AURORA NLP

IMPORTING LIBS

In [1]:

```
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.tokenize import RegexpTokenizer
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

In [1]:

```
import textract
import re
import PyPDF2
# import fitz
```

FUNCTION: ARRAY STEMMER

In [2]:

```
def arrayStemmer(textArray):
    lematized_wordArray=[]
    porter_wordArray=[]
    porter = PorterStemmer()
    for text in textArray:
        text=text.lower()
        porter_wordArray.append(porter.stem(text))
    return (porter_wordArray)
```

FUNCTION: STEMMING TEXT FROM PAGE

In [3]:

```
def textAnalyze(text):
    text=text.lower()
#     tokenized_word=word_tokenize(text)
    tokenizer = RegexpTokenizer(r'\w+')
    tokenized_word=tokenizer.tokenize(text)
    stop_words=set(stopwords.words("english"))
    filtered_sent=[]
    for w in tokenized_word:
        if w not in stop_words:
            filtered_sent.append(w)
    porter_wordArray=[]
    porter = PorterStemmer()
    for text in filtered_sent:
        text=text.lower()
        porter_wordArray.append(porter.stem(text))
    finalWords=[]
    for w in porter_wordArray:
        try:
            width = float(w)
        except ValueError:
            finalWords.append(w)
    return finalWords
```

FUNCTION: SCORING SENTENCES BASED ON BAG OF WORDS

In [4]:

```
def scoreGenie(bag_of_words, stemmed_array):
    final_score = 0
    frequency_score = 0
    match_score = 0
    match_array = []
    for word in stemmed_array:
        for w in bag_of_words:
            if word == w:
                match_score+=1
                match_array.append(w)
    frequency_score = len(set(match_array))
    final_score = frequency_score * match_score
    return final_score
```

FUNCTION: CREATE PDFS OF INDIVIDUAL PAGES

In [5]:

```
def createPDF(pageNo):
    pdf = PyPDF2.PdfFileReader(pdf_document)
    pdf_writer = PyPDF2.PdfFileWriter()
    pdf_writer.addPage(pdf.getPage(pageNo))
    output = f'temp/{pageNo}.pdf'
    with open(output, 'wb') as output_pdf:
            pdf_writer.write(output_pdf)
```

READING/PARSING PDF

In [6]:

```
pdf_document = "./DEL/Del Monte Pacific Ltd AR 2018.pdf" #Change name here for other PDF
```

In [7]:

```
doc = fitz.open(pdf_document)
print ("Number of pages: %i" % doc.pageCount)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-7-0f969109e5c2> in <module>
----> 1 doc = fitz.open(pdf_document)
      2 print ("Number of pages: %i" % doc.pageCount)

NameError: name 'fitz' is not defined
```

BAG OF WORDS

In [165]:

```
profit_bag_of_words = ['Profit','Income','Margins','Gross','Margin','profits','Revenues','I
redflag_bag_of_words = ['disputes', 'difficulty', 'serious', 'adverse', 'unexpected', 'irre
cashflow_bag_of_words = ['cash flows', 'cash flow', 'free cash flow increase', 'free cash f
revenue_bag_of_words = ['New revenue','Expansion','Acquisitions','Acquired','Growth','Earni

test_bag_of_words = profit_bag_of_words #Change here for the required bag of words
bag_of_words = arrayStemmer(test_bag_of_words)
bag_of_words = list(set(bag_of_words))
print(bag_of_words)
```

```
['improv', 'achiev', 'decreas', 'incom', 'gain', 'increas', 'net', 'greate
r', 'better', 'profit', 'revenu', 'margin', 'gross', 'earn']
```

SEARCHING THROUGH PAGES

In [166]:

```
matchedPageArray = []
for current_page in range(len(doc)):
    page = doc.loadPage(current_page)
    for word in test_bag_of_words:
        if page.searchFor(word):
            matchedPageArray.append(current_page)
matchedPageArray = list(set(matchedPageArray))
```

In [167]:

