Exercise 2

In [2]:

```
import nltk
from nltk.corpus import movie_reviews
import matplotlib.pyplot as plt
import random
```

In [3]:

```
nltk.download('movie_reviews')
```

```
[nltk_data] Downloading package movie_reviews to
[nltk_data]     /home/nihar/nltk_data...
[nltk_data]   Unzipping corpora/movie_reviews.zip.
```

Out[3]:

```
True
```

In [4]:

```
print (movie_reviews.categories())
```

```
['neg', 'pos']
```

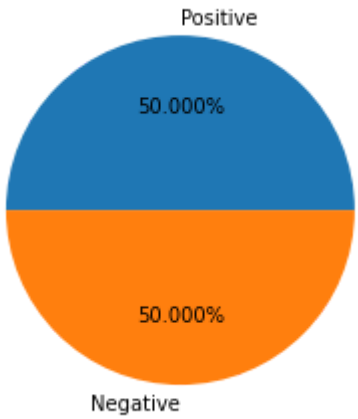
In [5]:

```
print (len(movie_reviews.fileids('pos')))
print (len(movie_reviews.fileids('neg')))
```

```
1000
1000
```

In [6]:

```
labels=['Positive','Negative']
list_count=[len(movie_reviews.fileids('pos')),len(movie_reviews.fileids('neg'))]
plt.pie(list_count,labels=labels,autopct='%.3f%%')
plt.show()
```



In [8]:

```
docs = []

for category in movie_reviews.categories():
    for fileid in movie_reviews.fileids(category):
        docs.append((movie_reviews.words(fileid), category))

print (docs[0])
```

```
(['plot', ':', 'two', 'teen', 'couples', 'go', 'to', ...], 'neg')
```

In [9]:

```
all_words = [word.lower() for word in movie_reviews.words()]
print(all_words[:10])
```

```
['plot', ':', 'two', 'teen', 'couples', 'go', 'to', 'a', 'church',
'party']
```

In [10]:

```
import string
nltk.download('stopwords')
from nltk.corpus import stopwords
stopwords_english = stopwords.words('english')
print('Stop Words\n')
print (stopwords_english)
print('\nPunctuation\n')
print(string.punctuation)
```

```
[nltk_data] Downloading package stopwords to /home/nihar/nltk_dat
a...

Stop Words

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you',
"you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself',
'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her',
'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 't
heir', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'thi
s', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was',
'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do',
'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with',
'about', 'against', 'between', 'into', 'through', 'during', 'befor
e', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'ou
t', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'onc
e', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'b
oth', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no',
'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should'v
e", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "are
n't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'h
adn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "is
n't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "need
n't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'w
eren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]

Punctuation

!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~


[nltk_data]   Unzipping corpora/stopwords.zip.
```

In [13]:

```
all_words = [word for word in all_words if word not in stopwords_english and wor
d not in string.punctuation]

print(all_words[:10])
```

```
['plot', 'two', 'teen', 'couples', 'go', 'church', 'party', 'drink',
'drive', 'get']
```

In [14]:

```
from nltk.stem import PorterStemmer

stemmer = PorterStemmer()

review_stem = []
limit_list=all_words[:500]
for word in limit_list:
    stem_word = stemmer.stem(word)
    review_stem.append(stem_word)

print('stemmed words:')
print(review_stem[:10])
```

```
stemmed words:
['plot', 'two', 'teen', 'coupl', 'go', 'church', 'parti', 'drink',
'drive', 'get']
```

In [ ]:

```

```