NLP LAB 02 Théo Ripoll - Quentin Fish - Nicolas Fidel

Part 1: The Dataset

In [8]:
```
from datasets import load_dataset
from datasets import get_dataset_split_names
import pandas as pd
```

In [9]:
```
dataset = load_dataset("imdb")
dataset
```

```
Found cached dataset imdb (/Users/quentinfisch/.cache/huggingface/datasets/
imdb/plain_text/1.0.0/d613c88cf8fa3bab83b4ded3713f1f74830d1100e171db75bbddb
80b3345c9c0)
  0%|          | 0/3 [00:00<?, ?it/s]
```

Out[9]:
```
DatasetDict({
    train: Dataset({
        features: ['text', 'label'],
        num_rows: 25000
    })
    test: Dataset({
        features: ['text', 'label'],
        num_rows: 25000
    })
    unsupervised: Dataset({
        features: ['text', 'label'],
        num_rows: 50000
    })
})
```

In [10]:
```
get_dataset_split_names("imdb")
```

Out[10]:
```
['train', 'test', 'unsupervised']
```

Let's count the number of labs in each dataset

In [11]:
```
train_labels = pd.DataFrame(dataset["train"]['label'], columns=["label"])
print(train_labels.groupby("label")["label"].count())

test_labels = pd.DataFrame(dataset["test"]['label'], columns=["label"])
print(test_labels.groupby("label")["label"].count())
```

```
label
0    12500
1    12500
Name: label, dtype: int64
label
0    12500
1    12500
Name: label, dtype: int64
```

Question 1: How many splits does the dataset has?

There are 3 splits: `train`, `test` and `unsupervised`

Question 2: How big are the splits ?

train: 25000 test: 25000 unsupervised: 50000

Question 3: What is the proportion of each class on the supervised splits?

train: 50% positive, 50% negative test: 50% positive, 50% negative

Partie 2: Naive Bayes classifier

In [12]:
```
from string import punctuation
import re

def preprocess(dataset: pd.DataFrame) -> pd.DataFrame :
    """
    Preprocess the dataset by lowercasing the text and removing the punctuat

    Parameters
    ----------
    dataset : pd.DataFrame
        The dataset to preprocess

    Returns
    -------
    pd.DataFrame
        The preprocessed dataset
    """
    # First lower the case
    dataset["document"] = dataset["document"].apply(lambda x: x.lower())
    # Replace the punctuation with spaces. We keep the ' - that may give rev
    # Replace HTML tag <br />
    punctuation_to_remove = '|'.join(map(re.escape, sorted(list(filter(lambd
    print(f"Deleting all these punctuation: {punctuation_to_remove}")
    dataset["document"] = dataset["document"].apply(lambda x: re.sub(punctua
    return dataset
```

Apply the preprocessing steps to both the training and test sets. We choose to save them in a pandas DataFrame.

In [13]:
```
train_raw = pd.DataFrame(dataset["train"], columns=["text", "label"]).rename
preprocessed_train = preprocess(train_raw)
preprocessed_train
```

```
Deleting all these punctuation: \~|\}|\||\{|\`|_|\^|\]|\\|\[|@|\?|>|=|<|;|:
|/|\.|,|\+|\*|\)|\(|\&|%|\$|\#|"|!
```

Out[13]:

	document	class
0	i rented i am curious-yellow from my video sto...	0
1	i am curious yellow is a risible and preten...	0
2	if only to avoid making this type of film in t...	0
3	this film was probably inspired by godard's ma...	0
4	oh brother after hearing about this ridicul...	0
...	...	...
24995	a hit at the time but now better categorised a...	1
24996	i love this movie like no other another time ...	1
24997	this film and it's sequel barry mckenzie holds...	1
24998	'the adventures of barry mckenzie' started lif...	1
24999	the story centers around barry mckenzie who mu...	1

25000 rows × 2 columns

In [14]:
```
test_raw = pd.DataFrame(dataset["test"], columns=["text", "label"]).rename(c
preprocessed_test = preprocess(test_raw)
preprocessed_test
```

```
Deleting all these punctuation: \~|\}|\||\{|\`|_|\^|\]|\\|\[|@|\?|>|=|<|;|:
|/|\.|,|\+|\*|\)|\(|\&|%|\$|\#|"|!
```

Out[14]:

	document	class
0	i love sci-fi and am willing to put up with a ...	0
1	worth the entertainment value of a rental esp...	0
2	its a totally average film with a few semi-alr...	0
3	star rating saturday night friday ...	0
4	first off let me say if you haven't enjoyed a...	0
...	...	...
24995	just got around to seeing monster man yesterda...	1
24996	i got this as part of a competition prize i w...	1
24997	i got monster man in a box set of three films ...	1
24998	five minutes in i started to feel how naff th...	1
24999	i caught this movie on the sci-fi channel rece...	1

25000 rows × 2 columns

Question 2: Naive Bayes Classifier using pseudo-code

In [43]:
```
import numpy as np
from typing import List
```

```python
from sklearn.metrics import accuracy_score, precision_score, recall_score
from sklearn.feature_extraction.text import CountVectorizer

def get_vocabulary(d: pd.DataFrame) -> List[str]:
    """
    Return the vocabulary of the dataset

    Parameters
    ----------
    d : pd.DataFrame

    Returns
    -------
    List[str]
        The vocabulary
    """
    res = list(set(" ".join(d["document"]).split(" ")))
    # Remove empty string and words without any letter
    res = list(filter(lambda x: x != "" and re.search("[a-zA-Z]", x), res))
    return res

def train_naive_bayes(d: pd.DataFrame):
    """
    Train a Naive Bayes classifier
    Apply pseudo code from lecture 2

    Parameters
    ----------
    d : pd.DataFrame

    Returns
    -------
    logprior : dict
        The log prior of each class
    loglikelihood : dict
        The log likelihood of each word for each class
    V : List[str]
        The vocabulary
    """
    classes = d["class"].unique()
    logprior = {}
    bigdoc = {}
    loglikelihood = {}
    V = get_vocabulary(d)
    for c in classes:
        count = {}
        n_doc = len(d)
        n_c = len(d[d["class"] == c])
        logprior[c] = np.log(n_c / n_doc)
        bigdoc[c] = list(" ".join(d[d["class"] == c]["document"]).split(" ")
        for word in V:
            count[(word, c)] = bigdoc[c].count(word)
        for word in V:
            loglikelihood[(word, c)] = np.log((count[(word, c)] + 1) / (sum(
    return logprior, loglikelihood, V
```

```python
def test_naive_bayes(testdoc, classes, logprior, loglikelihood, V) -> int:
    """
    Test a Naive Bayes classifier

    Parameters
    ----------
    testdoc : str
        The document to classify
    classes : List[int]
        The list of classes
    logprior : dict
        The log prior of each class
    loglikelihood : dict
        The log likelihood of each word for each class
    V : List[str]
        The vocabulary

    Returns
    -------
    int
        The predicted class
    """
    sum_loglikelihood = {}
    for c in classes:
        sum_loglikelihood[c] = logprior[c]
        for word in testdoc.split(" "):
            if word in V:
                sum_loglikelihood[c] += loglikelihood[(word, c)]
    return max(sum_loglikelihood, key=sum_loglikelihood.get)
```

