# ICA 2 Stemming and Lemmatizing

(September 16th)

In this activity, we will work on stemming and lemmatizing.

**(Same as ICA 1) From NLTK import gutenberg corpus and open up the book *Alice in Wonderland* by Lewis Carol.**

**(Same as ICA 1)** Step 0: Preprocessing - strip the text of punctuation, and lower case the words - obtaining the original text as a list of words.

Step 1: From NLTK import PorterStemmer and LancasterStemmer. Also import WordnetLemmatizer.

Step 2: Run the sentences in the first three paragraphs of *Alice in Wonderland* through PorterStemmer and LancasterStemmer.

Identify at least three main differences that you can observe in the output of the two stemmers.

Step 3: Run the sentences in the first three paragraphs of *Alice in Wonderland* through the WordnetLemmatizer.

Step 4: From a language modeling perspective, when might it be advantageous to include stemming as a pre-processing step? When might it be disadvantageous?
Hint: Think of NLP applications such as information retrieval or sentiment analysis.

**In Class Special Question (not to be included in ICA file upload):**

The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs would you argue shouldn't be conflated. Give your reasoning.

1. abandon/abandonment
2. absorbency/absorbent
3. marketing/markets
4. university/universe
5. volume/volumes