TITLE : Text mining: Implement a text mining model to analyze customer reviews and identify sentiments (positive, negative, neutral) and preferences using techniques like sentiment analysis and keyword extraction.

NAME : Shinde Shubham Dnyandev,

ROLL NO : 23107121,

BATCH : B.

In [1]:

```
import pandas as pd
import nltk
import re
import spacy
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
```

In [3]:

```
df = pd.read_csv("/home/admin1/flipkart.csv")
df
```

Out[3]:

	review	rating
0	It was nice produt. I like it's design a lot. ...	5
1	awesome sound....very pretty to see this nd th...	5
2	awesome sound quality. pros 7-8 hrs of battery...	4
3	I think it is such a good product not only as ...	5
4	awesome bass sound quality very good bettary l...	5
...	...	...
9971	GoodREAD MORE	5
9972	Everything is amazimg but the built is very li...	5
9973	GoodREAD MORE	5
9974	Best headphone i have ever used....READ MORE	5
9975	NiceREAD MORE	5

9976 rows × 2 columns

In [5]:

```
df = df.drop("rating",axis=1)
df
```

Out[5]:

	review
0	It was nice produt. I like it's design a lot. ...
1	awesome sound....very pretty to see this nd th...
2	awesome sound quality. pros 7-8 hrs of battery...
3	I think it is such a good product not only as ...
4	awesome bass sound quality very good bettary l...
...	...
9971	GoodREAD MORE
9972	Everything is amazimg but the built is very li...
9973	GoodREAD MORE
9974	Best headphone i have ever used....READ MORE
9975	NiceREAD MORE

9976 rows × 1 columns

In [7]:

```python
lemmatizer = WordNetLemmatizer()
```

In [9]:

```python
def preprocess_review(text):
    if pd.isna(text):
        return ""

    text = re.sub(r'<.*?>', ' ', text)

    text = re.sub(r'[^a-zA-Z]', ' ', text)

    text = text.lower()

    words = text.split()

    words = [lemmatizer.lemmatize(word)
             for word in words
             if word not in set(stopwords.words('english'))]

    return ' '.join(words)
```

In [11]:

```python
df['clean_review'] = df['review'].apply(preprocess_review)
```

In [13]:

```python
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('punkt_tab')
```

```
[nltk_data] Downloading package stopwords to /home/admin1/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /home/admin1/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package punkt_tab to /home/admin1/nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
```

Out[13]:

```
True
```

In [15]:

```python
df
```

Out[15]:

	review	clean_review
0	It was nice produt. I like it's design a lot. ...	nice produt like design lot easy carry looked ...
1	awesome sound....very pretty to see this nd th...	awesome sound pretty see nd sound quality good...
2	awesome sound quality. pros 7-8 hrs of battery...	awesome sound quality pro hr battery life incl...
3	I think it is such a good product not only as ...	think good product per quality also design qui...
4	awesome bass sound quality very good bettary l...	awesome bass sound quality good bettary long l...
...	...	...
9971	GoodREAD MORE	goodread
9972	Everything is amazimg but the built is very li...	everything amazimg built light read
9973	GoodREAD MORE	goodread
9974	Best headphone i have ever used....READ MORE	best headphone ever used read
9975	NiceREAD MORE	niceread

9976 rows × 2 columns

In [17]:

```
from textblob import TextBlob

def get_sentiment(text):
    polarity = TextBlob(text).sentiment.polarity
    if polarity > 0.1:
        return "Positive"
    elif polarity < -0.1:
        return "Negative"
    else:
        return "Neutral"

df['predicted_sentiment'] = df['clean_review'].apply(get_sentiment)
```

In [19]:

```
df
```

Out[19]:

	review	clean_review	predicted_sentiment
0	It was nice produt. I like it's design a lot. ...	nice produt like design lot easy carry looked ...	Positive
1	awesome sound....very pretty to see this nd th...	awesome sound pretty see nd sound quality good...	Positive
2	awesome sound quality. pros 7-8 hrs of battery...	awesome sound quality pro hr battery life incl...	Positive
3	I think it is such a good product not only as ...	think good product per quality also design qui...	Positive
4	awesome bass sound quality very good bettary l...	awesome bass sound quality good bettary long l...	Positive
...	...	...	...
9971	GoodREAD MORE	goodread	Neutral
9972	Everything is amazimg but the built is very li...	everything amazimg built light read	Positive
9973	GoodREAD MORE	goodread	Neutral
9974	Best headphone i have ever used....READ MORE	best headphone ever used read	Positive
9975	NiceREAD MORE	niceread	Neutral

