05 |
| 'm | 2.9938e-05 |
| was | 8.6892e-06 |
| feel | 8.5510e-06 |

|  |  |
| --- | --- |
| sorry | 0.6065 |
| proud | 0.1276 |
| grateful | 0.1142 |
| happy | 0.0881 |
| blessed | 0.0636 |

1. We find two sentences that share the same word, such that the cosine similarity between the word vectors in the two sentences is **very high**:

Sentence1: 'I love you' Sentence2: 'I love him' similarity: 0.9897

1. We find two sentences that share the same word, such that the cosine similarity between the word vectors in the two sentences is **very low** (low is relative):

Sentence1: 'The fission of the cell could be inhibited with certain chemicals.'  
Sentence2: 'His cell phone worked, so he spoke with his parents and sister-in-law.'  
similarity: 0.8418

1. We find a sentence with n words, that is tokenized into m > n tokens by the tokenizer:

original sentence: Didn't I tell you it's gonna be a rock 'n' roll weekend with lots o' fun, and we'll gather 'round the campfire, singin' our favorite songs 'til the break o' dawn? **(n=31)**

tokenized sentence: ['<s>', 'Did', 'n', "'t", ' I', ' tell', ' you', ' it', "'s", ' gonna', ' be', ' a', ' rock', " '", 'n', "'", ' roll', ' weekend', ' with', ' lots', ' o', "'", ' fun', ',', ' and', ' we', "'ll", ' gather', " '", 'round', ' the', ' camp', 'fire', ',', ' sing', 'in', "'", ' our', ' favorite', ' songs', " '", 'til', ' the', ' break', ' o', "'", ' dawn', '?', '</s>'] **(m=49)**

## 1) Part-of-speech tagging

In this part of the assignment, we will explore the notion of part-of-speech tagging.  
The “catch” in this assignment is that we don’t do it in the standard way.  
Instead of train a classifier to predict the correct part-of-speech tag from vector representation, in this assignment we will experiment with predicting parts of speech of words without training any classifiers.

### No word vectors

In this section we are not allowed to use any word vectors at all.

### Static word vectors

As before, but now we are allowed to use static vectors.

### Contextualized word vectors

As before, but now we are allowed to use the output of roberta-base model.