Задача 1: сравнение предложений

Объявление функций и импорт библиотек

In [52]:

```
import re
import numpy as np
from scipy import spatial
```

In [53]:

```
# dictionary of uniq words
words_dict = dict()
```

In [54]:

```
# func for add new uniq word in list

def uniq_words(list_of_word, word_for_append):
    isOnList = False
    for word in list_of_word:
        if word_for_append == word:
            isOnList = True
            break
    if isOnList == False:
        list_of_word.append(word_for_append)
    return list_of_word
```

In [55]:

```
# func for return list wiht count of words

def count_words(inpt_sentense):
    sentense_counts = []
    for i in range(0, len(words_dict)):
        count = 0
        for word in inpt_sentense:
            if words_dict[i] == word:
                count += 1
        sentense_counts.append(count)
    return sentense_counts
```

Чтение из файла

In [56]:

```
sent_l = [] # list of sentences

with open('sentences.txt') as f_sentences:
    for line in f_sentences:
        sent_l.append(line.strip().lower())
```

Токенизация и создание списка слов

In [57]:

```
sentences_split = [re.split('[^a-z]', lin) for lin in sent_l]

words_duty_list = [] # list of word with repeat

for sen_word in sentences_split:
    for word in sen_word:
        if word:
            words_duty_list.append(word)

uniq_word_list = [] # list of uniq words

for w in words_duty_list:
    uniq_word_list = uniq_words(uniq_word_list, w)
```

Сопоставление индексов словам

In [58]:

```
iter = 0;
for word in uniq_word_list:
    words_dict[iter] = word
    iter += 1
```

Создание матрицы

In [59]:

```
senWordMatrix = np.array(map(count_words, sentences_split)) # matrix with words entries
#print senWordMatrix
#print senWordMatrix.shape

#matrix_f = open('matrix.txt', 'a')
#for l in senWordMatrix:
#    matrix_f.write(str(l))
#matrix_f.close()
```

```
[[1 1 1 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 [0 0 2 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [1 0 1 ... 0 0 0]
 [0 0 1 ... 1 1 1]]
(22L, 254L)
```

Вычисление косинусного растояния

In [60]:

```
dict_distance = dict() # dictionary of sentences with distance

for i in range(1,senWordMatrix.shape[0]):
    dict_distance[scipy.spatial.distance.cosine(senWordMatrix[0],senWordMatrix[i])] = i
```

просто вывод всех растояний

In [61]:

```
list_keys = list(dict_distance.keys())
list_keys.sort()

for k in list_keys:
    print 'строка с растоянием = ' + str(k) + ' и индексом = ' + str(dict_distance[k])
```

```
строка с растоянием = 0.7327387580875756 и индексом = 6
строка с растоянием = 0.7770887149698589 и индексом = 4
строка с растоянием = 0.8250364469440588 и индексом = 21
строка с растоянием = 0.8328165362273942 и индексом = 10
строка с растоянием = 0.8396432548525454 и индексом = 12
строка с растоянием = 0.8406361854220809 и индексом = 16
строка с растоянием = 0.8427572744917122 и индексом = 20
строка с растоянием = 0.8644738145642124 и индексом = 2
строка с растоянием = 0.8703592552895671 и индексом = 13
строка с растоянием = 0.8740118423302576 и индексом = 14
строка с растоянием = 0.8804771390665607 и индексом = 11
строка с растоянием = 0.8842724875284311 и индексом = 8
строка с растоянием = 0.8885443574849294 и индексом = 19
строка с растоянием = 0.8951715163278082 и индексом = 3
строка с растоянием = 0.9055088817476932 и индексом = 9
строка с растоянием = 0.9258750683338899 и индексом = 7
строка с растоянием = 0.9402385695332803 и индексом = 5
строка с растоянием = 0.9442721787424647 и индексом = 18
строка с растоянием = 0.9527544408738466 и индексом = 1
строка с растоянием = 0.956644501523794 и индексом = 17
```

Ближайшие по косинусному растоянию

Запись в файл

In [62]:

```
first = 1.0
second = 1.0

for k in list_keys:
    if float(k)<first:
        first = k
for k in list_keys:
    if float(k)<second and float(k)>first:
        second = k

task_first_answer = open('task_first_answer.txt', 'w')
task_first_answer.write(str(dict_distance[first]) + ' ' + str(dict_distance[second]))
task_first_answer.close()
```

исходная строка:

In comparison to dogs, cats have not undergone major changes during the domestication process.

ближайшая по косинусному растоянию:

Domestic cats are similar in size to the other members of the genus Felis, typically weighing between 4 and 5 kg (8.8 and 11.0 lb).

следующая:

In one, people deliberately tamed cats in a process of artificial selection, as they were useful predators of vermin.

темтики данных предложений действительно близки к исходной

среди приведенных предложений в файле существуют менее схожие по смыслу предложения, однако, на мой взгляд, результаты таких вычеслений действительно не очень точны (ввиду приведенных в условиях задачи обстоятельсв).