544_project_data_preparation

December 1, 2021

1 Data Preparation For Bart

[ ]:

```
from google.colab import drive
drive.mount('/content/drive')
root_dir = "/content/drive/MyDrive/Colab Notebooks/544_bart/"
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
drive.mount("/content/drive", force_remount=True).
```

[ ]:

```
!pip install rake-nltk
import pandas as pd
import numpy as np
from rake_nltk import Rake
import nltk

nltk.download('stopwords')
nltk.download('punkt')

from nltk.corpus import stopwords
```

```
Requirement already satisfied: rake-nltk in /usr/local/lib/python3.7/dist-
packages (1.0.6)
Requirement already satisfied: nltk<4.0.0,>=3.6.2 in /usr/local/lib/python3.7
/dist-packages (from rake-nltk) (3.6.5)
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages
(from nltk<4.0.0,>=3.6.2->rake-nltk) (4.62.3)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages
(from nltk<4.0.0,>=3.6.2->rake-nltk) (1.1.0)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.7/dist-
packages (from nltk<4.0.0,>=3.6.2->rake-nltk) (2021.11.10)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages
(from nltk<4.0.0,>=3.6.2->rake-nltk) (7.1.2)
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

[ ]:

```
df = pd.read_csv(root_dir + 'ROCStories_winter2017 - ROCStories_winter2017.csv')
df
```

[ ]:

```
                                    storyid  ...
sentence5
0      8bbe6d11-1e2e-413c-bf81-eaea05f4f1bd  ...  After a few weeks, he started
to feel much bet...
1      0beabab2-fb49-460e-a6e6-f35a202e3348  ...  Tom sat on his couch filled
with regret about ...
2      87da1a22-df0b-410c-b186-439700b70ba6  ...  Marcus was happy to have the
right clothes for...
3      2d16bcd6-692a-4fc0-8e7c-4a6f81d9efa9  ...  He ended up buying the truck
he wanted despite...
4      c71bb23b-7731-4233-8298-76ba6886cee1  ...      His congregation was
delighted and so was he.
...                                     ...  ...
...
52660  134e8636-3617-43d8-ba6a-9a11b3b115b1  ...       The owner of the flavor
sold him the recipe.
52661  4c317f76-ca42-4024-a4c2-12ec911cf89b  ...                     I skipped
detention all week.
52662  a18fd0d2-4d0c-4316-befe-e3d827fe699b  ...  She breaks her wrist in the
process and goes t...
52663  2c14252b-4080-4fca-8765-537772018508  ...  Jamie gets married and they
spent the rest of ...
52664  19d92cee-110d-4cd3-90b3-fe9aee5577f6  ...                       That
orange broke her nose.

[52665 rows x 7 columns]
```

[ ]:

```
# df2 = pd.read_csv(root_dir + 'ROCStories__spring2016 - ROCStories_spring2016.
 ↪csv')
# df2
```

[ ]:

```
# df = df.append(df2, ignore_index=True)
# df
```

[ ]:

```
# Uses stopwords for english from NLTK, and all punctuation characters by
# default
#max_length=1 so that we only get one word
r = Rake(max_length=1, min_length=1, include_repeated_phrases=False)

def rake_implement(s1, s2, s3, s4, s5, r):
    result = []
    result.append(get_word(r, s1))
    result.append(get_word(r, s2))
    result.append(get_word(r, s3))
    result.append(get_word(r, s4))
    result.append(get_word(r, s5))
    return ' '.join(result)

def get_word(r, s):
  r.extract_keywords_from_text(s)
  phrases = r.get_ranked_phrases()
  if(len(phrases) == 0):
    #if nothing is extracted, take first word that is not in stopwords (from␣
 ↪the middle)
    sentence = s.split()
    for i in range(int(len(sentence)/2), len(sentence)):
      if not sentence[i] in stopwords.words('english'):
        return sentence[i].lower()
  else:
    return phrases[0]

#https://stackoverflow.com/questions/56836477/
 ↪apply-nltk-rake-to-each-row-in-dataframe
df['storyline'] = df.apply(lambda x: rake_implement(x.sentence1, x.sentence2, x.
 ↪sentence3, x.sentence4, x.sentence5, r), axis=1)
```

[ ]:

```
df
```

[ ]:

