108_text_to_numbers

December 20, 2018

1 Converting text to numbers

In [1]:
```
import nltk
import numpy as np
import pandas as pd

def text_to_numbers(text, cutoff_for_rare_words = 1):
    """Function to convert text to numbers. Text must be tokenzied so that
    test is presented as a list of words. The index number for a word
    is based on its frequency (words occuring more often have a lower index).
    If a word does not occur as many times as cutoff_for_rare_words,
    then it is given a word index of zero. All rare words will be zero.
    """

    # Flatten list if sublists are present
    if len(text) > 1:
        flat_text = [item for sublist in text for item in sublist]
    else:
        flat_text = text

    # get word freuqncy
    fdist = nltk.FreqDist(flat_text)

    # Convert to Pandas dataframe
    df_fdist = pd.DataFrame.from_dict(fdist, orient='index')
    df_fdist.columns = ['Frequency']

    # Sort by word frequency
    df_fdist.sort_values(by=['Frequency'], ascending=False, inplace=True)

    # Add word index
    number_of_words = df_fdist.shape[0]
    df_fdist['word_index'] = list(np.arange(number_of_words)+1)

    # replace rare words with index zero
    frequency = df_fdist['Frequency'].values
    word_index = df_fdist['word_index'].values
```

```
            mask = frequency <= cutoff_for_rare_words
            word_index[mask] = 0
            df_fdist['word_index'] =  word_index

            # Convert pandas to dictionary
            word_dict = df_fdist['word_index'].to_dict()

            # Use dictionary to convert words in text to numbers
            text_numbers = []
            for string in text:
                string_numbers = [word_dict[word] for word in string]
                text_numbers.append(string_numbers)

            return (text_numbers)
```

In [5]:
```
# An example tokenised list

text = [['hello', 'world', 'Michael'],
          ['hello', 'world', 'sam'],
          ['hello', 'universe'],
          ['michael', 'makes', 'a', 'good', 'cup', 'of', 'tea'],
          ['tea', 'is', 'nice'],
          ['michael', 'is', 'nice']]

text_numbers = text_to_numbers(text)
print (text_numbers, 3)
```

```
[[1, 2, 0], [1, 2, 0], [1, 0], [3, 0, 0, 0, 0, 0, 4], [4, 5, 6], [3, 5, 6]] 3
```

In [ ]: