**Team Assignment**

Given the following corpus, made up of 5 documents:

|  |  |  |
| --- | --- | --- |
| **#** | **Description** | **Category** |
| 1 | Peter Piper picked a peck of pickled peppers. | C1 |
| 2 | A peck of pickled peppers Peter Piper picked. | C1 |
| 3 | Where is the peck of pickled peppers Peter Piper picked? | C1 |
| 4 | How much wood would a woodchuck chuck if a woodchuck could chuck wood? | C2 |
| 5 | As a woodchuck would if a woodchuck could chuck wood! | C2 |

1. Using spaCy, lemmatize the corpus of documents and remove stop words, punctuation marks and spaces (this should reduce the vocabulary)
2. Save the processed document back to a text file
3. Manually build a multinomial NB model (create the probability tables) using the corpus created in #1 and #2.
4. Predict whether the following document belongs to category C1 or C2

**The woodchuck picked peppers and chucked wood.**

* Complete part 1 and 2 in Python (submit a notebook)
* For parts 3 and 4, you can use Excel (you can also verify your results using Python)
* Remember that in 4, you need to pre-process the document for which you are predicting the category (use Spacy once again)