```
                                    storyid  ...
storyline
0      8bbe6d11-1e2e-413c-bf81-eaea05f4f1bd  ...         put try realized
started weeks
1      0beabab2-fb49-460e-a6e6-f35a202e3348  ...              tom angry wall tom
regret
2      87da1a22-df0b-410c-b186-439700b70ba6  ...  business formal pair perfectly
marcus
3      2d16bcd6-692a-4fc0-8e7c-4a6f81d9efa9  ...         trailer needed much
ways truck
4      c71bb23b-7731-4233-8298-76ba6886cee1  ...    pastor tried sing sundays
delighted
...                                     ...  ...
...
52660  134e8636-3617-43d8-ba6a-9a11b3b115b1  ...         flavor tried get recipe
recipe
52661  4c317f76-ca42-4024-a4c2-12ec911cf89b  ...            trouble day found
told week
52662  a18fd0d2-4d0c-4316-befe-e3d827fe699b  ...        janice legs working
wrist wrist
52663  2c14252b-4080-4fca-8765-537772018508  ...       jamie married man
marrying spent
52664  19d92cee-110d-4cd3-90b3-fe9aee5577f6  ...                tree hit tree
tree nose

[52665 rows x 8 columns]
```

```
[ ]: def combineSen(row):
       return row['sentence1'] + row['sentence2'] + row['sentence3'] +␣
      ↪row['sentence4'] + row['sentence5']
     df['story'] = df.apply (lambda row: combineSen(row), axis=1)
     df = df[['storytitle', 'storyline', 'story']]
     df
```

```
[ ]:                     storytitle  ...
     story
     0       David Drops the Weight  ...  David noticed he had put on a lot of weight
     re...
     1                  Frustration  ...  Tom had a very short temper.One day a guest
     ma...
     2           Marcus Buys Khakis  ...  Marcus needed clothing for a business
     casual e...
     3           Different Opinions  ...  Bobby thought Bill should buy a trailer and
     ha...
     4      Overcoming shortcomings  ...  John was a pastor with a very bad memory.He
     tr...
     ...                        ...  ...
     ...
     52660                   Flavor  ...  The man liked the flavor.He tried to
     recreate ...
     52661              After Death  ...  After my friend's dad's funeral, I got in
     trou...
     52662  Janice breaks her wrist  ...  Janice was out exercising for her big
     soccer g...
     52663   Jamie marries for love  ...  Jamie is an american girl.Jamie wants to
     get m...
     52664                   Orange  ...  The orange fell from the tree.It hit a girl
     on...

     [52665 rows x 3 columns]
```

```
[ ]: def concat_title_line(x):
         l = x['storytitle'].split()
         l.append('<EOT>')
         l.extend(x['storyline'].split())
         return ' '.join(l)

     df['titleLineConcat'] = df.apply(lambda x: concat_title_line(x), axis=1)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:7:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
```

```
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  import sys
```

```
[ ]: df['titleLineConcat'][0]
```

```
[ ]: 'David Drops the Weight <EOT> put try realized started weeks'
```

```
[ ]: df
```

```
[ ]:                       storytitle  ...
     titleLineConcat
     0       David Drops the Weight  ...  David Drops the Weight <EOT> put try
     realized ...
     1                  Frustration  ...        Frustration <EOT> tom angry wall tom
     regret
     2           Marcus Buys Khakis  ...  Marcus Buys Khakis <EOT> business formal
     pair ...
     3           Different Opinions  ...  Different Opinions <EOT> trailer needed
     much w...
     4      Overcoming shortcomings  ...  Overcoming shortcomings <EOT> pastor tried
     sin...
     ...                         ...  ...
     ...
     52660                   Flavor  ...        Flavor <EOT> flavor tried get recipe
     recipe
     52661              After Death  ...      After Death <EOT> trouble day found
     told week
     52662   Janice breaks her wrist  ...  Janice breaks her wrist <EOT> janice legs
     work...
     52663    Jamie marries for love  ...  Jamie marries for love <EOT> jamie married
     man...
     52664                   Orange  ...              Orange <EOT> tree hit tree
     tree nose

     [52665 rows x 4 columns]
```

```
[ ]: df.to_csv(root_dir+'train_title_line_story.csv', encoding='utf-8', index=False)
```

```
[ ]: #output notebook as pdf
     !wget -nc https://raw.githubusercontent.com/brpy/colab-pdf/master/colab_pdf.py
     from colab_pdf import colab_pdf
     colab_pdf('544_project_data_preparation.ipynb')
```

```
--2021-12-01 05:22:26--  https://raw.githubusercontent.com/brpy/colab-
pdf/master/colab_pdf.py
Resolving raw.githubusercontent.com (raw.githubusercontent.com)...
185.199.110.133, 185.199.109.133, 185.199.111.133, ...
Connecting to raw.githubusercontent.com
(raw.githubusercontent.com)|185.199.110.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
```

```
Length: 1864 (1.8K) [text/plain]
Saving to: colab_pdf.py

colab_pdf.py        100%[===================>]   1.82K  --.-KB/s    in 0s

2021-12-01 05:22:26 (28.6 MB/s) - colab_pdf.py saved [1864/1864]

Mounted at /content/drive/

WARNING: apt does not have a stable CLI interface. Use with caution in scripts.


WARNING: apt does not have a stable CLI interface. Use with caution in scripts.

Extracting templates from packages: 100%
```