In [17]:
```python
# We reduce the dataset to 10% of the original size to speed up the training
train_dataset_reduced = preprocessed_train.loc[::10, :]
test_dataset_reduced = preprocessed_test.loc[::10, :]
logprior_r, loglikelyhood_r, V_r = train_naive_bayes(train_dataset_reduced)

all_res = []
for row in test_dataset_reduced.iterrows():
    test_doc = row[1]["document"]
    res = test_naive_bayes(test_doc, preprocessed_test["class"].unique(), lo
    all_res.append(res)

print("Manual Naive Bayes Accuracy Score -> ",accuracy_score(test_dataset_re
print("Manual Naive Bayes Precision Score -> ",precision_score(test_dataset_
print("Manual Naive Bayes Recall Score -> ",recall_score(test_dataset_reduce
```

```
Manual Naive Bayes Accuracy Score ->  80.12
Manual Naive Bayes Precision Score ->  84.44647758462946
Manual Naive Bayes Recall Score ->  73.83999999999999
```

Question 3: Naive Bayes Classifier using sklearn (Pipeline with CountVectorizer and MultinomialNB)

We will create a pipeline with a CountVectorizer and a MultinomialNB. We will use the default parameters for both of them as a first try.

In [18]:
```
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import Pipeline
```

In [19]:
```
def sklearn_naive_bayes(d_train: pd.DataFrame, pipeline_params: dict = {}) -
    """
    Train a Naive Bayes classifier using sklearn

    Parameters
    ----------
    d_train : pd.DataFrame
        The training dataset
    pipeline_params : dict, optional
        The parameters of the pipeline, by default {}

    Returns
    -------
    Pipeline
        The trained pipeline
    """
    # create pipeline
    pipeline = Pipeline([
        ('vectorizer', CountVectorizer()),
        ('classifier', MultinomialNB())
    ])
    pipeline.set_params(**pipeline_params)

    # train the model
    pipeline.fit(d_train["document"], d_train["class"])
    return pipeline

def test_sklearn_naive_bayes(pipeline: Pipeline, d_test: pd.DataFrame) -> Li
    """
    Test a Naive Bayes classifier using sklearn

    Parameters
    ----------
    pipeline : Pipeline
        The trained pipeline
    d_test : pd.DataFrame
        The test dataset

    Returns
    -------
    List[int]
        The predicted classes
    """
    # predict the labels on validation dataset
    predictions = pipeline.predict(d_test["document"])

    print("Sklearn Naive Bayes Accuracy Score -> ",accuracy_score(d_test["cl
    print("Sklearn Naive Bayes Precision Score -> ",precision_score(d_test["
    print("Sklearn Naive Bayes Recall Score -> ",recall_score(d_test["class"

    return predictions
```

In [20]:
```
pipeline = sklearn_naive_bayes(preprocessed_train)
predictions = test_sklearn_naive_bayes(pipeline, preprocessed_test)
```

```
Sklearn Naive Bayes Accuracy Score ->  81.44
Sklearn Naive Bayes Precision Score ->  86.05504587155963
Sklearn Naive Bayes Recall Score ->  75.03999999999999
```

Question 4: Report the accuracy on the test set

See prints above

Question 5: Most likely, the scikit-learn implementation will give better results. Looking at the documentation, explain why it could be the case.

The scikit-learn implementation is better because it uses a MultinomialNB which is a more efficient way to compute the probabilities. It also uses a CountVectorizer which is a more efficient way to count the words in the dataset.

Question 6: Why is accuracy a sufficient measure of evaluation here?

Because the dataset is balanced, we have the same number of positive and negative reviews. So the accuracy is a good measure of evaluation.

Question 7: Using one of the implementation, take at least 2 wrongly classified example from the test set and try explaining why the model failed.

In [21]:
```
# We will take a look at the sklearn implementation
# First we need to get the wrongly classified examples
wrongly_classified = preprocessed_test[preprocessed_test["class"] != predict

# We will take the first 2 examples
# We can see that the first example is a negative review but the model predi
# The second example is a positive review but the model predicted it as a ne
print(wrongly_classified.iloc[0]["document"])
print(wrongly_classified.iloc[1]["document"])
print()

# Let's see the probability of each class for the first example
print(pipeline.predict_proba([wrongly_classified.iloc[0]["document"]]))
# Let's see the probability of each class for the second example
print(pipeline.predict_proba([wrongly_classified.iloc[1]["document"]]))
```

```
blind date  columbia pictures  1934  was a decent film  but i have a few i
ssues with this film  first of all  i don't fault the actors in this film a
t all  but more or less  i have a problem with the script  also  i understa
nd that this film was made in the 1930's and people were looking to escape
reality  but the script made ann sothern's character look weak  she kept go
ing back and forth between suitors and i felt as though she should have sta
yed with paul kelly's character in the end  he truly did care about her and
her family and would have done anything for her and he did by giving her up
in the end to fickle neil hamilton who in my opinion was only out for a goo
d time  paul kelly's character  although a workaholic was a man of integrit
y and truly loved kitty  ann sothern  as opposed to neil hamilton  while he
did like her a lot  i didn't see the depth of love that he had for her char
acter  the production values were great  but the script could have used a l
ittle work
ben  rupert grint  is a deeply unhappy adolescent  the son of his unhappi
ly married parents  his father  nicholas farrell  is a vicar and his moth
er  laura linney  is  well  let's just say she's a somewhat hypocritic
al soldier in jesus' army  it's only when he takes a summer job as an assis
tant to a foul-mouthed  eccentric  once-famous and now-forgotten actress ev
ie walton  julie walters  that he finally finds himself in true 'harold a
nd maude' fashion  of course  evie is deeply unhappy herself and it's only
when these two sad sacks find each other that they can put their mutual mis
ery aside and hit the road to happiness of course it's corny and sentimenta
l and very predictable but it has a hard side to it  too and walters  who c
ould sleep-walk her way through this sort of thing if she wanted  is excell
ent  it's when she puts the craziness to one side and finds the pathos in t
he character  like hitting the bottle and throwing up in the sink  that s
he's at her best  the problem is she's the only interesting character in th
e film  and it's not because of the script which doesn't do anybody any fav
ours  grint  on the other hand  isn't just unhappy  he's a bit of a bore a
s well while linney's starched bitch is completely one-dimensional  still
she's got the english accent off pat  the best that can be said for it is
that it's mildly enjoyable - with the emphasis on the mildly

[[4.22158007e-06 9.99995778e-01]]
[[0.00150068 0.99849932]]
```

We can see that the model is very confident about its prediction for the two examples (0.99...) but it's wrong. These examples are very hard to classify because they are very close to the decision boundary and also mixing a movie description (which can have positive or negative connotations due to the life of the main character, etc) and a review. So the model is not able to classify them correctly because of the confusing bundary between description and facts and the opinion.

