National University of Sciences and Technology (NUST)
School of Electrical Engineering and Computer Science

Department of Computing

CS370: Artificial Intelligence

Class: BSCS-6AB

Lab 07: Sentiment Analysis (Part 1)

Date: 21 Mar 2019

Time: 10am-1pm & 2pm-5pm

Lab Engineer: Syed Muhammad Ali Musa

Lab 07: Sentiment Analysis (Part 1)

Introduction

Sentiment Analysis, also known as opinion mining refers to the use of natural language processing, text analysis to identify and extract subjective information in source materials. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.

Objective

In this lab you will use IMDB database that contains 25000 movie reviews. Each movie review is labeled with a positive or a negative sentiment. This is your training dataset. You will use IMDB dataset to train a classifier. The trained classifier when presented with a new review will predict if the review is positive or negative.

In this lab you will clean the dataset so that you can use it for training purposes. In the next lab you will use the clean dataset for training.

Tools/Software Requirement

Python, pandas, re, BeautifulSoup4, nltk

Lab Tasks

Download labeledTrainData.tsv, which contains 25000 IMDB movie reviews.

Use Pandas http://pandas.pydata.org python package to read this file. Pandas package is preinstalled in your canopy python distribution.

```
import pandas as pd
train = pd.read_csv('labeledTrainData', header=0, delimiter='\t', quoting=3)
```

Task 1:

How is the data stored in the variable 'train'?

Answer: Data is basically stored in the tabular form. Pandas is basically placing the data in three columns.

What is the shape of the variable 'train'?

Answer: The shape of the variable train is (25000, 3)

How do you read the first few reviews from the variable 'train'?

Answer: We can read the first few review by using command *train["review"][0]* and we can also use pandas function *train.head()*.

There are HTML tags in the review. HTML tags won't help us in sentiment analysis. So we remove them. We will use Beautiful Soup

http://www.crummy.com/software/BeautifulSoup/bs4/doc/ package to do that. First install it in your canopy distribution using the following command

```
pip install BeautifulSoup4
```

Remember to restart the kernel after the installation. Now execute the following to remove HTML tags from the training reviews:

```
from bs4 import BeautifulSoup
# Run the BeautifulSoup object on a single movie review
example1 = BeautifulSoup(train["review"][0])
print example1.get_text()
```

Punctuation and numbers also don't help in deciding the sentiment of a review. We will remove them using the package **re** (regular expression). **Re** is a built in python package. See the package documentation to complete the next task.

Answer:
To clean the data from HTML tags we I have reused the above code in the following manner:

```
for i in range(len(train)):
    train["review"][i] = BeautifulSoup(train["review"][i]).get_text()
print(train["review"][0])
```

```
"With all this stuff going down at the moment with MJ i've started listening to his music, watching the odd documentary here an
d there, watched The Wiz and watched Moonwalker again. Maybe i just want to get a certain insight into this guy who i thought w
as really cool in the eighties just to maybe make up my mind whether he is guilty or innocent. Moonwalker is part biography, pa
rt feature film which i remember going to see at the cinema when it was originally released. Some of it has subtle messages abo
ut MJ's feeling towards the press and also the obvious message of drugs are bad m'kay.Visually impressive but of course this is
all about Michael Jackson so unless you remotely like MJ in anyway then you are going to hate this and find it boring. Some may
call MJ an egotist for consenting to the making of this movie BUT MJ and most of his fans would say that he made it for the fan
s which if true is really nice of him.The actual feature film bit when it finally starts is only on for 20 minutes or so exclud
ing the Smooth Criminal sequence and Joe Pesci is convincing as a psychopathic all powerful drug lord. Why he wants MJ dead so
bad is beyond me. Because MJ overheard his plans? Nah, Joe Pesci's character ranted that he wanted people to know it is he who
is supplying drugs etc so i dunno, maybe he just hates MJ's music.Lots of cool things in this like MJ turning into a car and a
robot and the whole Speed Demon sequence. Also, the director must have had the patience of a saint when it came to filming the
kiddy Bad sequence as usually directors hate working with one kid let alone a whole bunch of them performing a complex dance sc
ene.Bottom line, this movie is for people who like MJ on one level or another (which i think is most people). If not, then stay
away. It does try and give off a wholesome message and ironically MJ's bestest buddy in this movie is a girl! Michael Jackson i
s truly one of the most talented people ever to grace this planet but is he guilty? Well, with all the attention i've gave this
subject....hmmm well i don't know because people can be different behind closed doors, i know this for a fact. He is either an
extremely nice but stupid guy or one of the most sickest liars. I hope he is not the latter."
```

