# CS6370 Assignment 1

**CS14B043 - Harshal Gawai, CS14B017 - Mulle Venkata Manohar Reddy**

Ans. 1)
The obvious top-down approach to sentence segmentation for English texts would be segmentation using a set of delimiters such as full-stop/period (.), commas(,), question mark(?), exclamation(!), next line(\n)

Ans.2)
No, top-down approach will not always do correct segmentation.
Abbreviations are used in written language to denote the shortened form of a word. In many cases abbreviations are written as a sequence of characters terminated with a period.
Titles such as Mrs., Mr., Ms., Dr., Prof., Capt., Gen., Sen., Rev., Hon., and St.

Eg. “Mrs. Rita is an M.P. She still works till 6 p.m.” have 2 sentences.
Would be segmented as “Mrs”, “Rita is an M”, “P”, “She still works till 6 p”, “m”

Ans 3)
PunktSentenceTokenizer is a sentence boundary detection algorithm that must be trained to be used. NLTK already includes a pre-trained version of the PunktSentenceTokenizer. Punkt is designed to learn parameters (a list of abbreviations, etc.) unsupervised from a corpus similar to the target domain. This is more of a bottom-up approach(facilitated by top down knowledge).
The top-down approach fails to do correct segmentation on abbreviations unlike PunktSentenceTokenizer.

Ans 5)

Splitting a text(string) using whitespace as delimiter.

Ans 6)
The Penn Treebank tokenizer uses regular expressions to tokenize text. It assumes that text has already been split into sentences.

Ans 8)
**Stemming** is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.

**Lemmatization**, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization the root word is called Lemma. A lemma (plural lemmas or lemmata) is the canonical form, dictionary form, or citation form of a set of words.

Ans 9)
Both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.
*Stemming* does crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time,
*Lemmatization* does things properly with the use of a vocabulary and morphological analysis of words, normally aiming to return to the base or dictionary form of a word.
For search engine application , lemmatization is preferred.
eg.
Stemming: Dancing => Danc
Lemmatization: Dancing =>Dance

In the above question, the list of stopwords denotes top-down knowledge. Can you think
of a bottom-up approach for stopword removal?

Ans 12)
The amount of the documents and material we encounter each day is simply beyond our processing capacity. To reduce size and eliminate unnecessary words(stopwords) is essential. But having a list of stopwords then working with a search engine is not the best way.
Instead we can create probabilistic model for how often a words appears in those documents. And eliminate words(stopwords) above some threshold frequency of this model.