In [45]:
```
import json
import pandas as pd
import numpy as np
from tqdm import tqdm_notebook as tqdm
import os
```

In [46]:
```
dir_bb = 'data/years/'
filelist = os.listdir(dir_bb)
master_df = pd.DataFrame()
```

In [47]:
```
for year in tqdm(filelist):
    path = dir_bb + year
    with open(path, 'r', encoding='cp850', errors = 'ignore') as f:
        j = json.load(f)
        temp_df = pd.DataFrame(j)
        temp_df['pos'] = temp_df['sentiment'].apply(lambda x : x['pos'])
        temp_df['neg'] = temp_df['sentiment'].apply(lambda x : x['neg'])
        master_df = pd.concat([master_df,temp_df], sort=False)

master_df.head()
```

Out[47]:

	lyrics	tags	num_syllables	pos	year	fog_index	flesch_index	num_words
0	Mona Lisa, Mona Lisa, men have named you\nYou'...	[american, death by lung cancer, easy listenin...	189.0	0.199	1950	5.2	88.74	145
1	I wanna be Loved\nBy Andrews Sisters\n\nOooo-o...	[andrews sisters]	270.9	0.224	1950	4.4	82.31	189
2	I was dancing with my darling to the Tennessee...	[country, pop]	174.6	0.351	1950	5.2	88.74	138
3	Each time I hold someone new\nMy arms grow col...	[death by liver failure, spiritual]	135.9	0.231	1950	4.4	99.23	117
4	Unfortunately, we are not licensed to display ...	[country, pop]	46.8	0.079	1950	6.0	69.79	32

In [48]:
```
cleaned_df = master_df.drop(['flesch_index','f_k_grade','difficult_words','fog_index','num_dupes','num_lines','num_syllables','num_words','tags'],
                            axis=1)
```

In [59]:
```
final_df = cleaned_df[['artist','title','year','pos','neg','lyrics']]
```

In [65]:
```
final_df.to_csv(index=False, path_or_buf = 'song_sentiment.csv')
```