Task 2:
Use re package to find every thing that is not a lowercase letter or upper case letter and replace it with a space for each review in the training data.
For example the following code finds the alphabet a and v and replaces it with b.

```
import re
example = 'This is a car, very good car'
example_ = re.sub('[av]',"b",example)
print example_
```

Solution: To remove the non-letters I have used the following regex:

example = re.sub('[^a-zA-Z]+', ' ', example)

```
' With all this stuff going down at the moment with MJ i ve started listening to his music watching the odd documentary here an
d there watched The Wiz and watched Moonwalker again Maybe i just want to get a certain insight into this guy who i thought was
really cool in the eighties just to maybe make up my mind whether he is guilty or innocent Moonwalker is part biography part fe
ature film which i remember going to see at the cinema when it was originally released Some of it has subtle messages about MJ
s feeling towards the press and also the obvious message of drugs are bad m kay Visually impressive but of course this is all a
bout Michael Jackson so unless you remotely like MJ in anyway then you are going to hate this and find it boring Some may call
MJ an egotist for consenting to the making of this movie BUT MJ and most of his fans would say that he made it for the fans whi
ch if true is really nice of him The actual feature film bit when it finally starts is only on for minutes or so excluding the
Smooth Criminal sequence and Joe Pesci is convincing as a psychopathic all powerful drug lord Why he wants MJ dead so bad is be
yond me Because MJ overheard his plans Nah Joe Pesci s character ranted that he wanted people to know it is he who is supplying
drugs etc so i dunno maybe he just hates MJ s music Lots of cool things in this like MJ turning into a car and a robot and the
whole Speed Demon sequence Also the director must have had the patience of a saint when it came to filming the kiddy Bad sequen
ce as usually directors hate working with one kid let alone a whole bunch of them performing a complex dance scene Bottom line
this movie is for people who like MJ on one level or another which i think is most people If not then stay away It does try and
give off a wholesome message and ironically MJ s bestest buddy in this movie is a girl Michael Jackson is truly one of the most
talented people ever to grace this planet but is he guilty Well with all the attention i ve gave this subject hmmm well i don t
know because people can be different behind closed doors i know this for a fact He is either an extremely nice but stupid guy o
r one of the most sickest liars I hope he is not the latter '
```

TOKENIZATION: We will also convert every thing into lower case and split the reviews into individual words using following commands

words = example.lower().split()

Finally, we need to decide how to deal with frequently occurring words that don't carry much meaning. Such words are called stop words; in English they include words such as "a", "and", "is", and "the". We will use Natural Language Toolkit (nltk) http://www.nltk.org package for this purpose. First install the package and download the stop word list as follows.

pip install nltk

Now execute following in the shell

```
import nltk
nltk.download('stopwords')
from nltk import stopwords
print stopwords.words('english')
```

Now remove the stop words from all the reviews. The following will remove the stop words from the variable 'words'. Remember 'words' contains tokenized review. **Understand the syntax below (how for loop is used.)**

```
stops = set(stopwords.words('english'))
words = [w for w in words if not w in stops]
```