9976 rows × 3 columns

In [21]:

```
df['predicted_sentiment'].value_counts()
```

Out[21]:

```
Positive    7112
Neutral     2391
Negative     473
Name: predicted_sentiment, dtype: int64
```

In [23]:

```
from rake_nltk import Rake
```

In [25]:

```
rake = Rake()
```

In [27]:

```
def extract_rake_keywords(text, top_n=5):
    rake.extract_keywords_from_text(text)
    keywords = rake.get_ranked_phrases()
    return keywords[:top_n]
```

In [29]:

```
df['rake_keywords'] = df['clean_review'].apply(extract_rake_keywords)
```

In [31]:

```
df
```

Out[31]:

	review	clean_review	predicted_sentiment	rake_keywords
0	It was nice produt. I like it's design a lot. ...	nice produt like design lot easy carry looked ...	Positive	[nice produt like design lot easy carry looked...
1	awesome sound....very pretty to see this nd th...	awesome sound pretty see nd sound quality good...	Positive	[awesome sound pretty see nd sound quality goo...
2	awesome sound quality. pros 7-8 hrs of battery...	awesome sound quality pro hr battery life incl...	Positive	[awesome sound quality pro hr battery life inc...
3	I think it is such a good product not only as ...	think good product per quality also design qui...	Positive	[think good product per quality also design qu...
4	awesome bass sound quality very good bettary l...	awesome bass sound quality good bettary long l...	Positive	[awesome bass sound quality good bettary long ...
...	...	...	...	...
9971	GoodREAD MORE	goodread	Neutral	[goodread]
9972	Everything is amazimg but the built is very li...	everything amazimg built light read	Positive	[everything amazimg built light read]
9973	GoodREAD MORE	goodread	Neutral	[goodread]
9974	Best headphone i have ever used....READ MORE	best headphone ever used read	Positive	[best headphone ever used read]
9975	NiceREAD MORE	niceread	Neutral	[niceread]

9976 rows × 4 columns

In [33]:

```python
from collections import Counter

def top_preferences(sentiment, n=10):
    keywords = []
    for kws in df[df['predicted_sentiment'] == sentiment]['rake_keywords']:
        keywords.extend(kws)
    return Counter(keywords).most_common(n)
```

In [35]:

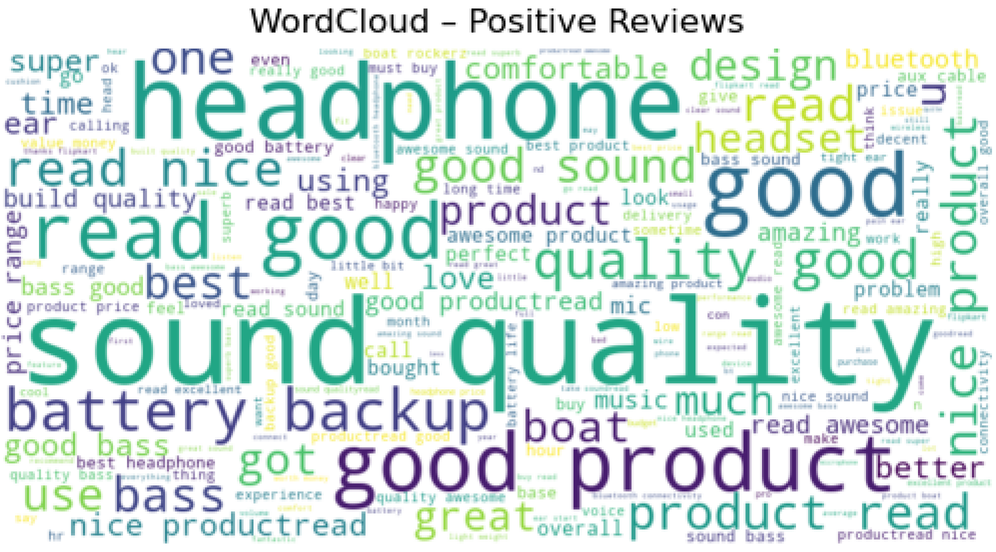
```python
from wordcloud import WordCloud
import matplotlib.pyplot as plt
```

In [37]:

```python
def generate_wordcloud(keywords, title):
    text = " ".join(keywords)

    wc = WordCloud(
        width=800,
        height=400,
        background_color="white"
    ).generate(text)

    plt.figure()
    plt.imshow(wc, interpolation="bilinear")
    plt.axis("off")
    plt.title(title)
    plt.show()
```