Question 8: What are the top 10 most important words (features) for each class? (bonus points)

In [22]:

```python
# We will use the sklearn implementation to get the top 10 most important wo

def get_top_10_words(pipeline: Pipeline) -> dict:
    """
    Get the top 10 words for each class
```

```
    Parameters
    ----------
    pipeline : Pipeline
        The trained pipeline

    Returns
    -------
    dict
        The top 10 words for each class
    """
    top_10_words = {}
    for c in preprocessed_test["class"].unique():
        loglikelihood = pipeline.named_steps["classifier"].feature_log_prob_
        V = pipeline.named_steps["vectorizer"].vocabulary_
        top_10_words[c] = [list(V.keys())[list(V.values()).index(i)] for i i
    return top_10_words
```

In [23]:
```
get_top_10_words(pipeline)
```

Out[23]:
```
{0: ['was', 'that', 'this', 'in', 'it', 'is', 'to', 'of', 'and', 'the'],
 1: ['as', 'this', 'that', 'it', 'in', 'is', 'to', 'of', 'and', 'the']}
```

The words we retreive are stop words, so they are not very meaningful. Let's try to remove them and see if we get better results.

In [24]:
```
pipeline_without_stopwords = sklearn_naive_bayes(preprocessed_train, {"vecto
predictions_without_stopwords = test_sklearn_naive_bayes(pipeline_without_st

get_top_10_words(pipeline_without_stopwords)
```

```
Sklearn Naive Bayes Accuracy Score ->  81.976
Sklearn Naive Bayes Precision Score ->  86.22439731738264
Sklearn Naive Bayes Recall Score ->  76.112
```

Out[24]:
```
{0: ['story',
  'don',
  'time',
  'really',
  'bad',
  'good',
  'just',
  'like',
  'film',
  'movie'],
 1: ['people',
  'really',
  'great',
  'time',
  'story',
  'just',
  'good',
  'like',
  'movie',
  'film']}
```

We see that the top 10 words are more unique using stopwords, but the results are

pretty equivalent with or without stopwords.

Question 9: Play with scikit-learn's version parameters. For example, see if you can consider unigram and bigram instead of only unigrams.

We will compare previous results using sklearn with the results using unigram and bigram, and with/without removing stopwords.

In [25]:
```
# Unigram and bigram
pipeline_bigram = sklearn_naive_bayes(preprocessed_train, {"vectorizer__ngra
predictions_bigram = test_sklearn_naive_bayes(pipeline_bigram, preprocessed_
```

```
Sklearn Naive Bayes Accuracy Score ->  84.244
Sklearn Naive Bayes Precision Score ->  87.4857693318154
Sklearn Naive Bayes Recall Score ->  79.92
```

In [26]:
```
# Unigram and bigram with stopwords
pipeline_bigram_stopwords = sklearn_naive_bayes(preprocessed_train, {"vector
predictions_bigram_stopwords = test_sklearn_naive_bayes(pipeline_bigram_stop
```

```
Sklearn Naive Bayes Accuracy Score ->  85.672
Sklearn Naive Bayes Precision Score ->  88.62612612612612
Sklearn Naive Bayes Recall Score ->  81.848
```

In [27]:
```
# Only bigram
pipeline_only_bigram = sklearn_naive_bayes(preprocessed_train, {"vectorizer_
predictions_only_bigram = test_sklearn_naive_bayes(pipeline_only_bigram, pre
```

```
Sklearn Naive Bayes Accuracy Score ->  82.952
Sklearn Naive Bayes Precision Score ->  87.63018454229857
Sklearn Naive Bayes Recall Score ->  76.736
```

In [28]:
```
# Only bigram with stopwords
pipeline_only_bigram_stopwords = sklearn_naive_bayes(preprocessed_train, {"v
predictions_only_bigram_stopwords = test_sklearn_naive_bayes(pipeline_only_b
```

```
Sklearn Naive Bayes Accuracy Score ->  86.952
Sklearn Naive Bayes Precision Score ->  89.35753237900477
Sklearn Naive Bayes Recall Score ->  83.896
```

The accuracy is better with only bigrams and without removing stopwords.

Part 3: Stemming & Lemmatization

In this part we will add preprocessing, including stemming and leammatization.

We need to add an extra module for spacy.

In [ ]:
```
! python -m spacy download en_core_web_sm
```

Lemmatization preprocessing

Let's start with a small example to understand how to recover a lem.

In this case we will use Spacy, especially its pipeline features to do preprocessing.

In [29]:
```
# Setup spacy
import spacy
nlp = spacy.load('en_core_web_sm')
```

In [30]:
```
# Take a 20 characters sentence example from the test dataset
test_list = dataset['train']['text'][0].split()[:20]
test_sentence = ' '.join(test_list)

# Lemmatize the sentence
doc = nlp(test_sentence)

# Get all token
tokens = [token.text for token in doc]

print(f'Original Sentence: {test_sentence}')
for token in doc:
    if token.text != token.lemma_:
        print(f'Original : {token.text}, New: {token.lemma_}')
```

```
Original Sentence: I rented I AM CURIOUS-YELLOW from my video store because
of all the controversy that surrounded it when it was
Original : rented, New: rent
Original : AM, New: be
Original : CURIOUS, New: curious
Original : surrounded, New: surround
Original : was, New: be
```

Results look good, words are reduced to their root form.

Let's define a preprocessing function.

In [31]:
```
def lemma_preprocessor(x_list: List[str]) -> List[str]:
    """
    Preprocessing function to lowercase and remove punctuation
    of a list of string and lemmatize each string.

    Args:
        x_list: List of strings

    Returns:
        List of preprocessed strings.
    """
    no_punc_lower = [x.lower().translate(str.maketrans("", "", punctuation))
    spacy_nlp = spacy.load('en_core_web_sm')
    res = []
    for sentence in no_punc_lower:
        doc = spacy_nlp(sentence)
        s = []
        for word in doc:
            s.append(word.lemma_)
        s = ' '.join(s)
```

```
        res.append(s)
    return res
```

Print a example of the result :

In [32]:
```
print(dataset['train']['text'][:10])
lemma_preprocessor(dataset['train']['text'][:10])
```

['I rented I AM CURIOUS-YELLOW from my video store because of all the contr
oversy that surrounded it when it was first released in 1967. I also heard
that at first it was seized by U.S. customs if it ever tried to enter this
country, therefore being a fan of films considered "controversial" I really
had to see this for myself.

The plot is centered around a young
Swedish drama student named Lena who wants to learn everything she can abou
t life. In particular she wants to focus her attentions to making some sort
of documentary on what the average Swede thought about certain political is
sues such as the Vietnam War and race issues in the United States. In betwe
en asking politicians and ordinary denizens of Stockholm about their opinio
ns on politics, she has sex with her drama teacher, classmates, and married
men.

What kills me about I AM CURIOUS-YELLOW is that 40 years ag
o, this was considered pornographic. Really, the sex and nudity scenes are
few and far between, even then it\'s not shot like some cheaply made porno.
While my countrymen mind find it shocking, in reality sex and nudity are a
major staple in Swedish cinema. Even Ingmar Bergman, arguably their answer
to good old boy John Ford, had sex scenes in his films.

