1. Importing Packages

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os

from sklearn.model_selection import train_test_split
```

```
PKG_DIR="/content/drive/MyDrive/offline_packages"

!pip install --no-index --find-links="$PKG_DIR" transformers datasets evaluate
```

```
Looking in links: /content/drive/MyDrive/offline_packages
Requirement already satisfied: transformers in /usr/local/lib/python3.12/dist-packages (4.57.3)
Requirement already satisfied: datasets in /usr/local/lib/python3.12/dist-packages (4.0.0)
Processing ./drive/MyDrive/offline_packages/evaluate-0.4.6-py3-none-any.whl
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-packages (from transformers) (3.20.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.34.0 in /usr/local/lib/python3.12/dist-packages (from transformers) (
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.12/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from transformers) (25.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.12/dist-packages (from transformers) (6.0.3)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.12/dist-packages (from transformers) (2025.11.3)
Requirement already satisfied: requests in /usr/local/lib/python3.12/dist-packages (from transformers) (2.32.4)
Requirement already satisfied: tokenizers<=0.23.0,>=0.22.0 in /usr/local/lib/python3.12/dist-packages (from transformers) (0
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.12/dist-packages (from transformers) (0.7.0)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.12/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.12/dist-packages (from datasets) (18.1.0)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /usr/local/lib/python3.12/dist-packages (from datasets) (0.3.8)
Requirement already satisfied: pandas in /usr/local/lib/python3.12/dist-packages (from datasets) (2.2.2)
Requirement already satisfied: xxhash in /usr/local/lib/python3.12/dist-packages (from datasets) (3.6.0)
Requirement already satisfied: multiprocess<0.70.17 in /usr/local/lib/python3.12/dist-packages (from datasets) (0.70.16)
Requirement already satisfied: fsspec<=2025.3.0,>=2023.1.0 in /usr/local/lib/python3.12/dist-packages (from fsspec[http]<=20
Requirement already satisfied: aiohttp!=4.0.0a0,!=4.0.0a1 in /usr/local/lib/python3.12/dist-packages (from fsspec[http]<=202
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.12/dist-packages (from huggingface-hub<1
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.12/dist-packages (from huggingface-hub<1.0,>=0
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests->transform
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests->transformers) (3.11)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests->transformers) (
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests->transformers) (
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.12/dist-packages (from pandas->datasets) (2.
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: aiohappyeyeballs>=2.5.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4
Requirement already satisfied: aiosignal>=1.4.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1-
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fs
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1-
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.8.2->pandas->dat
Installing collected packages: evaluate
Successfully installed evaluate-0.4.6
```

```
from transformers import AutoConfig,AutoModelForSequenceClassification, AutoTokenizer
from datasets import Dataset, DatasetDict
from transformers import TrainingArguments, Trainer
import evaluate
```

2. EDA

```
df = pd.read_excel("https://github.com/laxmimerit/All-CSV-ML-Data-Files-Download/raw/master/fake_news.xlsx")
df.head()
```

	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1

```
df[['title','text']]
```

	title	text
0	House Dem Aide: We Didn’t Even See Comey’s Let...	House Dem Aide: We Didn’t Even See Comey’s Let...
1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Ever get the feeling your life circles the rou...
2	Why the Truth Might Get You Fired	Why the Truth Might Get You Fired October 29, ...
3	15 Civilians Killed In Single US Airstrike Hav...	Videos 15 Civilians Killed In Single US Airstr...
4	Iranian woman jailed for fictional unpublished...	Print \nAn Iranian woman has been sentenced to...
...	...	...
20795	Rapper T.I.: Trump a ’Poster Child For White S...	Rapper T. I. unloaded on black celebrities who...
20796	N.F.L. Playoffs: Schedule, Matchups and Odds -...	When the Green Bay Packers lost to the Washing...
20797	Macy’s Is Said to Receive Takeover Approach by...	The Macy’s of today grew from the union of sev...
20798	NATO, Russia To Hold Parallel Exercises In Bal...	NATO, Russia To Hold Parallel Exercises In Bal...
20799	What Keeps the F-35 Alive	David Swanson is an author, activist, journa...

20800 rows × 2 columns

