```
from ast import literal_eval
import pandas as pd
import numpy as np
```

Файл унших

```
def read_data(filename):
  data = pd.read_csv(filename, sep='\t')
  data['tags'] = data['tags'].apply(literal_eval)
  return data
```

```
train = read_data('train.tsv')
validation = read_data('validation.tsv')
test = pd.read_csv('test.tsv', sep='\t')
```

Уншсан файлын толгой хэсэг

```
train.head()
```

	title	tags
0	How to draw a stacked dotplot in R?	[r]
1	mysql select all records where a datetime fiel...	[php, mysql]
2	How to terminate windows phone 8.1 app	[c#]
3	get current time in a specific country via jquery	[javascript, jquery]
4	Configuring Tomcat to Use SSL	[java]

```
validation.head()
```

	title	tags
0	Why odbc_exec always fail?	[php, sql]
1	Access a base classes variable from within a c...	[javascript]
2	Content-Type "application/json" not required i...	[ruby-on-rails, ruby]
3	Sessions in Sinatra: Used to Pass Variable	[ruby, session]
4	Getting error - type "json" does not exist - i...	[ruby-on-rails, ruby, json]

```
import re
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
```

```
    [nltk_data]   Package stopwords is already up-to-date!
```

Текст нормалчлах функц

```
REPLACE_BY_SPACE_RE = re.compile('[/(){}\[\]\|@,;]')
BAD_SYMBOLS_RE = re.compile('[^0-9a-z #+_]')
STOPWORDS = set(stopwords.words('english'))

def text_prepare(text):
  text = text.lower()
  text = REPLACE_BY_SPACE_RE.sub(" ", text)

  text = BAD_SYMBOLS_RE.sub("", text)

  text = re.sub(r'\s+'," ", text)

  text = ' '.join([word for word in text.split() if word not in STOPWORDS])

  return text
```

TfidfVectorizer функц оруулах

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

Сургалтын өгүүлбэрүүд

```
text = [
        "good movie",
        "not a good movie",
        "did not like",
        "i like it",
        "good one"
]
```

Юниграм болон биграммаар хамгийн багадаа 2 удаа байх өгөгдлөөр вектор үүсгэх, сургах

```
tfidf = TfidfVectorizer(min_df=2, max_df=0.5, ngram_range=(1,2))
features = tfidf.fit_transform(text)
```

Үр дүн

```
pd.DataFrame(
    features.todense(),
    columns = tfidf.get_feature_names()
)
```

	good movie	like	movie	not
0	0.707107	0.000000	0.707107	0.000000
1	0.577350	0.000000	0.577350	0.577350
2	0.000000	0.707107	0.000000	0.707107
3	0.000000	1.000000	0.000000	0.000000
4	0.000000	0.000000	0.000000	0.000000

Сургасан векторыг ашиглаж тестийн өгүүлбэрийг вектор болгох

```
test = ["did good good movie"]
test_vec = tfidf.transform(test)
```

```
print(test_vec)
```

```
      (0, 2)        0.7071067811865476
      (0, 0)        0.7071067811865476
```

✓ 0s completed at 4:31 PM