I do com
mend the filmmakers for the fact that any sex shown in the film is shown fo
r artistic purposes rather than just to shock people and make money to be s
hown in pornographic theaters in America. I AM CURIOUS-YELLOW is a good fil
m for anyone wanting to study the meat and potatoes (no pun intended) of Sw
edish cinema. But really, this film doesn\'t have much of a plot.', '"I Am
Curious: Yellow" is a risible and pretentious steaming pile. It doesn\'t ma
tter what one\'s political views are because this film can hardly be taken
seriously on any level. As for the claim that frontal male nudity is an aut
omatic NC-17, that isn\'t true. I\'ve seen R-rated films with male nudity.
Granted, they only offer some fleeting views, but where are the R-rated fil
ms with gaping vulvas and flapping labia? Nowhere, because they don\'t exis
t. The same goes for those crappy cable shows: schlongs swinging in the bre
eze but not a clitoris in sight. And those pretentious indie movies like Th
e Brown Bunny, in which we\'re treated to the site of Vincent Gallo\'s thro
bbing johnson, but not a trace of pink visible on Chloe Sevigny. Before cry
ing (or implying) "double-standard" in matters of nudity, the mentally obtu
se should take into account one unavoidably obvious anatomical difference b
etween men and women: there are no genitals on display when actresses appea
rs nude, and the same cannot be said for a man. In fact, you generally won
\'t see female genitals in an American film in anything short of porn or ex
plicit erotica. This alleged double-standard is less a double standard than
an admittedly depressing ability to come to terms culturally with the insid
es of women\'s bodies.', "If only to avoid making this type of film in the
future. This film is interesting as an experiment but tells no cogent stor
y.

One might feel virtuous for sitting thru it because it touche
s on so many IMPORTANT issues but it does so without any discernable motiv
e. The viewer comes away with no new perspectives (unless one comes up with
one while one's mind wanders, as it will invariably do during this pointles
s film).

One might better spend one's time staring out a window
at a tree growing.

", "This film was probably inspired by Godar
d's Masculin, féminin and I urge you to see that film instead.

T
he film has two strong elements and those are, (1) the realistic acting (2)
the impressive, undeservedly good, photo. Apart from that, what strikes me
most is the endless stream of silliness. Lena Nyman has to be most annoying
actress in the world. She acts so stupid and with all the nudity in this fi
lm,...it's unattractive. Comparing to Godard's film, intellectuality has be
en replaced with stupidity. Without going too far on this subject, I would
say that follows from the difference in ideals between the French and the S
wedish society.

A movie of its time, and place. 2/10.", 'Oh, bro

ther...after hearing about this ridiculous film for umpteen years all I can think of is that old Peggy Lee song..

"Is that all there is??" ...I was just an early teen when this smoked fish hit the U.S. I was too young to get in the theater (although I did manage to sneak into "Goodbye Columbus"). Then a screening at a local film museum beckoned - Finally I could see this film, except now I was as old as my parents were when they schlepped to see it!!

The ONLY reason this film was not condemned to the anonymous sands of time was because of the obscenity case sparked by its U.S. release. MILLIONS of people flocked to this stinker, thinking they were going to see a sex film...Instead, they got lots of closeups of gnarly, repulsive Swedes, on-street interviews in bland shopping malls, asinie political pretension...and feeble who-cares simulated sex scenes with saggy, pale actors.

Cultural icon, holy grail, historic artifact..whatever this thing was, shred it, burn it, then stuff the ashes in a lead box!

Elite esthetes still scrape to find value in its boring pseudo revolutionary political spewings..But if it weren\'t for the censorship scandal, it would have been ignored, then forgotten.

Instead, the "I Am Blank, Blank" rhythymed title was repeated endlessly for years as a titilation for porno films (I am Curious, Lavender - for gay films, I Am Curious, Black - for blaxploitation films, etc..) and every ten years or so the thing rises from the dead, to be viewed by a new generation of suckers who want to see that "naughty sex film" that "revolutionized the film industry"...

Yeesh, avoid like the plague..Or if you MUST see it - rent the video and fast forward to the "dirty" parts, just to get it over with.

', "I would put this at the top of my list of films in the category of unwatchable trash! There are films that are bad, but the worst kind are the ones that are unwatchable but you are suppose to like them because they are supposed to be good for you! The sex sequences, so shocking in its day, couldn't even arouse a rabbit. The so called controversial politics is strictly high school sophomore amateur night Marxism. The film is self-consciously arty in the worst sense of the term. The photography is in a harsh grainy black and white. Some scenes are out of focus or taken from the wrong angle. Even the sound is bad! And some people call this art?

", "Whoever wrote the screenplay for this movie obviously never consulted any books about Lucille Ball, especially her autobiography. I've never seen so many mistakes in a biopic, ranging from her early years in Celoron and Jamestown to her later years with Desi. I could write a whole list of factual errors, but it would go on for pages. In all, I believe that Lucille Ball is one of those inimitable people who simply cannot be portrayed by anyone other than themselves. If I were Lucie Arnaz and Desi, Jr., I would be irate at how many mistakes were made in this film. The filmmakers tried hard, but the movie seems awfully sloppy to me.", 'When I first saw a glimpse of this movie, I quickly noticed the actress who was playing the role of Lucille Ball. Rachel York\'s portrayal of Lucy is absolutely awful. Lucille Ball was an astounding comedian with incredible talent. To think about a legend like Lucille Ball being portrayed the way she was in the movie is horrendous. I cannot believe out of all the actresses in the world who could play a much better Lucy, the producers decided to get Rachel York. She might be a good actress in other roles but to play the role of Lucille Ball is tough. It is pretty hard to find someone who could resemble Lucille Ball, but they could at least find someone a bit similar in looks and talent. If you noticed York\'s portrayal of Lucy in episodes of I Love Lucy like the chocolate factory or vitavetavegamin, nothing is similar in any way-her expression, voice, or movement.

To top it all off, Danny Pino playing Desi Arnaz is horrible. Pino does not qualify to play as Ricky. He\'s small and skinny, his accent is unreal, and once again, his acting is unbelievable. Although Fred a

nd Ethel were not similar either, they were not as bad as the characters of Lucy and Ricky.

Overall, extremely horrible casting and the stor y is badly told. If people want to understand the real life situation of Lu cille Ball, I suggest watching A&E Biography of Lucy and Desi, read the boo k from Lucille Ball herself, or PBS\' American Masters: Finding Lucy. If yo u want to see a docudrama, "Before the Laughter" would be a better choice. The casting of Lucille Ball and Desi Arnaz in "Before the Laughter" is much better compared to this. At least, a similar aspect is shown rather than no thing.', 'Who are these "They"- the actors? the filmmakers? Certainly could n\'t be the audience- this is among the most air-puffed productions in exis tence. It\'s the kind of movie that looks like it was a lot of fun to shoot \x97 TOO much fun, nobody is getting any actual work done, and that almost always makes for a movie that\'s no fun to watch.