```
df.info()
```

Show hidden output

```
print(df.isna().sum())
df.dropna(inplace=True)
print("\nDone Deleting\n")
print(df.isna().sum())
```

```
id           0
title      558
author    1957
text        43
label        0
dtype: int64

Done Deleting

id        0
title     0
author    0
text      0
label     0
dtype: int64
```

```
df.duplicated().sum()
```

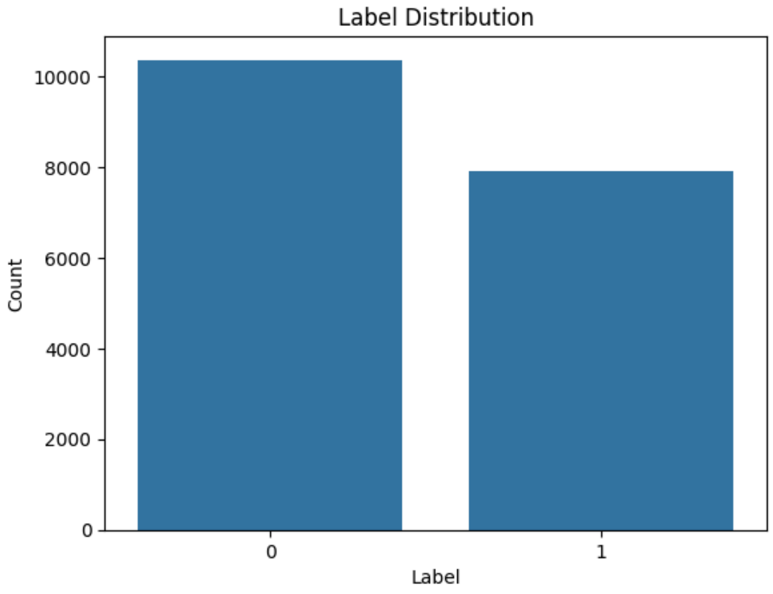
```
np.int64(0)
```

```
df['label'].value_counts()
```

	count
label	
0	10361
1	7920

dtype: int64

```
sns.barplot(x=df['label'].value_counts().index, y=df['label'].value_counts().values)
plt.xlabel("Label")
plt.ylabel("Count")
plt.title("Label Distribution")
plt.show()
```



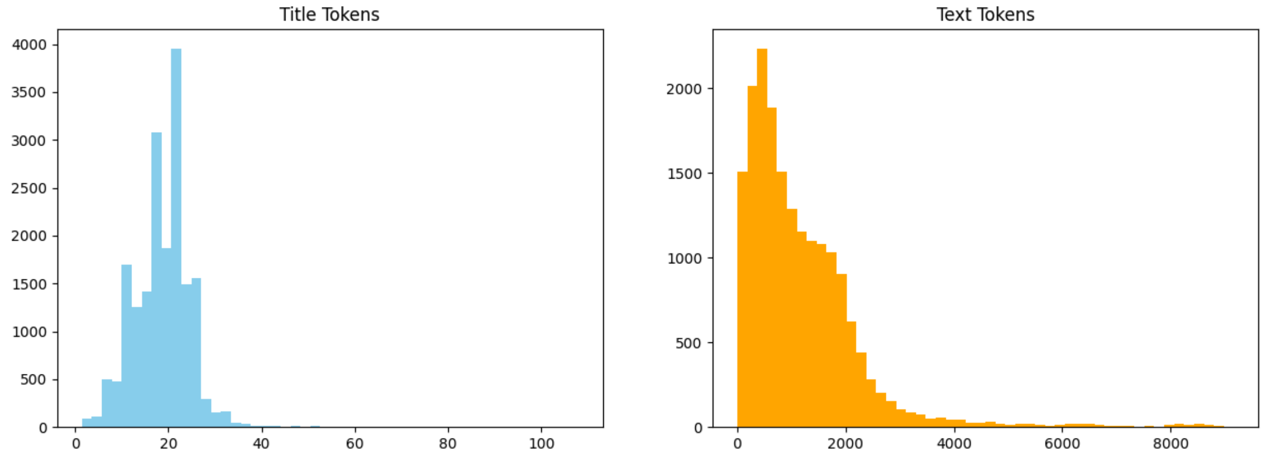
```
# 1.5 tokens per word on average
df['title_tokens'] = df['title'].apply(lambda x: len(x.split())*1.5)
df['text_tokens'] = df['text'].apply(lambda x: len(x.split())*1.5)