print words

Solution:
Stop words removed

```
['stuff', 'going', 'moment', 'mj', 'started', 'listening', 'music', 'watching', 'odd', 'documentary', 'watched', 'wiz', 'watch
d', 'moonwalker', 'maybe', 'want', 'get', 'certain', 'insight', 'guy', 'thought', 'really', 'cool', 'eighties', 'maybe', 'mak
e', 'mind', 'whether', 'guilty', 'innocent', 'moonwalker', 'part', 'biography', 'part', 'feature', 'film', 'remember', 'going'
'see', 'cinema', 'originally', 'released', 'subtle', 'messages', 'mj', 'feeling', 'towards', 'press', 'also', 'obvious', 'mess
ge', 'drugs', 'bad', 'kay', 'visually', 'impressive', 'course', 'michael', 'jackson', 'unless', 'remotely', 'like', 'mj', 'any
ay', 'going', 'hate', 'find', 'boring', 'may', 'call', 'mj', 'egotist', 'consenting', 'making', 'movie', 'mj', 'fans', 'would'
'say', 'made', 'fans', 'true', 'really', 'nice', 'actual', 'feature', 'film', 'bit', 'finally', 'starts', 'minutes', 'excludin
g', 'smooth', 'criminal', 'sequence', 'joe', 'pesci', 'convincing', 'psychopathic', 'powerful', 'drug', 'lord', 'wants', 'mj',
'dead', 'bad', 'beyond', 'mj', 'overheard', 'plans', 'nah', 'joe', 'pesci', 'character', 'ranted', 'wanted', 'people', 'know',
'supplying', 'drugs', 'etc', 'dunno', 'maybe', 'hates', 'mj', 'music', 'lots', 'cool', 'things', 'like', 'mj', 'turning', 'ca
r', 'robot', 'whole', 'speed', 'demon', 'sequence', 'also', 'director', 'must', 'patience', 'saint', 'came', 'filming', 'kidd
y', 'bad', 'sequence', 'usually', 'directors', 'hate', 'working', 'one', 'kid', 'let', 'alone', 'whole', 'bunch', 'performing'
'complex', 'dance', 'scene', 'bottom', 'line', 'movie', 'people', 'like', 'mj', 'one', 'level', 'another', 'think', 'people',
'stay', 'away', 'try', 'give', 'wholesome', 'message', 'ironically', 'mj', 'bestest', 'buddy', 'movie', 'girl', 'michael', 'ja
kson', 'truly', 'one', 'talented', 'people', 'ever', 'grace', 'planet', 'guilty', 'well', 'attention', 'gave', 'subject', 'hmm
m', 'well', 'know', 'people', 'different', 'behind', 'closed', 'doors', 'know', 'fact', 'either', 'extremely', 'nice', 'stupi
d', 'guy', 'one', 'sickest', 'liars', 'hope', 'latter']
```

Task 4
Multiply each element of the list A = {2, 3, 4, 5, 7, 8, 9, 2, 5} with 5 using the for loop syntax above.
Multiply each element of the list A = {2, 3, 4, 5, 7, 8, 9, 2, 5}, except 2, with 5 using the for loop syntax above.

Now join the words back into one string separated by space.
sentence = " ".join(words);

Solution:
Joined the separated words:

```
stuff going moment mj started listening music watching odd documentary watched wiz watched moonwalker maybe want get certain in
sight guy thought really cool eighties maybe make mind whether guilty innocent moonwalker part biography part feature film reme
mber going see cinema originally released subtle messages mj feeling towards press also obvious message drugs bad kay visually
impressive course michael jackson unless remotely like mj anyway going hate find boring may call mj egotist consenting making m
ovie mj fans would say made fans true really nice actual feature film bit finally starts minutes excluding smooth criminal sequ
ence joe pesci convincing psychopathic powerful drug lord wants mj dead bad beyond mj overheard plans nah joe pesci character r
anted wanted people know supplying drugs etc dunno maybe hates mj music lots cool things like mj turning car robot whole speed
demon sequence also director must patience saint came filming kiddy bad sequence usually directors hate working one kid let alo
ne whole bunch performing complex dance scene bottom line movie people like mj one level another think people stay away try giv
e wholesome message ironically mj bestest buddy movie girl michael jackson truly one talented people ever grace planet guilty w
ell attention gave subject hmmm well know people different behind closed doors know fact either extremely nice stupid guy one s
ickest liars hope latter
```

Task 5
How can you join the words back into one string separated by colon (:)?

Solution:
Already done in task 2.

Task 6
You have learned how to take a review, remove HTML tags, remove punctuations, convert it to lower case, split it into words, remove stop words and finally join the words back separated by

space. Write a function that combines all these steps so that you can reuse that for all the reviews.

def review_to_words(raw_review)
#1. Remove HTML
#2. Remove non letters
#3. Convert to lowercase and split it into words
#4. Remove stops words
#5. Joint back and return the joined sentence

Solution:

I have merged all the above functionality in one function below. Basically, I am changing the train variables review column to reuse it for the training.

```
def review_to_words(train):
    for i in range(len(train)):
        train["review"][i] = BeautifulSoup(train["review"][i]).get_text()
        train["review"][i] = re.sub('[^a-zA-Z]+', ' ', train["review"][i])
        words = train["review"][i].lower().split()
        words = [w for w in words if not w in stops]
        sentence = " ".join(words)
        train["review"][i] = sentence
    return train
```

Task 7
Run the above function for each review in your training data and store the output in one list.

In the next lab you will use this list to create a Bag of Words representation and machine learning for sentiment analysis.

Deliverables: Upload Word file containing all the tasks.
Time: **End of lab.**