```
print(matchedPageArray)
```

```
[1, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 17, 18, 21, 23, 25, 27, 28, 31, 3
9, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 54, 56, 57, 59, 60, 61, 62, 6
3, 64, 66, 68, 69, 72, 73, 78, 80, 81, 82, 83, 84, 85, 87, 88, 89, 90, 91, 9
2, 93, 96, 97, 98, 99, 101, 102, 103, 104, 105, 106, 107, 108, 109, 111, 11
2, 113, 114, 115, 116, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 12
8, 129, 130, 133, 141, 143, 144, 145, 147, 148, 149, 150, 151, 152, 153, 15
4, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 16
9, 170, 171, 172, 173, 174, 176, 183, 184, 185, 186, 187, 188, 189, 190, 19
1, 192, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 21
1, 212, 213, 215, 216, 217, 219, 220, 221, 222, 223, 224, 225, 227, 229, 23
0, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 247, 24
8, 249, 250, 251, 254, 257]
```

SPLITTING MATCHED PAGES INTO SEPARATE PDFS

In [168]:

```
import shutil
import os

shutil.rmtree('temp')
if not os.path.exists('temp'):
    os.mkdir('temp')

for page in matchedPageArray:
    createPDF(page)
```

PARSING TEXT FROM INDIVIDUAL PDFS

In [169]:

```
paraDictionary = []
for page in matchedPageArray:
    processedText = textract.process(f'temp/{page}.pdf')
    paragraphArray = processedText.decode().split('\n\n')
    for paragraph in paragraphArray:
        paraObj = {
            "page": page,
            "weight": scoreGenie(bag_of_words,textAnalyze(paragraph)),
            "paragraph": paragraph
        }
        paraDictionary.append(paraObj)
```

SORTING THE PARAGRAPHS BASED ON WEIGHTS & GETTING THE TOP 5 PARAGRAPHS

In [170]:

```
paraDictionary = [i for n, i in enumerate(paraDictionary) if i not in paraDictionary[n + 1:
```

In [171]:

```
sortedParaDictionary = sorted(paraDictionary, key = lambda i: i['weight'],reverse=True)
```

In [172]:

```
top5Paragraphs = sortedParaDictionary[:5]
```

In [173]:

```
for item in top5Paragraphs:
    item['paragraph'] = item['paragraph'].replace('\n','')
```

In [174]:

```
top5Paragraphs
```

Out[174]:

```
[{'page': 124,
  'weight': 60,
  'paragraph': 'from device repayment plans. The increase was driven by stro
ng customer additions in mobile and fixed broadband, increased Equipment sal
es and higher National Broadband Network (NBN) migration revenues despite th
e temporary suspension in connecting and migrating customers to NBN’s HFC ne
twork. Outgoing mobile service revenue rose 1.7% and would be up 5.7% exclud
ing the service credits. Optus gained mobile market share with net addition
of 384,000 customers, underpinned by its investments in network and content.
Mass Market Fixed revenue grew 9.4% driven by higher NBN revenue from net ad
dition of 225,000 customers for the year. With higher operating revenue and
increase in other income from a dispute settlement, EBITDA grew by 4.0%.'},
 {'page': 119,
  'weight': 48,
  'paragraph': 'FY 2018The Group delivered record earnings for FY 2018 with
net profit of S$5.45 billion bolstered by exceptional gain of S$2.03 billion
from the divestment of NetLink Trust and a strong core performance. Operatin
g revenue was S$17.53 billion, 4.9% higher than FY 2017, while EBITDA rose
1.8% to S$5.09 billion reflecting strong customer gains in Australia and fir
st time contribution from Turn (acquired by Amobee in April 2017). In consta
nt currency terms, operating revenue and EBITDA increased by 4.7% and 1.5% r
espectively. '},
 {'page': 119,
  'weight': 48,
  'paragraph': 'The associates’ pre-tax contributions rose 8.2% to S$2.79 bi
llion and would have increased 9.7% excluding the currency translation impac
t. The regional associates recorded strong customer growth and robust mobile
data growth, with higher earnings from Telkomsel and Globe offsetting the de
cline in Airtel.  Underlying net profit was stable and net profit including
exceptional items increased 2.4% to S$3.87 billion. In constant currency ter
ms, underlying net profit and net profit would have increased 4.0% and 5.5%
respectively from FY 2015. '},
 {'page': 126,
  'weight': 35,
  'paragraph': 'In India, Airtel’s results were adversely impacted by intens
e competition with aggressive pricing by a new player and further aggravated
by mandated cuts in mobile termination rates, despite recording strong custo
mer additions and data usage growth. Consequently, Airtel’s revenue in India
fell 13% led by a drop in mobile revenue partly mitigated by growth in other
segments. EBITDA correspondingly declined 22%. In Africa, operating revenue
was stable in constant US Dollar terms and would have increased 5% across th
e 14 countries if excluding the divested operations, led by strong growth in
data and Airtel Money services. EBITDA was up a significant 46% with continu
ed strong cost control initiatives and efficiency gains, as well as improved
margins. '},
 {'page': 10,
  'weight': 30,
  'paragraph': 'ACCELERATING OUR DIGITAL TRANSFORMATIONThe past year is the
sixth since we embarked on our transformation journey, crossing the threshol
d into digital where disruption is rampant and change is constant. Despite t
he challenging operating environment and intensifying competition, we manage
d to accelerate the build out of our new digital businesses in cyber securit
y and digital marketing, and digitalise and strengthen our core business. Th
e resiliency of our earnings while we accelerated changes to our business sp
eaks to the success of our efforts thus far. Our net profit for FY 2018 was
S$5.45 billion on divestment gains from unlocking the value of NetLink Trust
and a strong performance by our core business. Our ICT and new digital busin
esses now represent a meaningful 24% of Group revenue and have helped change
our revenue profile.'}]
```

GETTING PARAGRAPH HEADINGS

In [176]:

```
for obj in top5Paragraphs:
    page = obj['page']
    print(page)
    pText = textract.process(f"/Users/abhigyansingh/Documents/aurora/aurora-nlp/temp/{page}
    pText = pText.decode()
    pText = pText.split('\n')
    print(pText[0])
```

```
124
GROUP CONSUMER
119
5-YEAR FINANCIAL REVIEW
119
5-YEAR FINANCIAL REVIEW
126
The regional associates continued
10
Dear Shareholders,
```

In [184]:

```
textract.process(f"/Users/abhigyansingh/Documents/aurora/aurora-nlp/temp/125.pdf")
```

Out[184]:

```
b"ASSOCIATES\n\nFinancial Year Ended 31 March\n\nGroup share of associates'
pre-tax profits (2)\nShare of post-tax profits \n Telkomsel \n AIS (2)\n Glo
be (3)\n- ordinary results\n- exceptional items \n\nIntouch (3) (4)\n- opera
ting results \n- amortisation of acquired intangibles\n\n Airtel (3)\n- ordi
nary results (India and South Asia) \n- ordinary results (Africa) \n- except
ional items\n\nBTL (5)\n\nRegional associates (2)\n\nNetLink NBN Trust/ NetL
ink Trust (6)\nOther associates \n\nGroup share of associates' post-tax prof
its (2)\n\n\xe2\x80\x9cnm\xe2\x80\x9d denotes not meaningful.\n\n2018\n(S$ m
iIlion)\n 2,461 \n\n 1,031 \n 292 \n\n 180 \n 22 \n 202 \n\n 106 \n (21)\n 8
6 \n\n (31)\n 145 \n (13)\n 101 \n (18)\n 83 \n 1,694 \n72 \n57 \n\n1,823 \n
\n2017\n(S$ million)\n 2,886 \n\n 1,071 \n 278 \n\n 208 \n - \n 208 \n\n 35
\n (7)\n 28 \n\n 364 \n (102)\n - \n 262 \n 8 \n 270 \n 1,855 \n130 \n64 \n
\n2,048 \n\nChange in \nconstant \ncurrency  \n(%)\n-13.5\n\n(1 )\n\nChange
(%)\n-14.7\n\n-3.7\n4.9\n\n-13.5\nnm\n-2.7\n\n204.0\n210.6\n202.5\n\nnm\nnm
\nnm\n-61.5\nnm\n-69.1\n-8.7\n-45.0\n-9.8\n\n-11.0\n\n-0.8\n0.4\n\n-7.1\nnm
\n4.6\n\n198.1\n207.5\n195.9\n\nnm\nnm\nnm\n-62.0\nnm\n-69.5\n-7.1\n-45.0\n-
9.8\n\n-9.6\n\nNotes:\n( 1 )     Assuming constant exchange rates for the re
gional currencies (Indian Rupee, Indonesian Rupiah, Philippine Peso and Thai
Baht) from FY 2017. \n(2)     The share of AIS\xe2\x80\x99 3G/4G handset sub
sidy costs in FY 2017 previously classified as exceptional items of the Grou
p have been reclassified to share of AIS\xe2\x80\x99 ordinary \n\nresults to
be consistent with FY 2018. \n\n(3)     Excluded the Group\xe2\x80\x99s shar
e of the associates\xe2\x80\x99 certain one-off items which have been classi
fied as exceptional items of the Group. \n(4)     Intouch, which Singtel acq
uired an equity interest of 21% in November 2016, has an equity interest of
40.5% in AIS. \n(5)     Bharti Telecom Limited (BTL) holds 50.1% equity inte
rest in Airtel as at 31 March 2018. In BTL\xe2\x80\x99s standalone books, it
s results for FY 2018 comprised mainly interest \n\ncharges on debt arising
from its acquisition of additional equity interest in Airtel.  \n\n(6)     S
ingtel ceased to own units in NetLink Trust following the sale to NetLink NB
N Trust in July 2017 but continues to have an interest of 24.8% in NetLink N
BN Trust, the \nholding company of NetLink Trust. The share of results inclu
ded Singtel\xe2\x80\x99s amortisation of deferred gain of S$26 million (FY 2
017: S$52 million) on assets transferred \nto NetLink Trust in prior years,
but excluded fair value adjustments recorded by NetLink NBN Trust in respect
of its acquisition of units in NetLink Trust.\n\nCountry mobile penetration
rate \nMarket share, 31 March 2018 (2)\nMarket share, 31 March 2017 (2)\nMar
ket position (2)\n\nMobile customers ('000) \n- Aggregate \n- Proportionate
\nGrowth in mobile customers (%) (3) \n\nTelkomsel\n154%\n47.0%\n46.0%\n#1\n
\n192,752\n67,463\n13.8%\n\nAIS\n136%\n44.8%\n44.8%\n#1\n\n 40,050 \n 9,340
\n-1.5%\n\nAirtel (1 )\n89%\n25.6%\n23.4%\n#1\n\n 395,722 \n 156,350 \n11.3%
\n\nGlobe\n116%\n52.1%\n48.1%\n#1\n\n 63,263 \n 29,816 \n8.0%\n\nNotes:\n( 1
)     Mobile penetration rate, market share and market position pertained to
India market only.\n(2)     Based on number of mobile customers.\n(3)     Co
mpared against 31 March 2017 and based on aggregate mobile customers. \n\n12
3\n\nManagement Discussion and Analysis\x0c"
```

SENTIMENT ANALYSIS

In [107]:

```
analyser = SentimentIntensityAnalyzer()
```

In [108]:

```
def sentiment_analyzer_scores(sentence):
    score = analyser.polarity_scores(sentence)
    print(score)
```

In [141]:

```
sentiment_analyzer_scores(top5Paragraphs[0]['paragraph'])
```

```
{'neg': 0.017, 'neu': 0.799, 'pos': 0.184, 'compound': 0.9882}
```