Lab 2
DATA 441/641
Applied Natural Language Processing

1. Data Acquisition–Back translation (2 points)

Translate a sentence to your favorite language using Backtranslation with googletrans. For the same sentence, try different languages and print your results. For this task you can use the BackTranslation python library, (`https://pypi.org/project/BackTranslation/`), which is implemented to back translate the words among any two languages.

2. HTML Parsing and Cleanup (5 points)

(a) In this exercise we will scrap data from webpages using the python library Beautiful Soup (`https://www.crummy.com/software/BeautifulSoup/bs4/doc/`). For the first task use StackOverflow as your main source to extract question and best-answer pairs from this website. Identify and extract the question-best answer for the question "What is the module/method used to get the current time?"

(b) Follow similar arguments as in part (a) and write the code to download at your local directory the following file (`https://zoisboukouvalas.github.io/COVID19_Twitter_Dataset.xlsx`).

3. Extract text from pdf files (3 points)

Open the 'sample.pdf' file which is available on Canvas and extract the plain text from the document. There are several libraries that can be used for this task such as PyPDF or PDFMiner. However, they are far from perfect. Go through the following article (`https://filingdb.com/b/pdf-text-extraction`) and list some of the most important difficulties when we have to extract text from pdf files.

4. Text pre-processing (10 points) For this exercise this will be our corpus.

Need to finalize the demo corpus which will be used for this notebook & should be done soon !!. It should be done by the ending of this month. But will it? This notebook has been run 4 times !!"

(a) Lower case the corpus.

(b) Remove digits from the corpus, punctuations, and trailing white-spaces.

(c) Tokenize the corpus using NLTK or Spacy and remove stop words.

(d) Perform stemming using NLTK and print the output.

(e) Perform lemmatization using NLTK and print the output.