Brody Vogel NLP Homework #3

September 20, 2018

```
In [70]: # Brody Vogel (Section 2) NLP Assignment 3 #
         import nltk
         from nltk import word_tokenize
         import re
         import pprint

In [71]: #9a

         # define the loading function
         def load(f):
             # open and return the text
             text = open(f)
             return(text.read())

         # here's my file
         text = load('/Users/brodyvogel/Desktop/corpus.txt')

         # (I think these are all of the punctuation characters)
         pattern = r"""(?x)
         \.+      # period or ellipsis
         |,       # comma
         |;       # semi-colon
         |\?       # question mark
         |!       # exclamation point
         |\"       # quotation mark
         |:       # colon
         |'       # apostrophe
         |\-       # hyphen
         """

         # get the tokens
         nltk.regexp_tokenize(text, pattern)

Out[71]: ['.', '.', '-', '-', ',', '.', '.', '-', '.']

In [76]: #9b
         # same thing, different regex
         pattern = r"""(?x)
         \$?[0-9]+\.[0-9]+\$?      # monetary amounts (dollars, at least, like $5.00 or 5.00$)
         |[0-9]+\-[0-9]+\-[0-9]+       # dates (in the format mm-dd-yyyy)
         |[A-Z][a-z]+(?:\s?[A-Z][a-z]+)(?:\s?[A-Z][a-z]+)?      # proper names (with or without
         """

         nltk.regexp_tokenize(text, pattern)
```

```
Out[76]: ['Brody Vogel', '$15.00', '09-20-2018', 'Georgetown Analytics Department']
```

```
In [109]: #43
          from nltk.corpus import udhr
          from nltk import FreqDist
          from nltk.metrics.spearman import *

          # create a dictionary of frequency distributions to loop through
          dict_of_freq_dists = {}
          # when I add more languages, it usually gives me the wrong answer
          for lang in ['English-Latin1', 'French_Francais-Latin1',
                       'German_Deutsch-Latin1', 'Icelandic_Yslenska-Latin1']:
              # add the frequency distribution to the dictionary
              dict_of_freq_dists[lang] = FreqDist([w.lower() for w
                                                   in udhr.raw(lang) if w.isalpha() == True])
          # function for guessing which of the 4 languages text came from
          def guess(text):
              # create a frequency distribution for the text
              text_freq_dist = FreqDist([w.lower() for w
                                         in text if w.isalpha() == True])
              # keep track of the best correlation so far
              best_corr = (0, '')
              # go through the frequency distributions of languages in the dictionary
              for language in dict_of_freq_dists.keys():
                  # update our best_corr if this language's is higher
                  if spearman_correlation(ranks_from_sequence(dict_of_freq_dists[language]),
                                          ranks_from_sequence(text_freq_dist)) > best_corr[0]:
                      best_corr = (spearman_correlation(
                          ranks_from_sequence(dict_of_freq_dists[language]),
                          ranks_from_sequence(text_freq_dist)), language)
              # return the best one
              return(best_corr)

          # text taken from BBC News
          text = ('''Police have arrested a suspect in the fatal stabbing of a recently
          engaged woman who was running near her home in a trendy Washington DC neighbourhood.
          Wendy Karina Martinez, 35, was attacked on Tuesday night only one mile from the Whit
          She died after seeking help at a nearby takeaway restaurant.
          Anthony Crawford, 23, has been arrested and charged with murder, police said.
          A motive for the attack is still unclear.
          Friends describe Ms Martinez as "brilliant and a hard worker".
          "Wendy should have been shopping for her wedding dress on Friday," said
          Kristina Moore, as she stood outside the restaurant where Ms Martinez sought help af
          What are the charges?
          Anthony Crawford has been charged with first degree murder for what police
          believe was a random knife attack on Tuesday night around 20:00 (01:00GMT).
          The Georgetown University alumna was stabbed multiple times,
          including in the neck, and went to a nearby Chinese restaurant for help.
          What US female joggers face on every run
          Bystanders at the business, including a nurse, attempted to save her life.
          Image copyright CBS
          Image caption The takeaway restaurant manager showed footage of the panic as
          patrons noticed the attack
          "We don't have a motive," Police Chief Peter Newsham said on Thursday,
          adding that it does not appear that there was any attempt to rob Ms Martinez.
          Police say Mr Crawford has a criminal history, and they will be examining him
          for any mental health issues.
          Mr Crawford was arrested in a park on Wednesday night and was
          transported to hospital with an injury to his hand, according to police.
          Authorities said they received tips from the community after a surveillance image wa
          Image copyright Police handout
          Image caption Police released grainy footage of the attacker
          "This is one of those types of unsettling incidents that sometimes happen in large c
          Who was the victim?
          ''')

          # call the function on our example text
          guess(text)
```

```
Out[109]: (0.6153846153846154, 'English-Latin1')
```