1 {"metadata":{"kernelspec":{"language":"python"," display_name":"Python 3","name":"python3"}," language_info":{"name":"python","version":"3.7.12"," mimetype":"text/x-python","codemirror_mode":{"name":" ipython", "version":3}, "pygments_lexer":"ipython3", " nbconvert_exporter":"python","file_extension":".py "}}, "nbformat_minor":4, "nbformat":4, "cells":[{" cell_type":"code","source":"# This Python 3 environment comes with many helpful analytics libraries installed\n# It is defined by the kaggle/ python Docker image: https://github.com/kaggle/docker -python\n# For example, here's several helpful packages to load\n\nimport numpy as np # linear algebra\nimport pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)\n\n# Input data files are available in the read-only \"../input/\" directory\n # For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory\n\nimport os\nfor dirname, _, filenames in os.walk('/kaggle/input'):\n filename in filenames:\n print(os.path.ioin(dirname, filename))\n\n# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using \"Save & Run All\" \n# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session", "metadata": {"_uuid":" 8f2839f25d086af736a60e9eeb907d3b93b6e0e5","_cell_guid ":"b1076dfc-b9ad-4769-8c92-a6c4dae69d19","execution ":{"iopub.status.busy":"2023-01-07T14:39:34.810845Z ","iopub.execute_input":"2023-01-07T14:39:34.811739Z ","iopub.status.idle":"2023-01-07T14:39:34.845561Z"," shell.execute_reply.started":"2023-01-07T14:39:34. 811617Z", "shell.execute_reply": "2023-01-07T14:39:34. 844320Z"}, "trusted":true}, "execution_count":1, " outputs":[{"name":"stdout","text":"/kaggle/input/imdb -dataset-of-50k-movie-reviews/IMDB Dataset.csv\n"," output_type":"stream"}]},{"cell_type":"code","source ":"df = pd.read_csv('/kagqle/input/imdb-dataset-of-50k-movie-reviews/IMDB Dataset.csv')", "metadata":{" execution":{"iopub.status.busy":"2023-01-07T14:39:54.

```
1 975371Z", "iopub.execute_input": "2023-01-07T14:39:54.
 975796Z", "iopub.status.idle": "2023-01-07T14:39:56.
 879528Z", "shell.execute_reply.started": "2023-01-07T14
 :39:54.975758Z", "shell.execute_reply": "2023-01-07T14:
 39:56.878400Z"}, "trusted":true}, "execution_count":2, "
 outputs":[]},{"cell_type":"code","source":"df.head
 ()","metadata":{"execution":{"iopub.status.busy":"
 2023-01-07T14:39:59.752890Z","iopub.execute_input":"
 2023-01-07T14:39:59.754202Z", "iopub.status.idle":"
 2023-01-07T14:39:59.775594Z", "shell.execute_reply.
 started": "2023-01-07T14:39:59.754155Z", "shell.
 execute_reply":"2023-01-07T14:39:59.773811Z"},"
 trusted":true}, "execution_count":3, "outputs":[{"
 execution_count":3, "output_type": "execute_result", "
 data":{"text/plain
 review sentiment\n0 One of the other reviewers has
 mentioned that ... positive\n1 A wonderful little
 production. <br /><br />The... positive\n2 I
 thought this was a wonderful way to spend ti...
 positive\n3 Basically there's a family where a
 little boy ...
               negative\n4 Petter Mattei's \"Love
 in the Time of Money\" is... positive", "text/html
 ":"<div>\n<style scoped>\n .dataframe tbody tr th:
                       vertical-align: middle;\n
 only-of-type {\n
             .dataframe tbody tr th {\n
 vertical-align: top;\n
                         }\n\n
                               .dataframe thead
              text-align: right;\n }\n</style>\n
 th {\n
 table border=\"1\" class=\"dataframe\">\n <thead>\n
     \n
         review\n
 >\n
                              sentiment\n
     \n </thead>\n \n
                                    \n
 >0\n
               One of the other reviewers has
 mentioned that ...
