**DVA256 Data Management and Datafication**

**Answers for Assignment Exercise 2 (OVN1)**

*Please use this template to answer the questions and upload this doc file during submission.*

| *Q. No* | *Answer* | *Comments* |
| --- | --- | --- |
| *1.* | [*https://countwordsfree.com/stopwords/swedish*](https://countwordsfree.com/stopwords/swedish) *Or, in code, using the: nltk.corpus.stopwords.words(‘swedish’)* |  |
| *2.* | *Lemmatization is more accurate, as it reverts the word back to its primary form.*  *Stemming does… not. Example: cries can be stemmed back to cri, which is wrong.* |  |
| *3.* | *There is no change in word count. Instead, the lemmatized words were reverted back to their original form.* |  |
| *4 (i).* | *First we use .split(‘’) to make a list of each text. We then match the words from each text in a for loop, while filtering any special characters, to see if they match, then calculate the normalized value and see how close each text is to each other.* |  |
| *4 (ii).* | *We use two lists within a list, to make a vector of each word. We then count each word in both texts as they appear and add it to the list in the relevant index.* |  |
| *5* | *No. As the lemmatized words revert to their original verb, adj, noun, the number of words matching will be higher than the non lemmatized text.* |  |