```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mour
```

```
lines = []
with open("/content/drive/MyDrive/Машин сургалт/Lab_02/SMSSpamCollection") as f:
  lines = f.readlines()
```

Хэрэглэгдэх нэмэлт функцууд:

```
# Тухайн тэмдэгт нь тоо агуулсан эсэхийг шалгана
def containNumber(s):
  for char in s:
    if char.isdigit():
      return 1
  return 0
```

```
# Тухайн тэмдэгт нь URL мөн эсэхийг шалгана
def containURL(s):
  s = s.lower()
  if (s.find("www") >= 0 or s.find("http") >=0 or s.find("https")>=0):
    return 1
  else:
    return 0
```

```
# dictionary дотор key нь агуулагдаж байгаа эсэхийг шалгана
def checkKey(dict, key):
  if key in dict:
    return 1
  else :
    return 0
```

```
#Тэмдэгт мөр нь том жижиг үсэг агуулсан эсэхийг шалгана
def isMixed(s):
  if not s.islower() and not s.isupper():
    return 1;
  else:
     return 0
```

```
#Тэмдэгт мөр нь тоо мөн эсэхийг шалгана
def isNumber(str):
  try:
```

```
    num = float(str)
    return 1
  except:
    return 0
```

TF-IDF

```
import nltk
nltk.download("stopwords")
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords

dataset = []
symbols = "!\"#$%&()*+-./:;<=>?@[\]^'_`{|}~\n"
ps = PorterStemmer()
for doc in lines:
  # табын зайгаар нь салгана
  text = doc.split("\t")
  # шинэ мөрнөөс эхлүүлэх тэмдэгтийг хассан
  sms = text[1][0:len(text[1])-1]
  # мессежийг үг бүрээр нь салгаж хүснэгт болгоно
  sms_list = sms.split()
  sms_without_sw = ""
  for word in sms_list:
    #өөр өөр нөхцөл авсан үгнүүдийг язгуур үгээр нь солино
    word = ps.stem(word)
    for char in symbols:
      # тэмдэгтүүдийг хоосон зайгаар солино
      word = word.replace(char, '')
    if len(word) > 2:
      # тоо агуулсан эсэхийг шалгана
      if not containNumber(word):
        # URL агуулсан эсэхийг шалгана
        if not containURL(word):
          if word.lower() not in stopwords.words('english'):
            # stopwords(the , a , an, in ...) дотор тухайн үг нь байхгүй
            # бол тэмдэгт мөр дээр нэмнэ
            sms_without_sw = sms_without_sw + ' ' + word
  if not len(sms_without_sw) == 0:
    # дата сет рүүгээ ялгаж авсан өгүүлбэрээ нэмнэ
    dataset.append(sms_without_sw)


wordset = {}
temp_list = []
for sms in dataset:
  # нэг мессежийг үг үгээр нь салгана
  sms_list = sms.split()
  # dataset дотор байгаа бүх үгийг нэгтгэнэ
```

```
    # dataset дотор байгаа бүх үгийн нэгтгэнэ
    temp_list = list(set().union(temp_list, sms_list))
wordset = set(temp_list)
del temp_list
len(wordset) # давхцал байхгүй нийт үг


# Term Frequency
# c = нэг document-д байгаа тухайн нэг үгийн давтамжийн тоо
# l = document-н урт
# tf = c/l
def TF(sms):
  tfDict = {}
   # мессежийн урт
  smsLength = len(sms)
   # гаргаж авсан wordset-р dictionary хийнэ
  wordDict = dict.fromkeys(wordset, 0)
  for word in sms.split():
    # wordset дотор тухайн word байгаа эсэхийг checkKey функцээр шалгана
    if checkKey(wordset, word):
      # мессеж болгонд dictionary үүсгэж үг бүрийг нь хэдэн ширхэг байгааг тоолно
      wordDict[word] += 1
  for word , count in wordDict.items():
    # tf
    tfDict[word] = count/float(smsLength)
  return tfDict


tf_list = [] # мессеж болгоны term frequency-г агуулсан лист
for sms in dataset:
  tf = TF(sms)
  tf_list.append(tf)


# Inverse Document Frequency
# l = нийт Document-н тоо (count of corpus)
# val =  wordset -н үгнүүд corpus дотор хэдэн ширхэг байгааг илэрхийлнэ
# idf = log(l/val)
def IDF():
  import math
  dataset_len = len(dataset)
  # wordset-н үг тус бүр хэдэн ширхэг байгааг олохын тулд dictionary үүсгэнэ
  idfDict = dict.fromkeys(tf_list[0].keys(), 0)
  for tf in tf_list:
    for word , val in tf.items():
      # word -н давтамжийн тоо 0-с их байвал idfDict[word]-н тоог нэмэгдүүлнэ
      if val > 0:
        idfDict[word] +=1
  for word, val in idfDict.items():
    # idf
    idfDict[word] = math.log(dataset_len/val)
  return idfDict
```

```
# idf-г олно
idf = IDF();


# tfidf = tf * idf
def TFIDF(tf):
  tfidf = {}
  for word , val in tf.items():
    if checkKey(idf, word):
      # тухайн үгний tf-н утгыг  харгалзах idf-н утгаар үржүүлнэ
      tfidf[word] = val*idf[word]
  return tfidf


tfidf_list = [] # мессежний tfidf-г агуулсан лист
for tf in tf_list:
  temp_tfidf  = TFIDF(tf)
  tfidf_list.append(temp_tfidf)
import pandas as pd
pd.DataFrame(tfidf_list)
# доорх хүснэгтэд бүгд 0 мэт харагдаж байгаа боловч
# мессеж тус бүрээр нь тулган харвал хариу нь гарсан байгаа
```