                             positive
               \n
                        1
     \n
 wonderful little production. <br /&gt;&lt;br /&gt;
 The...
                  positive
                                       \n
                             I thought this
     \n
                2

 was a wonderful way to spend ti...
 positive\n
                  \n
                           \n
                                     3
        Basically there's a family where a little
```

negative\n \n 1 boy ...\n \n 4 Petter Mattei's \"Love in the Time of Money\" is... \n \n\n</div positive\n >"}, "metadata":{}}]}, {"cell_type":"code", "source":"df ['review'][0]", "metadata": {"execution": {"iopub.status .busy":"2023-01-07T14:40:29.311312Z","iopub. execute_input":"2023-01-07T14:40:29.311710Z","iopub. status.idle":"2023-01-07T14:40:29.318991Z","shell. execute_reply.started":"2023-01-07T14:40:29.311677Z ", "shell.execute_reply": "2023-01-07T14:40:29.318058Z "},"trusted":true},"execution_count":5,"outputs":[{" execution_count":5,"output_type":"execute_result"," data":{"text/plain":"\"One of the other reviewers has mentioned that after watching just 1 Oz episode you' ll be hooked. They are right, as this is exactly what happened with me.

The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in the classic use of the word.

It is called OZ as that is the nickname given to the Oswald Maximum Security State Penitentary. It focuses mainly on Emerald City, an experimental section of the prison where all the cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many...Aryans, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings and shady agreements are never far away.
sbr />I would say the main appeal of the show is due to the fact that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget charm, forget romance...OZ doesn't mess around . The first episode I ever saw struck me as so nasty it was surreal, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to the high levels of graphic violence. Not just violence, but injustice (crooked

```
1 guards who'll be sold out for a nickel, inmates who'
  ll kill on order and get away with it, well mannered
  , middle class inmates being turned into prison
  bitches due to their lack of street skills or prison
  experience) Watching Oz, you may become comfortable
  with what is uncomfortable viewing....thats if you
 can get in touch with your darker side.\""}, "metadata
  ":{}}]},{"cell_type":"code","source":"df.info()","
  metadata":{"execution":{"iopub.status.busy":"2023-01-
  07T15:00:01.967084Z", "iopub.execute_input": "2023-01-
  07T15:00:01.967803Z", "iopub.status.idle": "2023-01-
  07T15:00:02.001320Z", "shell.execute_reply.started":"
  2023-01-07T15:00:01.967750Z", "shell.execute_reply":"
  2023-01-07T15:00:02.000065Z"}, "trusted":true}, "
  execution_count":6,"outputs":[{"name":"stdout","text
  ":"<class 'pandas.core.frame.DataFrame'>\nRangeIndex
  : 50000 entries, 0 to 49999\nData columns (total 2
                             Non-Null Count
  columns):\n # Column
                                              Dtype \n
                  _____
                                   ---- \n 0
                                                review
       50000 non-null object\n 1
                                     sentiment
            object\ndtypes: object(2)\nmemory usage:
  non-null
  781.4+ KB\n", "output_type": "stream"}]}, {"cell_type": "
  code", "source": "df.duplicated().sum()", "metadata":{"
  execution":{"iopub.status.busy":"2023-01-07T15:00:22.
  405132Z", "iopub.execute_input": "2023-01-07T15:00:22.
 406451Z", "iopub.status.idle": "2023-01-07T15:00:22.
  602368Z", "shell.execute_reply.started": "2023-01-07T15
  :00:22.406406Z", "shell.execute_reply": "2023-01-07T15:
  00:22.601149Z"}, "trusted":true}, "execution_count":7, "
  outputs":[{"execution_count":7,"output_type":"
  execute_result","data":{"text/plain":"418"},"metadata
  ":{}}]},{"cell_type":"code","source":"df = df.
  drop_duplicates()","metadata":{"execution":{"iopub.
  status.busy":"2023-01-07T15:00:48.108662Z","iopub.
  execute_input": "2023-01-07T15:00:48.109066Z", "iopub.
  status.idle":"2023-01-07T15:00:48.293952Z","shell.
  execute_reply.started":"2023-01-07T15:00:48.109032Z
  ", "shell.execute_reply": "2023-01-07T15:00:48.292741Z
  "}, "trusted":true}, "execution_count":9, "outputs
  ":[]},{"cell_type":"code","source":"df = df.iloc[:
  10000]", "metadata": {"execution": {"iopub.status.busy
```

```
1 ":"2023-01-07T15:01:56.330858Z", "iopub.execute_input
 ":"2023-01-07T15:01:56.331316Z", "iopub.status.idle":"
 2023-01-07T15:01:56.337229Z", "shell.execute_reply.
 started": "2023-01-07T15:01:56.331270Z", "shell.