fix,ax = plt.subplots(1,2,figsize=(15,5))

ax[0].hist(df['title_tokens'], bins=50, color = 'skyblue')
ax[0].set_title("Title Tokens")

ax[1].hist(df['text_tokens'], bins=50, color = 'orange')
ax[1].set_title("Text Tokens")

plt.show()
```



```
df.rename(columns={'label':'labels'}, inplace=True)
df.head()
```

	id	title	author	text	labels	title_tokens	text_tokens
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1	21.0	1230.0
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0	13.5	1065.0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1	10.5	1899.0

3. Train-Test-Split

```
train,test = train_test_split(df, test_size=0.3, random_state=42,stratify=df['labels'])
test,validation = train_test_split(test, test_size=1/3, random_state=42,stratify=test['labels'])

print(train.shape,test.shape,validation.shape)
```

```
(12796, 7) (3656, 7) (1829, 7)
```

4. Dataframe -> HF Dataset

```
dataset = DatasetDict({
    'train': Dataset.from_pandas(train,preserve_index=False),
    'test': Dataset.from_pandas(test,preserve_index=False),
    'validation': Dataset.from_pandas(validation,preserve_index=False)
})
dataset
```

```
DatasetDict({
    train: Dataset({
        features: ['id', 'title', 'author', 'text', 'labels', 'title_tokens', 'text_tokens'],
        num_rows: 12796
    })
    test: Dataset({
        features: ['id', 'title', 'author', 'text', 'labels', 'title_tokens', 'text_tokens'],
        num_rows: 3656
    })
    validation: Dataset({
        features: ['id', 'title', 'author', 'text', 'labels', 'title_tokens', 'text_tokens'],
        num_rows: 1829
    })
})
```

5. label2id, id2label

```
label2id = {"Real": 0, "Fake": 1}
id2label = {0:"Real", 1:"Fake"}
```

6. Model, Tokenizer

```
distilbert_dir = '/content/drive/MyDrive/offline_models/distilbert-base-uncased'
tinybert_dir = '/content/drive/MyDrive/offline_models/tinybert'
```

```
distilbert_tokenizer = AutoTokenizer.from_pretrained(distilbert_dir,local_files_only=True)
distilber_model = AutoModelForSequenceClassification.from_pretrained(distilbert_dir,local_files_only=True,num_labels=2,id2

tinybert_tokenizer = AutoTokenizer.from_pretrained(tinybert_dir,local_files_only=True)
tinybert_model = AutoModelForSequenceClassification.from_pretrained(tinybert_dir,local_files_only=True,num_labels=2,id2labe
```

```
Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at /content/drive/MyDrive
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of BertForSequenceClassification were not initialized from the model checkpoint at /content/drive/MyDrive/offli
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
```

7. Tokenization

1. combine title+text
2. text only

```
def tokenize(batch):
  return tinybert_tokenizer(batch['title'], batch['text'], truncation=True,padding=True,max_length = 512)
```

```
tokenized_dataframe = dataset.map(tokenize, batched=True, batch_size=None)
tokenized_dataframe
```

```
Map: 100%                                      12796/12796 [01:14<00:00, 171.95 examples/s]

Map: 100%                                      3656/3656 [00:20<00:00, 183.33 examples/s]

Map: 100%                                      1829/1829 [00:10<00:00, 171.43 examples/s]
DatasetDict({
    train: Dataset({
        features: ['id', 'title', 'author', 'text', 'labels', 'title_tokens', 'text_tokens', 'input_ids', 'token_type_ids',
'attention_mask'],
        num_rows: 12796
    })
    test: Dataset({
        features: ['id', 'title', 'author', 'text', 'labels', 'title_tokens', 'text_tokens', 'input_ids', 'token_type_ids',
'attention_mask'],
        num_rows: 3656
    })
    validation: Dataset({
```