Ritter dons gl asses so as to hammer home his character\'s status as a sort of dopplegange r of the bespectacled Bogdanovich; the scenes with the breezy Ms. Stratten are sweet, but have an embarrassing, look-guys-I\'m-dating-the-prom-queen f eel to them. Ben Gazzara sports his usual cat\'s-got-canary grin in a futil e attempt to elevate the meager plot, which requires him to pursue Audrey H epburn with all the interest of a narcoleptic at an insomnia clinic. In the meantime, the budding couple\'s respective children (nepotism alert: Bogdan ovich\'s daughters) spew cute and pick up some fairly disturbing pointers o n \'love\' while observing their parents. (Ms. Hepburn, drawing on her dign ity, manages to rise above the proceedings- but she has the monumental chal lenge of playing herself, ostensibly.) Everybody looks great, but so what? It\'s a movie and we can expect that much, if that\'s what you\'re looking for you\'d be better off picking up a copy of Vogue.

Oh- and it has to be mentioned that Colleen Camp thoroughly annoys, even apart from he r singing, which, while competent, is wholly unconvincing... the country an d western numbers are woefully mismatched with the standards on the soundtr ack. Surely this is NOT what Gershwin (who wrote the song from which the mo vie\'s title is derived) had in mind; his stage musicals of the 20\'s may h ave been slight, but at least they were long on charm. "They All Laughed" t ries to coast on its good intentions, but nobody- least of all Peter Bogdan ovich - has the good sense to put on the brakes.

Due in no small part to the tragic death of Dorothy Stratten, this movie has a special plac e in the heart of Mr. Bogdanovich- he even bought it back from its producer s, then distributed it on his own and went bankrupt when it didn\'t prove p opular. His rise and fall is among the more sympathetic and tragic of Holly wood stories, so there\'s no joy in criticizing the film... there _is_ real emotional investment in Ms. Stratten\'s scenes. But "Laughed" is a faint ec ho of "The Last Picture Show", "Paper Moon" or "What\'s Up, Doc"- following "Daisy Miller" and "At Long Last Love", it was a thundering confirmation of the phase from which P.B. has never emerged.

All in all, though, the movie is harmless, only a waste of rental. I want to watch people havin g a good time, I\'ll go to the park on a sunny day. For filmic expressions of joy and love, I\'ll stick to Ernest Lubitsch and Jaques Demy...', "This is said to be a personal film for Peter Bogdonavitch. He based it on his li fe but changed things around to fit the characters, who are detectives. The se detectives date beautiful models and have no problem getting them. Sound s more like a millionaire playboy filmmaker than a detective, doesn't it? T his entire movie was written by Peter, and it shows how out of touch with r eal people he was. You're supposed to write what you know, and he did that, indeed. And leaves the audience bored and confused, and jealous, for that m atter. This is a curio for people who want to see Dorothy Stratten, who was murdered right after filming. But Patti Hanson, who would, in real life, ma rry Keith Richards, was also a model, like Stratten, but is a lot better an

d has a more ample part. In fact, Stratten's part seemed forced; added. She doesn't have a lot to do with the story, which is pretty convoluted to begi n with. All in all, every character in this film is somebody that very few people can relate with, unless you're millionaire from Manhattan with beaut iful supermodels at your beckon call. For the rest of us, it's an irritatin g snore fest. That's what happens when you're out of touch. You entertain y our few friends with inside jokes, and bore all the rest."]

Out[32]: ['I rent I be curiousyellow from my video store because of all the controve
rsy that surround it when it be first release in 1967 I also hear that at f
irst it be seize by us customs if it ever try to enter this country therefo
re be a fan of film consider controversial I really have to see this for my
selfbr br the plot be center around a young swedish drama student name lena
who want to learn everything she can about life in particular she want to f
ocus her attention to make some sort of documentary on what the average swe
de think about certain political issue such as the vietnam war and race iss
ue in the united states in between ask politician and ordinary denizen of s
tockholm about their opinion on politic she have sex with her drama teacher
classmate and married menbr br what kill I about I be curiousyellow be that
40 year ago this be consider pornographic really the sex and nudity scene b
e few and far between even then its not shoot like some cheaply make porno
while my countryman mind find it shock in reality sex and nudity be a major
staple in swedish cinema even ingmar bergman arguably their answer to good
old boy john ford have sex scene in his filmsbr br I do commend the filmmak
er for the fact that any sex show in the film be show for artistic purpose
rather than just to shock people and make money to be show in pornographic
theater in america I be curiousyellow be a good film for anyone want to stu
dy the meat and potatoe no pun intend of swedish cinema but really this fil
m do not have much of a plot',
 'I be curious yellow be a risible and pretentious steaming pile it do not
matter what one political view be because this film can hardly be take seri
ously on any level as for the claim that frontal male nudity be an automati
c nc17 that be not true I ve see rrate film with male nudity grant they onl
y offer some fleeting view but where be the rrate film with gape vulvas and
flap labia nowhere because they do not exist the same go for those crappy c
able show schlong swinge in the breeze but not a clitoris in sight and thos
e pretentious indie movie like the brown bunny in which be treat to the sit
e of vincent gallos throb johnson but not a trace of pink visible on chloe
sevigny before cry or imply doublestandard in matter of nudity the mentally
obtuse should take into account one unavoidably obvious anatomical differen
ce between man and woman there be no genital on display when actress appear
nude and the same can not be say for a man in fact you generally will not s
ee female genital in an american film in anything short of porn or explicit
erotica this allege doublestandard be less a double standard than an admitt
edly depressing ability to come to term culturally with the inside of women
body',
 'if only to avoid make this type of film in the future this film be intere
sting as an experiment but tell no cogent storybr br one might feel virtuou
s for sit thru it because it touch on so many important issue but it do so
without any discernable motive the viewer come away with no new perspective
unless one come up with one while one mind wander as it will invariably do
during this pointless filmbr br one might well spend one time stare out a w
indow at a tree growingbr br',
 'this film be probably inspire by godard masculin féminin and I urge you t
o see that film insteadbr br the film have two strong element and those be
1 the realistic acting 2 the impressive undeservedly good photo apart from
that what strike I most be the endless stream of silliness lena nyman have
to be most annoying actress in the world she act so stupid and with all the
nudity in this filmits unattractive compare to godard film intellectuality
have be replace with stupidity without go too far on this subject I would s
ay that follow from the difference in ideal between the french and the swed
ish societybr br a movie of its time and place 210',
 'oh brotherafter hear about this ridiculous film for umpteen year all I ca
n think of be that old peggy lee songbr br be that all there be I be just a