	skye	cake,	matter	long,	american	dryer	allalo	dehydrated	attract	freemess
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	
5530	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
5531	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
5532	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
5533	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
5534	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

5535 rows × 7892 columns

Hand features

Features

- URL агуулсан эсэх
- reward
- money
- sms-ний урт
- free
- $
- тоон цуваа агуулсан эсэх
- prize
- congratulations
- cash
- won
- Том жижиг үсэг агуулсан үг байгаа эсэх

Дээрх tf-idf-г ашиглан гаргаж авсан feature дээр гараар нэмсэн feature-үүд.

```
del dataset
del tfidf_list
del idf
del tf_list
del wordset
```

```
features_set = {'URL?','length?', 'reward', 'money', 'free', '$', 'numbers?', 'prize', 'congr
```

```
dataset = []
for doc in lines:
  # файлаас үншсан датанаас мессежийг ялган авна
  text = doc.split("\t")
  sms = text[1][0: len(text[1])-1]
  dataset.append(sms)
dataset
```

```
#true = 1 , false = 0
def computeFeatures(sms):
  word_list = sms.split()
  word_dict = dict.fromkeys(features_set , 0)
  word_dict['length?'] = len(sms)

  for word in word_list:
    word.lower()
    # URL агуулсан эсэхийг шалгана
    if containURL(word):
      word_dict['URL?'] = 1
    # том жижиг үсэг холилдсон үг байгааг шалгана
    if isMixed(word):
      word_dict['mixed?'] = 1
```

```
        word_dict[ mixed? ] = 1
      # тоон цуваа байгаа эсэхийг шалгана
      if isNumber(word):
        word_dict['numbers?'] = 1
      if word == 'reward':
        word_dict[word] = 1
      if word =='money':
        word_dict[word] = 1
      if word == 'free':
        word_dict[word] = 1
      if "$" in word:
        word_dict['$'] = 1
      if word == 'prize':
        word_dict[word] = 1
      if "congratulation" in word:
        word_dict['congratulations'] = 1
      if word == 'cash':
        word_dict[word] = 1
      if (word == 'won' or  word == 'win'):
        word_dict['won'] = 1

    return word_dict
```

```
hand_feature = [] # мессеж болгоны feature-г агуулсан dictionary-уудын лист
for sms in dataset:
  features = computeFeatures(sms)
  hand_feature.append(features)

import pandas as pd
pd.DataFrame(hand_feature)
```

	money	won	numbers?	cash	mixed?	free	congratulations	length?	prize	reward
0	0	0	0	0	1	0	0	111	0	0
1	0	0	0	0	1	0	0	29	0	0
2	0	1	1	0	1	0	0	155	0	0
3	0	0	0	0	0	0	0	49	0	0
4	0	0	0	0	1	0	0	61	0	0
...	...	...	...	...	...	...	...	...	...	...
5569	0	1	1	0	1	0	0	160	0	0
5570	0	0	0	0	1	0	0	36	0	0
5571	0	0	0	0	1	0	0	57	0	0
5572	0	0	0	0	1	1	0	125	0	0
5573	0	0	0	0	1	0	0	26	0	0

5574 rows × 12 columns