 execute_reply":"2023-01-07T15:01:56.335746Z"},"
 trusted":true}, "execution_count":12, "outputs":[]}, {"
 cell_type":"code","source":"# Cleaning\n\n# lower
 case\ndf.head()","metadata":{"execution":{"iopub.
 status.busy":"2023-01-07T15:04:42.967125Z","iopub.
 execute_input":"2023-01-07T15:04:42.967535Z","iopub.
 status.idle": "2023-01-07T15:04:42.980358Z", "shell.
 execute_reply.started":"2023-01-07T15:04:42.967503Z
 ", "shell.execute_reply": "2023-01-07T15:04:42.979161Z
 "}, "trusted":true}, "execution_count":13, "outputs":[{"
 execution_count":13,"output_type":"execute_result","
 data":{"text/plain
 review sentiment\n0 One of the other reviewers has
 mentioned that ...
                   positive\n1 A wonderful little
 production. <br /><br />The... positive\n2
 thought this was a wonderful way to spend ti...
 positive\n3 Basically there's a family where a
 little boy ... negative\n4 Petter Mattei's \"Love
 in the Time of Money\" is... positive","text/html
 ":"<div>\n<style scoped>\n .dataframe tbody tr th:
 only-of-type {\n
                       vertical-align: middle;\n
     }\n\n
             .dataframe tbody tr th {\n
 vertical-align: top;\n
                                 .dataframe thead
                         }\n\n
 th \{ \n
              text-align: right;\n }\n</style>\n
 table border=\"1\" class=\"dataframe\">\n <thead>\n
     \n
                                          </th
 >\n
         review\n
                              sentiment
     \n </thead>\n \n
                                    \n
               One of the other reviewers has
 >0\n
 mentioned that ...
                             positive
     \n
               \n
                       1
                                          A
 wonderful little production. <br /&qt;&lt;br /&qt;
                   positive
 The...
                                       \n
     \n
                2\n
                                I thought this
 was a wonderful way to spend ti...\n
                                            3\
 positive\n
                  \n
                             \n
```

```
Basically there's a family where a little
boy ...\n
                   negative
                                          \n
   \n
               4\n
                                 Petter Mattei's
\"Love in the Time of Money\" is...\n
positive\n
                  \n \n\n</div
>"}, "metadata":{}}]}, {"cell_type":"code", "source":"df
['review'] = df['review'].str.lower()","metadata":{"
execution":{"iopub.status.busy":"2023-01-07T15:05:51.
478800Z", "iopub.execute_input": "2023-01-07T15:05:51.
480009Z", "iopub.status.idle": "2023-01-07T15:05:51.
515678Z", "shell.execute_reply.started": "2023-01-07T15
:05:51.479953Z", "shell.execute_reply": "2023-01-07T15:
05:51.514369Z"}, "trusted":true}, "execution_count":15
,"outputs":[{"name":"stderr","text":"/opt/conda/lib/
python3.7/site-packages/ipykernel_launcher.py:1:
SettingWithCopyWarning: \nA value is trying to be set
 on a copy of a slice from a DataFrame.\nTry using .
loc[row_indexer,col_indexer] = value instead\n\nSee
the caveats in the documentation: https://pandas.
pydata.org/pandas-docs/stable/user_quide/indexing.
html#returning-a-view-versus-a-copy\n \"\"\"Entry
point for launching an IPython kernel.\n","
output_type":"stream"}]},{"cell_type":"code","source
":"df.head()", "metadata":{"execution":{"iopub.status.
busy":"2023-01-07T15:05:55.574585Z","iopub.
execute_input":"2023-01-07T15:05:55.575148Z","iopub.
status.idle":"2023-01-07T15:05:55.589399Z","shell.
execute_reply.started":"2023-01-07T15:05:55.575092Z
", "shell.execute_reply": "2023-01-07T15:05:55.588214Z
"}, "trusted":true}, "execution_count":16, "outputs":[{"
execution_count":16, "output_type": "execute_result", "
data":{"text/plain
                    one of the other reviewers has
review sentiment\n0
mentioned that ...
                   positive\n1
                                a wonderful little
production. <br /><br />the...
