Yotam Ben Dov 316387950

Ex.2

1. Warmup

The word ‘am’ isn’t masked and so the top 5 words are ‘am’ with a score of 1.00, and the top 5 words for ‘mask’ are:

{'score': 0.33508574962615967, 'token': 6661, 'token\_str': ' sorry', 'sequence': 'I am so sorry'},

{'score': 0.06557566672563553, 'token': 2602, 'token\_str': ' proud', 'sequence': 'I am so proud'},

{'score': 0.05825677886605263, 'token': 6161, 'token\_str': ' grateful', 'sequence': 'I am so grateful'},

{'score': 0.044130850583314896, 'token': 1372, 'token\_str': ' happy', 'sequence': 'I am so happy'},

{'score': 0.03153745457530022, 'token': 12230, 'token\_str': ' blessed', 'sequence': 'I am so blessed'}

The sentences I picked for the similarity segment are "I can take a right" and "you should turn right", who have a 0.9448 cosine similarity.

The sentences I picked for the dissimilarity segment are NOTSIMILAR = "Thou shall carry me like a princess" and "princesses are royalty", who have a 0.7478 cosine similarity.

Any sentences turned into vectors has m > n tokens, since the tokenization creates a starting point token and an end point token.

1. Part-of-Speech Tagging