```
print(dataset['train'][0],"\n")
print(tokenize(dataset['train'][0]))
```

```
{'id': 20451, 'title': 'Donald Trump Gettysburg Address RECAP', 'author': 'Truth Broadcast Network', 'text': "7 hours ago 3

{'input_ids': [101, 6221, 8398, 22577, 4769, 28667, 9331, 102, 1021, 2847, 3283, 1017, 4311, 2006, 2054, 2017, 2342, 2000, 2
```

```
final_dataset = tokenized_dataframe.remove_columns(['title','text','id','author'])
final_dataset
```

```
DatasetDict({
    train: Dataset({
        features: ['labels', 'title_tokens', 'text_tokens', 'input_ids', 'token_type_ids', 'attention_mask'],
        num_rows: 12796
    })
    test: Dataset({
        features: ['labels', 'title_tokens', 'text_tokens', 'input_ids', 'token_type_ids', 'attention_mask'],
        num_rows: 3656
    })
    validation: Dataset({
        features: ['labels', 'title_tokens', 'text_tokens', 'input_ids', 'token_type_ids', 'attention_mask'],
        num_rows: 1829
    })
})
```

8. Compute metrics

```
accuracy = evaluate.load('accuracy')

def compute_metrics(eval_pred):
  predictions,labels = eval_pred
  predictions = np.argmax(predictions, axis=1)
  return accuracy.compute(predictions=predictions, references=labels)
```

```
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
Downloading builder script:      4.20k/? [00:00<00:00, 82.7kB/s]
```

9. Training args and trainer - TinyBERT

```
batch_size = 32
training_dir = "train_dir"

training_args = TrainingArguments(
                                  output_dir=training_dir,
                                  overwrite_output_dir = True,
                                  eval_strategy = 'epoch',
                                  num_train_epochs = 3,
                                  learning_rate = 2e-5,
                                  per_device_train_batch_size = batch_size,
                                  per_device_eval_batch_size = batch_size,
                                  weight_decay = 0.01,
)
```

```
trainer = Trainer(
    model = tinybert_model,
    args = training_args,
    train_dataset = final_dataset['train'],
    eval_dataset = final_dataset['validation'],
    tokenizer = tinybert_tokenizer,
    compute_metrics = compute_metrics,
)
```

```
/tmp/ipython-input-1295999482.py:1: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for `Trair
  trainer = Trainer(
```

10. Model Training

```
trainer.train()
```

```
/usr/local/lib/python3.12/dist-packages/notebook/notebookapp.py:191: SyntaxWarning: invalid escape sequence '\/'
  | |_| | '_ \/ _` / _` |  _/ -_)
wandb: (1) Create a W&B account
wandb: (2) Use an existing W&B account
wandb: (3) Don't visualize my results
wandb: Enter your choice: 1
wandb: You chose 'Create a W&B account'
wandb: Create an account here: https://wandb.ai/authorize?signup=true&ref=models
wandb: Paste an API key from your profile and hit enter: ··········
wandb: No netrc file found, creating one.
wandb: Appending key for api.wandb.ai to your netrc file: /root/.netrc
wandb: Currently logged in as: pavan220405 (pavan220405-iit-ropar-tif) to https://api.wandb.ai. Use `wandb login --relogin`
```

Tracking run with wandb version 0.23.1
Run data is saved locally in /content/wandb/run-20251216_083523-awimk8xc
Syncing run **visionary-fog-8** to Weights & Biases (docs)
View project at https://wandb.ai/pavan220405-iit-ropar-tif/huggingface
View run at https://wandb.ai/pavan220405-iit-ropar-tif/huggingface/runs/awimk8xc

[1200/1200 07:53, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Accuracy
1	No log	0.046162	0.989612
2	0.218000	0.016922	0.996720
3	0.017500	0.015735	0.997266