                               positive\n2 i
thought this was a wonderful way to spend ti...
positive\n3 basically there's a family where a
little boy ...
               negative\n4 petter mattei's \"love
in the time of money\" is... positive","text/html
":"<div>\n<style scoped>\n
                             .dataframe thody tr th:
```

```
vertical-align: middle;\n
1 only-of-type {\n
             .dataframe tbody tr th {\n
     }\n\n
 vertical-align: top;\n
                        }\n\n
                               .dataframe thead
 th \{ \n
              text-align: right;\n
                                   }\n</style>\n
 table border=\"1\" class=\"dataframe\">\n <thead>\n
     \n
                                         </th
         review\n
                             sentiment
 >\n
     \n
             </thead>\n <tbody>\n
                                   \n
              one of the other reviewers has
 >0
 mentioned that ...
                            positive
     \n
              \n
                         1
                                         a
 wonderful little production. <br /&qt;&lt;br /&qt;
 the...
                  positive
                                      \n
     \n
               2
                               i thought this
 was a wonderful way to spend ti...
                                           \n
 positive\n
                  \n
                                      3\
       basically there's a family where a little
                   negative
 boy ...\n
                                       \n
     \n
               4
                               petter mattei's
  \"love in the time of money\" is...\n
 positive\n
                  \n \n\n</div
 >"}, "metadata":{}}]}, {"cell_type":"code", "source":"#
 remove whitespaces\ndf['review'] = df['review'].str.
 strip()","metadata":{"execution":{"iopub.status.busy
 ":"2023-01-07T15:07:20.669811Z","iopub.execute_input
 ":"2023-01-07T15:07:20.670232Z","iopub.status.idle":"
 2023-01-07T15:07:20.683166Z", "shell.execute_reply.
 started": "2023-01-07T15:07:20.670195Z", "shell.
 execute_reply":"2023-01-07T15:07:20.681847Z"},"
 trusted":true}, "execution_count":18, "outputs":[{"name
 ":"stderr", "text": "/opt/conda/lib/python3.7/site-
 packages/ipykernel_launcher.py:2:
 SettingWithCopyWarning: \nA value is trying to be set
  on a copy of a slice from a DataFrame.\nTry using .
 loc[row_indexer,col_indexer] = value instead\n\nSee
 the caveats in the documentation: https://pandas.
 pydata.org/pandas-docs/stable/user_quide/indexing.
 html#returning-a-view-versus-a-copy\n
 output_type":"stream"}]}, {"cell_type":"code", "source
 ":"df.head()", "metadata":{"execution":{"iopub.status.
 busy":"2023-01-07T15:07:26.167513Z","iopub.
```

```
1 execute_input":"2023-01-07T15:07:26.168031Z","iopub.
 status.idle": "2023-01-07T15:07:26.181478Z", "shell.
 execute_reply.started":"2023-01-07T15:07:26.167995Z
 ", "shell.execute_reply": "2023-01-07T15:07:26.180323Z
 "}, "trusted":true}, "execution_count":19, "outputs":[{"
 execution_count":19, "output_type": "execute_result", "
 data":{"text/plain
 review sentiment\n0 one of the other reviewers has
 mentioned that ... positive\n1 a wonderful little
 production. <br /><br />the...
                             positive\n2
 thought this was a wonderful way to spend ti...
 positive\n3 basically there's a family where a
 little boy ... negative\n4 petter mattei's \"love
 in the time of money\" is... positive","text/html
 ":"<div>\n<style scoped>\n .dataframe tbody tr th:
 only-of-type {\n
                      vertical-align: middle;\n
     }\n\n
             .dataframe tbody tr th {\n
 vertical-align: top;\n
                        }\n\n
                                .dataframe thead
             text-align: right;\n }\n</style>\n
 th {\n
 table border=\"1\" class=\"dataframe\">\n <thead>\n
     \n
                                        </th
         review\n
 >\n
                             sentiment\n
     \n </thead>\n \n
                                  \n
              one of the other reviewers has
 >0\n
 mentioned that ...
                            positive
                       1
     \n
              \n
 wonderful little production. <br /&qt;&lt;br /&qt;
 the...
                  positive
                                      \n
     \n
               2\n
                               i thought this
 was a wonderful way to spend ti...
 positive\n
                  \n
                           \n
                                      3\
       basically there's a family where a little
                   negative
 boy ...\n
     \n
               4
                               petter mattei's
  \"love in the time of money\" is...\n
 positive\n \n \n\n</div</pre>
 >"}, "metadata":{}}]}, {"cell_type":"code", "source":"#
 remove html tags\ndf['review'] = df['review'].str.
 replace(r'<.*?>','')","metadata":{"execution":{"iopub
  .status.busy":"2023-01-07T15:14:03.025370Z","iopub.
