Capstone Project Report

Name: **Pooja Ramdas Kadam**
Course: AI & ML (Batch - 4)

Problem Statement
Perform topic modelling using the **20 Newsgroup dataset** (the dataset is also available in sklearn datasets sub-module). Perform the required data cleaning steps using NLP and then model the topics

1. Using Latent Dirichlet Allocation (LDA)
2. Using Probabilistic Latent Semantic Analysis (PLSA)

Prerequisites
Along with Python below packages needed to be installed

Pandas
Sklearn

Dataset Used

Sklearn dataset 20newsgroups

Implementation
Import required libraries and load data

```
In [9]: from sklearn.datasets import fetch_20newsgroups
        from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
        from sklearn.decomposition import NMF, LatentDirichletAllocation
        import pandas as pd
```

Load data

```
In [19]: data, _ = fetch_20newsgroups(shuffle=True, random_state=1,
                                     remove=('headers', 'footers', 'quotes'),
                                     return_X_y=True)
         data_samples = data[:n_samples]
```

Vectorize using TFIDVectorizer

In [22]:
```
tfidf_vectorizer = TfidfVectorizer(max_df=0.95, min_df=2, max_features=n_features, stop_words='english')
tfidf = tfidf_vectorizer.fit_transform(data_samples)
```

Implement PLSA

```
nmf = NMF(n_components=n_components, random_state=1,
          beta_loss='kullback-leibler', solver='mu', max_iter=1000, alpha=.1,
          l1_ratio=.5).fit(tfidf)

tfidf_feature_names = tfidf_vectorizer.get_feature_names()

word_dict ={}
for topic_idx, topic in enumerate(nmf.components_):
        top_features_ind = topic.argsort()[:-n_top_words - 1:-1]
        top_features = [tfidf_feature_names[i] for i in top_features_ind]
        word_dict['Topic'+str(topic_idx)] = top_features
topics = pd.DataFrame(word_dict)
print('\n\n PLSA: \n\n',topics.head(10))
```

```
 PLSA:

    Topic0   Topic1      Topic2      Topic3 Topic4      Topic5    Topic6  \
0  people  windows         god      thanks     10       space       edu
1     don     help        does        know     00  government      file
2    just   thanks       jesus        bike   sale      number       com
3    like    using        true  interested   time      public   program
4   think       hi        book        mail  power        data      soon
5     did  looking   christian        like     12      states       try
6     say     info       bible         new    new       earth    window
7    time    video  christians         car     15    security   problem
8    make      dos    religion         edu   year       water  remember
9    know       pc       faith       heard     30    research     files

     Topic7    Topic8   Topic9
0      game     drive      use
1      team     think     good
2      year      hard     just
3     games  software      key
4      play      disk     chip
5       win    drives      got
6    season     apple     like
7    points  computer       ll
8     world       mac      way
9  division      need  clipper
```

Implement LDA

In [25]:
```
lda = LatentDirichletAllocation(
    n_components=n_components,
    max_iter=5,
    learning_method='online',
    learning_offset=50.,
    random_state=0
)
topics = pd.DataFrame(word_dict)
lda.fit(tfidf)


tf_feature_names = tf_vectorizer.get_feature_names()

word_dict ={}
for topic_idx, topic in enumerate(lda.components_):
        top_features_ind = topic.argsort()[:-n_top_words - 1:-1]
        top_features = [tf_feature_names[i] for i in top_features_ind]
        word_dict['Topic'+str(topic_idx)] = top_features
topics = pd.DataFrame(word_dict)
print('\n\n LDA: \n\n',topics.head(10))
```

LDA:

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9
0	windows	dog	god	pp	2nd	cases	mike	assume	just	wanted
1	thanks	attack	accept	22	math	00	love	magi	don	1992
2	file	head	read	18	ground	soon	graphics	home	people	gm
3	edu	drive	clock	19	value	edu	hear	order	like	season
4	use	talking	driving	11	said	sale	heard	card	think	1993
5	drive	human	stuff	23	leafs	condition	try	right	know	wouldn
6	software	maybe	nature	26	display	effective	state	lot	good	vs
7	mail	disk	port	55	long	asking	time	supposed	god	john
8	help	printer	think	van	try	consider	looking	hit	time	30
9	does	drivers	error	10	lunar	good	good	better	new	price