n early teen when this smoke fish hit the us I be too young to get in the t
heater although I do manage to sneak into goodbye columbus then a screening
at a local film museum beckon finally I could see this film except now I
be as old as my parent be when they schleppe to see itbr br the only reason
this film be not condemn to the anonymous sand of time be because of the ob
scenity case spark by its us release million of people flock to this stinke
r think they be go to see a sex filminstead they get lot of closeup of gnar
ly repulsive swede onstreet interview in bland shopping mall asinie politic
al pretensionand feeble whocare simulate sex scene with saggy pale actorsbr
br cultural icon holy grail historic artifactwhatever this thing be shre it
burn it then stuff the ashe in a lead boxbr br elite esthete still scrape t
o find value in its boring pseudo revolutionary political spewingsbut if it
be not for the censorship scandal it would have be ignore then forgottenbr
br instead the I be blank blank rhythymed title be repeat endlessly for yea
r as a titilation for porno film I be curious lavender for gay film I be
curious black for blaxploitation film etc and every ten year or so the th
ing rise from the dead to be view by a new generation of sucker who want to
see that naughty sex film that revolutionize the film industrybr br yeesh a
void like the plagueor if you must see it rent the video and fast forward
to the dirty part just to get it over withbr br',
 'I would put this at the top of my list of film in the category of unwatch
able trash there be film that be bad but the bad kind be the one that be un
watchable but you be suppose to like they because they be suppose to be goo
d for you the sex sequence so shocking in its day could not even arouse a r
abbit the so call controversial politic be strictly high school sophomore a
mateur night marxism the film be selfconsciously arty in the bad sense of t
he term the photography be in a harsh grainy black and white some scene be
out of focus or take from the wrong angle even the sound be bad and some pe
ople call this artbr br',
 'whoever write the screenplay for this movie obviously never consult any b
ook about lucille ball especially her autobiography I ve never see so many
mistake in a biopic range from her early year in celoron and jamestown to h
er later year with desi I could write a whole list of factual error but it
would go on for page in all I believe that lucille ball be one of those ini
mitable people who simply can not be portray by anyone other than themselve
s if I be lucie arnaz and desi jr I would be irate at how many mistake be m
ake in this film the filmmaker try hard but the movie seem awfully sloppy t
o I',
 'when I first see a glimpse of this movie I quickly notice the actress who
be play the role of lucille ball rachel yorks portrayal of lucy be absolute
ly awful lucille ball be an astounding comedian with incredible talent to t
hink about a legend like lucille ball be portray the way she be in the movi
e be horrendous I can not believe out of all the actress in the world who c
ould play a much well lucy the producer decide to get rachel york she might
be a good actress in other role but to play the role of lucille ball be tou
gh it be pretty hard to find someone who could resemble lucille ball but th
ey could at least find someone a bit similar in look and talent if you noti
ce york portrayal of lucy in episode of I love lucy like the chocolate fact
ory or vitavetavegamin nothing be similar in any wayher expression voice or
movementbr br to top it all off danny pino play desi arnaz be horrible pino
do not qualify to play as ricky he s small and skinny his accent be unreal
and once again his acting be unbelievable although fred and ethel be not si
milar either they be not as bad as the character of lucy and rickybr br ove
rall extremely horrible casting and the story be badly tell if people want
to understand the real life situation of lucille ball I suggest watch ae bi
ography of lucy and desi read the book from lucille ball herself or pbs ame

rican master find lucy if you want to see a docudrama before the laughter w
ould be a well choice the casting of lucille ball and desi arnaz in before
the laughter be much well compare to this at least a similar aspect be show
rather than nothing',
 'who be these they the actor the filmmaker certainly could not be the audi
ence this be among the most airpuffe production in existence its the kind o
f movie that look like it be a lot of fun to shoot\x97 too much fun nobody
be get any actual work do and that almost always make for a movie that s no
fun to watchbr br ritter don glass so as to hammer home his character statu
s as a sort of doppleganger of the bespectacle bogdanovich the scene with t
he breezy ms stratten be sweet but have an embarrassing lookguysimdatingthe
promqueen feel to they ben gazzara sport his usual catsgotcanary grin in a
futile attempt to elevate the meager plot which require he to pursue audrey
hepburn with all the interest of a narcoleptic at an insomnia clinic in the
meantime the bud couple respective child nepotism alert bogdanovich daughte
r spew cute and pick up some fairly disturbing pointer on love while observ
e their parent ms hepburn draw on her dignity manage to rise above the proc
eeding but she have the monumental challenge of play herself ostensibly eve
rybody look great but so what its a movie and we can expect that much if th
at s what you re look for you d be well off pick up a copy of voguebr br oh
and it have to be mention that colleen camp thoroughly annoy even apart fro
m her singing which while competent be wholly unconvincing the country and
western number be woefully mismatch with the standard on the soundtrack sur
ely this be not what gershwin who write the song from which the movie title
be derive have in mind his stage musical of the 20 may have be slight but a
t least they be long on charm they all laugh try to coast on its good inten
tion but nobody least of all peter bogdanovich have the good sense to put
on the brakesbr br due in no small part to the tragic death of dorothy stra
tten this movie have a special place in the heart of mr bogdanovich he even
buy it back from its producer then distribute it on his own and go bankrupt
when it do not prove popular his rise and fall be among the more sympatheti
c and tragic of hollywood story so there s no joy in criticize the film the
re be real emotional investment in ms stratten scene but laugh be a faint e
cho of the last picture show paper moon or what s up doc follow daisy mille
r and at long last love it be a thunder confirmation of the phase from whic
h pb have never emergedbr br all in all though the movie be harmless only a
waste of rental I want to watch people have a good time ill go to the park
on a sunny day for filmic expression of joy and love ill stick to ernest lu
bitsch and jaque demy',
 'this be say to be a personal film for peter bogdonavitch he base it on hi
s life but change thing around to fit the character who be detective these
detective date beautiful model and have no problem get they sound more like
a millionaire playboy filmmaker than a detective do not it this entire movi
e be write by peter and it show how out of touch with real people he be you
re suppose to write what you know and he do that indeed and leave the audie
nce bored and confused and jealous for that matter this be a curio for peop
le who want to see dorothy stratten who be murder right after film but patt
i hanson who would in real life marry keith richard be also a model like st
ratten but be a lot well and have a more ample part in fact stratten part s
eem force add she do not have a lot to do with the story which be pretty co
nvoluted to begin with all in all every character in this film be somebody
that very few people can relate with unless you re millionaire from manhatt
an with beautiful supermodel at your beckon call for the rest of we its an
irritating snore fest that s what happen when you re out of touch you enter
tain your few friend with inside joke and bear all the rest']

We see that the preprocessing is working well: words are reduced to their lemma.

Stemming preprocessing

Let's start with a small example to understand how to recover a lem.

In this case we will use NLTK, another library than Spacy, but it offers stemming unlike Spacy

In [2]:
```
import nltk

from nltk.stem import PorterStemmer
nltk.download("punkt")
```

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

Out[2]: True

In [34]:
```
# Initialize Python porter stemmer
ps = PorterStemmer()

test_list = dataset['train']['text'][0].split()[:20]
test_sentence = ' '.join(test_list)

# Example inflections to reduce
example_words = ["program","programming","programer","programs","programmed"

print(f'Original Sentence: {test_sentence}')
# Perform stemming
print("{0:20}{1:20}".format("--Word--","--Stem--"))
for word in test_list:
    print ("{0:20}{1:20}".format(word, ps.stem(word)))
```

```
Original Sentence: I rented I AM CURIOUS-YELLOW from my video store because
of all the controversy that surrounded it when it was
--Word--            --Stem--
I                   i
rented              rent
I                   i
AM                  am
CURIOUS-YELLOW      curious-yellow
from                from
my                  my
video               video
store               store
because             becaus
of                  of
all                 all
the                 the
controversy         controversi
that                that
surrounded          surround
it                  it
when                when
it                  it
was                 wa
```

Again, results are stasisfyng. However, we observe some errors, such as "becaus" instead of "because", or "wa" instead of "was".

Let's define a preprocessing function.