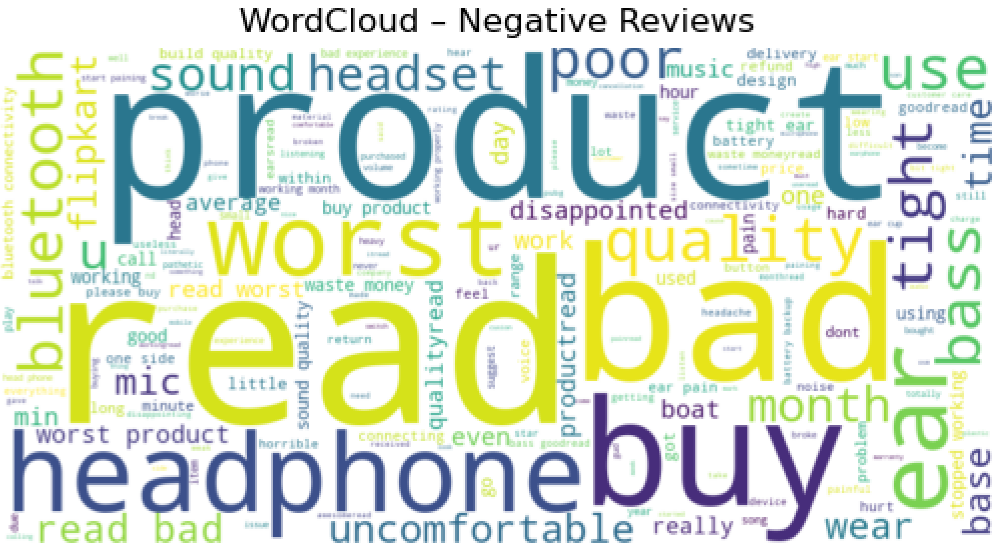
In [39]:

```python
positive_keywords = []
for kws in df[df['predicted_sentiment'] == 'Positive']['rake_keywords']:
    positive_keywords.extend(kws)

generate_wordcloud(positive_keywords, "WordCloud – Positive Reviews")
```



In [41]:

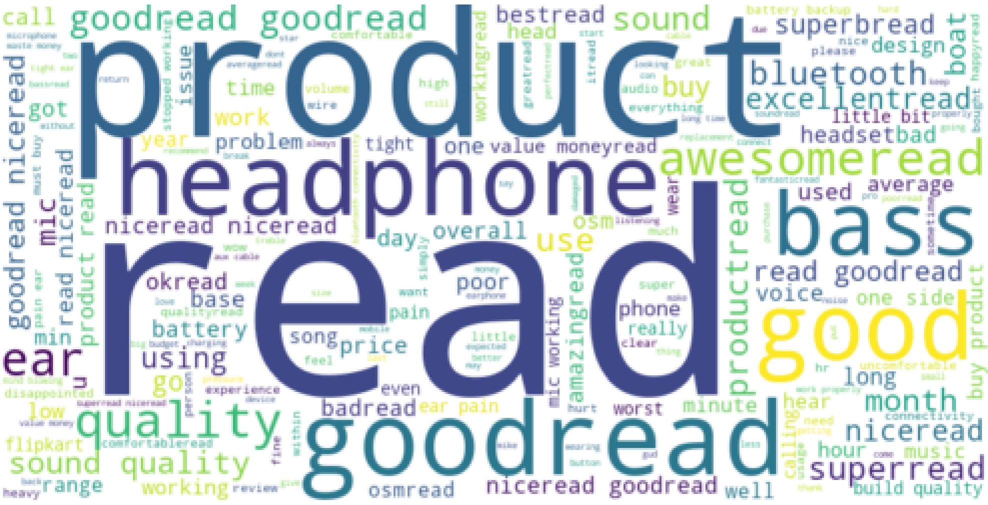
```python
negative_keywords = []
for kws in df[df['predicted_sentiment'] == 'Negative']['rake_keywords']:
    negative_keywords.extend(kws)

generate_wordcloud(negative_keywords, "WordCloud – Negative Reviews")
```



In [43]:

```python
neutral_keywords = []
for kws in df[df['predicted_sentiment'] == 'Neutral']['rake_keywords']:
    neutral_keywords.extend(kws)

generate_wordcloud(neutral_keywords, "WordCloud – Neutral Reviews")
```

WordCloud – Neutral Reviews



In [ ]: