STA 6714: Data Preparation

Assignment 4: Text Analytics- Extracting a social media data

Dataset used: BBC News Archive

It is a collection of BBC News Content and their associated labels. Data has the news title, the related text file name along with news content and its category. Dataset consists of 2225 documents related to 5 categories such as business, entertainment, politics, sport, tech (class labels) from the BBC news website from 2004 to 2005.

[6]:
```
data = pd.read_csv("../input/bbcnewsarchive/bbc-news-data.csv",sep='\t')
```

```
data.drop('filename',axis=1,inplace=True)
data.head()
```

[7]:

	category	title	content
0	business	Ad sales boost Time Warner profit	Quarterly profits at US media giant TimeWarne...
1	business	Dollar gains on Greenspan speech	The dollar has hit its highest level against ...
2	business	Yukos unit buyer faces loan claim	The owners of embattled Russian oil giant Yuk...
3	business	High fuel prices hit BA's profits	British Airways has blamed high fuel prices f...
4	business	Pernod takeover talk lifts Domecq	Shares in UK drinks and food firm Allied Dome...

[13]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2225 entries, 0 to 2224
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   category  2225 non-null   object
 1   title     2225 non-null   object
 2   content   2225 non-null   object
dtypes: object(3)
memory usage: 52.3+ KB
```

[11]:

```
px.pie(data.category.value_counts().to_frame().reset_index(),
       values='category',names='index',
       color_discrete_sequence=px.colors.sequential.Darkmint,
      title = 'Count and frequency news category in DataFrame',
      labels={'category':'count','index':'category'})
```