In [35]:
```python
def stem_preprocessor(x_list: List[str]) -> List[str]:
    """
    Preprocessing function to stem each string.

    Args:
        x_list: List of strings

    Returns:
        List of preprocessed strings.
    """
    spacy_nlp = spacy.load('en_core_web_sm')
    res = []
    ps = PorterStemmer()
    for sentence in x_list:
        doc = spacy_nlp(sentence)
        s = []
        for word in doc:
            s.append(ps.stem(str(word)))
        s = ' '.join(s)
        res.append(s)
    return res
```

In [36]:
```python
example_words = ["program","programming","programer","programs","programmed"
stem_preprocessor(example_words)
```

Out[36]:
```
['program', 'program', 'program', 'program', 'program']
```

Training with Stem and Lemmatize

Lemma training

Both are working well. Now let's try to use lemmatization in our pipeline

In [37]:
```
# use stem_preprocessor to preprocess the training and test data
preprocessed_train_stem = lemma_preprocessor(train_raw["document"][:10])
preprocessed_train_stem
```

Out[37]: ['I rent I be curiousyellow from my video store because of all the controve
rsy that surround it when it be first release in 1967 I also hear that at
first it be seize by u s custom if it ever try to enter this country th
erefore be a fan of film consider controversial I really have to see th
is for myself the plot be center around a young swedish drama student name
lena who want to learn everything she can about life in particular she wa
nt to focus her attention to make some sort of documentary on what the aver
age swede think about certain political issue such as the vietnam war and r
ace issue in the united states in between ask politician and ordinary den
izen of stockholm about their opinion on politic she have sex with her dr
ama teacher classmate and married man what kill I about I be curiousyel
low be that 40 year ago this be consider pornographic really the sex
and nudity scene be few and far between even then its not shoot like some
cheaply make porno while my countryman mind find it shocking in reality
sex and nudity be a major staple in swedish cinema even ingmar bergman
arguably their answer to good old boy john ford have sex scene in his fil
m I do commend the filmmaker for the fact that any sex show in the film be
show for artistic purpose rather than just to shock people and make money t
o be show in pornographic theater in america I be curiousyellow be a good
film for anyone want to study the meat and potato no pun intend of swed
ish cinema but really this film do not have much of a plot',
' I be curious yellow be a risible and pretentious steaming pile it
do not matter what one political view be because this film can hardly be ta
ke seriously on any level as for the claim that frontal male nudity be an
automatic nc17 that be not true I ve see rrate film with male nudity
grant they only offer some fleeting view but where be the rrate film wi
th gape vulvas and flap labia nowhere because they do not exist the s
ame go for those crappy cable show schlong swinge in the breeze but not a
clitoris in sight and those pretentious indie movie like the brown bunny
in which be treat to the site of vincent gallos throb johnson but not a
trace of pink visible on chloe sevigny before cry or imply doublesta
ndard in matter of nudity the mentally obtuse should take into account
one unavoidably obvious anatomical difference between man and woman there
be no genital on display when actress appear nude and the same can not be
say for a man in fact you generally will not see female genital in an a
merican film in anything short of porn or explicit erotica this allege do
ublestandard be less a double standard than an admittedly depressing abilit
y to come to term culturally with the inside of women body',
'if only to avoid make this type of film in the future this film be inte
resting as an experiment but tell no cogent story one might feel virtuous f
or sit thru it because it touch on so many important issue but it do so wit
hout any discernable motive the viewer come away with no new perspective
unless one come up with one while one mind wander as it will invariably
do during this pointless film one might well spend one time stare out a w
indow at a tree grow',
'this film be probably inspire by godard masculin féminin and I urge you
to see that film instead the film have two strong element and those be 1
the realistic act 2 the impressive undeservedly good photo apa
rt from that what strike I most be the endless stream of silliness lena
nyman have to be most annoying actress in the world she act so stupid and
with all the nudity in this film its unattractive compare to godard f
ilm intellectuality have be replace with stupidity without go too far o
n this subject I would say that follow from the difference in ideal betwe
en the french and the swedish society a movie of its time and place 2 1
0',
'oh brother after hear about this ridiculous film for umpteen year al

l I can think of be that old peggy lee song be that all there be
I be just an early teen when this smoke fish hit the u s I be too young t
o get in the theater although I do manage to sneak into goodbye columbu
s then a screening at a local film museum beckon finally I could see
this film except now I be as old as my parent be when they schleppe to se
e it the only reason this film be not condemn to the anonymous sand of ti
me be because of the obscenity case spark by its u s release million of
people flock to this stinker thinking they be go to see a sex film ins
tead they get lot of closeup of gnarly repulsive swede onstreet inter
view in bland shopping mall asinie political pretension and feeble who
care simulate sex scene with saggy pale actor cultural icon holy grail
 historic artifact whatever this thing be shre it burn it then stu
ff the ashe in a lead box elite esthete still scrape to find value in its b
oring pseudo revolutionary political spewing but if it be not for the cen
sorship scandal it would have be ignore then forget instead the I b
e blank blank rhythyme title be repeat endlessly for year as a titilati
on for porno film I be curious lavender for gay film I be curious
black for blaxploitation film etc and every ten year or so the thin
g rise from the dead to be view by a new generation of sucker who want to
see that naughty sex film that revolutionize the film industry ye
esh avoid like the plague or if you must see it rent the video and fa
st forward to the dirty part just to get it over with',
 'I would put this at the top of my list of film in the category of unwatch
able trash there be film that be bad but the bad kind be the one that b
e unwatchable but you be suppose to like they because they be suppose to be
good for you the sex sequence so shocking in its day could not even a
rouse a rabbit the so call controversial politic be strictly high school
sophomore amateur night marxism the film be selfconsciously arty in the b
ad sense of the term the photography be in a harsh grainy black and white
  some scene be out of focus or take from the wrong angle even the sound
be bad and some people call this art',
 'whoever write the screenplay for this movie obviously never consult any b
ook about lucille ball especially her autobiography I ve never see so m
any mistake in a biopic range from her early year in celoron and jamestow
n to her later year with desi I could write a whole list of factual error
  but it would go on for page in all I believe that lucille ball be on
e of those inimitable people who simply can not be portray by anyone other
than themselves if I be lucie arnaz and desi jr I would be irate at
how many mistake be make in this film the filmmaker try hard but the mo
vie seem awfully sloppy to I',
 'when I first see a glimpse of this movie I quickly notice the actress w
ho be play the role of lucille ball rachel yorks portrayal of lucy be abs
olutely awful lucille ball be an astounding comedian with incredible tale
nt to think about a legend like lucille ball be portray the way she be in
the movie be horrendous I can not believe out of all the actress in the w
orld who could play a much well lucy the producer decide to get rachel yo
rk she might be a good actress in other role but to play the role of luci
lle ball be tough it be pretty hard to find someone who could resemble lu
cille ball but they could at least find someone a bit similar in look and
talent if you notice york portrayal of lucy in episode of I love lucy lik
e the chocolate factory or vitavetavegamin nothing be similar in any wayh
er expression voice or movement to top it all off danny pino play des
i arnaz be horrible pino do not qualify to play as ricky he s small and
skinny his accent be unreal and once again his acting be unbelievable
  although fred and ethel be not similar either they be not as bad as th
e character of lucy and ricky overall extremely horrible casting and the

story be badly tell   if people want to understand the real life situation of lucille ball   I suggest watch a e biography of lucy and desi   read the book from lucille ball herself   or pbs american masters   find lucy   if y ou want to see a docudrama   before the laughter   would be a well choice   the casting of lucille ball and desi arnaz in   before the laughter   be much well compare to this   at least   a similar aspect be show rather than nothing',