```
TrainOutput(global_step=1200, training_loss=0.09965561777353286, metrics={'train_runtime': 494.3633,
'train_samples_per_second': 77.651, 'train_steps_per_second': 2.427, 'total_flos': 550445387046912.0, 'train_loss':
```

11. Model Evaluation

```
preds = trainer.predict(final_dataset['test'])
preds.metrics
```

```
{'test_loss': 0.011287638917565346,
 'test_accuracy': 0.9978118161925602,
 'test_runtime': 13.0795,
 'test_samples_per_second': 279.521,
 'test_steps_per_second': 8.792}
```

```
from sklearn.metrics import accuracy_score

y_preds = np.argmax(preds.predictions, axis=1)
y_true = dataset['test']['labels'][:]

print(accuracy_score(y_true, y_preds))
```

```
0.9978118161925602
```

12. Comparing with DistilBERT

```
distilbert_dir = '/content/drive/MyDrive/offline_models/distilbert-base-uncased'

distilbert_tokenizer = AutoTokenizer.from_pretrained(distilbert_dir,local_files_only=True)
distilbert_config = AutoConfig.from_pretrained(distilbert_dir,local_files_only=True,num_labels=2,id2label=id2label,label2
distilbert_model = AutoModelForSequenceClassification.from_pretrained(distilbert_dir,local_files_only=True,config = disti

def distil_tokenize(batch):
  return distilbert_tokenizer(batch['title'], batch['text'], truncation=True,padding=True,max_length = 512)
```

```
dataset_distil = dataset.map(distil_tokenize, batched=True, batch_size=None)
final_dataset_distil = dataset_distil.remove_columns(['title','text','id','author'])

trainer_distil = Trainer(
    model = distilbert_model,
    args = training_args,
    train_dataset = final_dataset_distil['train'],
    eval_dataset = final_dataset_distil['validation'],
    tokenizer = distilbert_tokenizer,
    compute_metrics = compute_metrics,
)

trainer_distil.train()
```

```
Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at /content/drive/MyDri
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Map: 100%                                     12796/12796 [00:47<00:00, 272.64 examples/s]
Map: 100%                                     3656/3656 [00:09<00:00, 368.84 examples/s]
Map: 100%                                     1829/1829 [00:06<00:00, 301.91 examples/s]
/tmp/ipython-input-81550133.py:13: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for `Tra
  trainer_distil = Trainer(
                                      [1200/1200 30:41, Epoch 3/3]
```

Epoch	Training Loss	Validation Loss	Accuracy
1	No log	0.009744	0.997813
2	0.059400	0.014965	0.996173
3	0.004500	0.008891	0.997813

```
TrainOutput(global_step=1200, training_loss=0.027100048462549844, metrics={'train_runtime': 1842.8176,
'train_samples_per_second': 20.831, 'train_steps_per_second': 0.651, 'total_flos': 5085158499606528.0, 'train_loss':
0.027100048462549844, 'epoch': 3.0})
```

```
preds_distil = trainer_distil.predict(final_dataset_distil['test'])
preds_distil.metrics
```

```
{'test_loss': 0.002549974247813225,
 'test_accuracy': 0.99945295404814,
 'test_runtime': 56.9653,
 'test_samples_per_second': 64.179,
 'test_steps_per_second': 2.019}
```

```
y_preds_distil = np.argmax(preds_distil.predictions, axis=1)
y_true_distil = dataset['test']['labels'][:]

print(accuracy_score(y_true_distil, y_preds_distil))
```

```
0.99945295404814
```

Comparision

```
comparision_table = pd.DataFrame({
    'Models' : ['DistilBERT','TinyBERT'],
    'Test Accuracy' : [preds_distil.metrics['test_accuracy'],preds.metrics['test_accuracy']],
    'Training Time' : [1842.8176,494.3633],
    'Inference Time' : [preds_distil.metrics['test_runtime'],preds.metrics['test_runtime']]
})

comparision_table
```

	Models	Test Accuracy	Training Time	Inference Time
0	DistilBERT	0.999453	1842.8176	56.9653
1	TinyBERT	0.997812	494.3633	13.0795