```

```
1 execute_input":"2023-01-07T15:14:03.025772Z","iopub.
  status.idle": "2023-01-07T15:14:03.069458Z", "shell.
  execute_reply.started":"2023-01-07T15:14:03.025741Z
  ", "shell.execute_reply": "2023-01-07T15:14:03.068007Z
  "},"trusted":true},"execution_count":24,"outputs":[{"
  name":"stderr","text":"/opt/conda/lib/python3.7/site-
  packages/ipykernel_launcher.py:2: FutureWarning: The
  default value of regex will change from True to False
   in a future version.\n \n/opt/conda/lib/python3.7/
  site-packages/ipykernel_launcher.py:2:
  SettingWithCopyWarning: \nA value is trying to be set
   on a copy of a slice from a DataFrame.\nTry using .
  loc[row_indexer,col_indexer] = value instead\n\nSee
  the caveats in the documentation: https://pandas.
  pydata.org/pandas-docs/stable/user_quide/indexing.
  html#returning-a-view-versus-a-copy\n \n","
  output_type":"stream"}]},{"cell_type":"code","source
  ":"df.head()", "metadata":{"execution":{"iopub.status.
  busy": "2023-01-07T15:14:07.727186Z", "iopub.
  execute_input":"2023-01-07T15:14:07.727586Z","iopub.
  status.idle": "2023-01-07T15:14:07.738922Z", "shell.
  execute_reply.started":"2023-01-07T15:14:07.727554Z
  ", "shell.execute_reply": "2023-01-07T15:14:07.737745Z
  "}, "trusted":true}, "execution_count":25, "outputs":[{"
  execution_count":25,"output_type":"execute_result","
  data":{"text/plain
  " : "
  review sentiment\n0 one of the other reviewers has
  mentioned that ... positive\n1 a wonderful little
  production. the filming tec...
                                 positive\n2
  thought this was a wonderful way to spend ti...
  positive\n3 basically there's a family where a
  little boy ...
                 negative\n4 petter mattei's \"love
  in the time of money\" is... positive", "text/html
  ":"<div>\n<style scoped>\n .dataframe tbody tr th:
  only-of-type {\n
                         vertical-align: middle;\n
               .dataframe tbody tr th {\n
      }\n\n
                                   .dataframe thead
  vertical-align: top;\n
                            }\n\n
                                       }\n</style>\n<</pre>
               text-align: right;\n
  table border=\"1\" class=\"dataframe\">\n <thead>\n
      \n
                                               </th
```

```
sentiment
       review\n
   \n </thead>\n
                      \n
                                 \n
                                            <th
>0\n
             one of the other reviewers has
mentioned that ...
                           positive
             \n
                       1
   \n
wonderful little production. the filming tec...
      positive
                          \n
                                    \n
n
                     i thought this was a
     2\n
wonderful way to spend ti...
                                    positive
\n
                   \n
                              3

td>basically there's a family where a little boy
               negative
 ...\n
                                   \n
         4\n
                         petter mattei's \"
tr>\n
love in the time of money\" is...
                \n \n\n</div
positive\n
>"}, "metadata":{}}]}, {"cell_type":"code", "source":"#
http://campusx.in www.campusx.in www.campusx.co.in\n\
ndef remove_url(data):\n data=re.sub(r\"https
?://\\S+|www\\.\\S+\",'',data)\n
                               return data","
metadata":{"execution":{"iopub.status.busy":"2023-01-
07T15:18:02.916624Z", "iopub.execute_input": "2023-01-
07T15:18:02.917643Z", "iopub.status.idle": "2023-01-
07T15:18:02.923258Z", "shell.execute_reply.started":"
2023-01-07T15:18:02.917599Z", "shell.execute_reply":"
2023-01-07T15:18:02.922079Z"}, "trusted":true}, "
execution_count":31, "outputs":[]}, {"cell_type":"code
", "source": "# code to remove a URL\ndf['review'] = df
['review'].str.replace(r\"https?://\\S+|www\\.\\S
+\",'')","metadata":{"execution":{"iopub.status.busy
":"2023-01-07T15:21:08.832002Z","iopub.execute_input
":"2023-01-07T15:21:08.832433Z","iopub.status.idle":"
2023-01-07T15:21:08.953450Z", "shell.execute_reply.
started": "2023-01-07T15:21:08.832389Z", "shell.
execute_reply":"2023-01-07T15:21:08.952096Z"},"
trusted":true}, "execution_count":43, "outputs":[{"name
```