 'who be these   they   the actor   the filmmaker   certainly could not be the audience this be among the most airpuffe production in existence   its the kind of movie that look like it be a lot of fun to shoot\x97 too much f un   nobody be get any actual work do   and that almost always make for a m ovie that s no fun to watch ritter don glass so as to hammer home his chara cter status as a sort of dopppleganger of the bespectacle bogdanovich   the scene with the breezy ms   stratten be sweet   but have an embarrassing   l ookguysimdatingthepromqueen feel to they   ben gazzara sport his usual cats gotcanary grin in a futile attempt to elevate the meager plot   which requi re he to pursue audrey hepburn with all the interest of a narcoleptic at an insomnia clinic   in the meantime   the bud couple respective child   nepot ism alert   bogdanovich daughter   spew cute and pick up some fairly distur bing pointer on love while observe their parent   ms   hepburn   draw on h er dignity   manage to rise above the proceeding but she have the monumenta l challenge of play herself   ostensibly   everybody look great   but so w hat   its a movie and we can expect that much   if that s what you re look for you d be well off pick up a copy of vogue oh and it have to be mention that colleen camp thoroughly annoy   even apart from her singing   which while competent   be wholly unconvincing   the country and western number be woefully mismatch with the standard on the soundtrack   surely this be n ot what gershwin   who write the song from which the movie title be derive   have in mind   his stage musical of the 20 may have be slight   but at le ast they be long on charm   they all laugh   try to coast on its good inte ntion   but nobody least of all peter bogdanovich   have the good sense to put on the brake due in no small part to the tragic death of dorothy stratt en   this movie have a special place in the heart of mr   bogdanovich he ev en buy it back from its producer   then distribute it on his own and go ban krupt when it do not prove popular   his rise and fall be among the more sy mpathetic and tragic of hollywood story   so there s no joy in criticize th e film   there   be   real emotional investment in ms   stratten scene   but   laugh   be a faint echo of   the last picture show   paper moon   o r   what s up   doc   follow   daisy miller   and   at long last love   it be a thunder confirmation of the phase from which p b   have never emerge a ll in all   though   the movie be harmless   only a waste of rental   I wan t to watch people have a good time   ill go to the park on a sunny day   fo r filmic expression of joy and love   ill stick to ernest lubitsch and jaqu es demy   ',

 'this be say to be a personal film for peter bogdonavitch   he base it on his life but change thing around to fit the character   who be detective   these detective date beautiful model and have no problem get they   sound m ore like a millionaire playboy filmmaker than a detective   do not it   thi s entire movie be write by peter   and it show how out of touch with real p eople he be   you re suppose to write what you know   and he do that   inde ed   and leave the audience bored and confuse   and jealous   for that matt er   this be a curio for people who want to see dorothy stratten   who be m urder right after film   but patti hanson   who would   in real life   marr y keith richard   be also a model   like stratten   but be a lot well and h ave a more ample part   in fact   stratten part seem force   add   she do n ot have a lot to do with the story   which be pretty convoluted to begin wi

```
th   all in all   every character in this film be somebody that very few pe
ople can relate with   unless you re millionaire from manhattan with beauti
ful supermodel at your beckon call   for the rest of we   its an irritating
snore fest   that s what happen when you re out of touch   you entertain yo
ur few friend with inside joke   and bear all the rest']
```

Now let's define a function that will drive the model by adding the preprocessor lemma to the pipeline

In [38]:
```python
from sklearn.preprocessing import FunctionTransformer

def sklearn_naive_bayes_lemma(d_train: pd.DataFrame, pipeline_params: dict =
    """
    Train a Naive Bayes classifier using sklearn with lemmatization.

    Parameters
    ----------
    d_train : pd.DataFrame
        The training dataset
    pipeline_params : dict, optional
        The parameters of the pipeline, by default {}

    Returns
    -------
    Pipeline
        The trained pipeline
    """
    # create pipeline with lemmatization, vectorizer and classifier
    pipeline = Pipeline([
        ('lemmatizer', FunctionTransformer(lemma_preprocessor)),
        ('vectorizer', CountVectorizer()),
        ('classifier', MultinomialNB())
    ])
    pipeline.set_params(**pipeline_params)

    # train the model
    pipeline.fit(d_train["document"], d_train["class"])
    return pipeline
```

Training and evaluation of the model again with these pretreatment :

In [40]:
```python
pipeline_lemma = sklearn_naive_bayes_lemma(train_raw[::10])
predictions_lemma = test_sklearn_naive_bayes(pipeline_lemma, test_raw[::10])
```

```
Sklearn Naive Bayes Accuracy Score ->  80.12
Sklearn Naive Bayes Precision Score ->  84.25841674249318
Sklearn Naive Bayes Recall Score ->  74.08
```

Results are not better than before (with default settings): 80.12% vs 81.44% accuracy. This is probably due to the fact that the lemmatization is not very efficient in this case. This can be caused by the fact the language is English, and the lemmatization is not very efficient for this language because of it's low morphology, removing information that could be useful for the classifier.

Let's try with stemming.

Stem training

Now let's define a function that will drive the model by adding the preprocessor stem to the pipeline

In [41]:
```
from sklearn.preprocessing import FunctionTransformer

def sklearn_naive_bayes_stem(d_train: pd.DataFrame, pipeline_params: dict =
    """
    Train a Naive Bayes classifier using sklearn with lemmatization.

    Parameters
    ----------
    d_train : pd.DataFrame
        The training dataset
    pipeline_params : dict, optional
        The parameters of the pipeline, by default {}

    Returns
    -------
    Pipeline
        The trained pipeline
    """
    # create pipeline with lemmatization, vectorizer and classifier
    pipeline = Pipeline([
        ('lemmatizer', FunctionTransformer(stem_preprocessor)),
        ('vectorizer', CountVectorizer()),
        ('classifier', MultinomialNB())
    ])
    pipeline.set_params(**pipeline_params)

    # train the model
    pipeline.fit(d_train["document"], d_train["class"])
    return pipeline
```

In [42]:
```
pipeline_stem = sklearn_naive_bayes_stem(train_raw[::10])
predictions_stem = test_sklearn_naive_bayes(pipeline_stem, test_raw[::10])
```

```
Sklearn Naive Bayes Accuracy Score ->  79.84
Sklearn Naive Bayes Precision Score ->  83.97085610200364
Sklearn Naive Bayes Recall Score ->  73.76
```

Here the results are even worse than before (with default settings): 79.84% vs 81.44% accuracy. Again, we surely have the same problem as before, the stemming is not very efficient in this case. Lemmatization is better than stemming in this case, because it's more aggressive on words changed.