module1-lab

September 11, 2023

[20]:
```
import spacy
import json
from nltk.stem import PorterStemmer
from collections import Counter
import matplotlib.pyplot as plt

# Load the English model
nlp = spacy.load("en_core_web_sm")

# Initialize Porter Stemmer from NLTK
stemmer = PorterStemmer()
```

[21]:
```
# This is just for testing purposes
data_dict = {'pass_summary': 'This is the first summary. This is the second␣
 ↪summary.'}

# Extract the 'pass_summary' field
pass_summary = data_dict['pass_summary']

# Tokenization
doc = nlp(pass_summary)
tokens = [token.text for token in doc]
print("Tokens:", tokens)

# Lemmas
lemmas = [token.lemma_ for token in doc]
print("Lemmas:", lemmas)

# Stop words
stop_words = [token.text for token in doc if token.is_stop]
print("Stop Words:", stop_words)
```

```
Tokens: ['This', 'is', 'the', 'first', 'summary', '.', 'This', 'is', 'the',
'second', 'summary', '.']
Lemmas: ['this', 'be', 'the', 'first', 'summary', '.', 'this', 'be', 'the',
'second', 'summary', '.']
Stop Words: ['This', 'is', 'the', 'first', 'This', 'is', 'the']
```

[22]:
```
# Read JSONL file and process each line
with open('amazon_product_pass_summary.jsonl', 'r') as f:
    for line_number, line in enumerate(f, start=1):
        data = json.loads(line.strip())

        # Extract pass_summary field for text analysis
        pass_summary = data.get('pass_summary', 'Key not found')

        # Initialize Spacy Doc
        doc = nlp(pass_summary)

        # Extract tokens, lemmas, and stop words
        tokens = [token.text for token in doc]
        lemmas = [token.lemma_ for token in doc]
        stop_words_removed = [token.text for token in doc if not token.is_stop]
        stemmed_words = [stemmer.stem(token.text) for token in doc]

        # Display the data
        print(f"===== Line {line_number} =====")
        print(f"Pass Summary: {pass_summary}")
        print(f"Tokens: {' | '.join(tokens)}")
        print(f"Lemmas: {' | '.join(lemmas)}")
        print(f"Stop Words Removed: {' | '.join(stop_words_removed)}")
        print(f"Stemmed Words: {' | '.join(stemmed_words)}")
        print("===== End of Line =====\n")
```

```
===== Line 1 =====
Pass Summary: These soft, breathable tights are great for transitioning from tap
to ballet. They fit snugly around the body and stay in place when worn with
ballet shoes. They are well made and well made, and can last longer than other
tights available. The colors are beautiful and will definitely be purchasing
again.
Tokens: These | soft | , | breathable | tights | are | great | for |
transitioning | from | tap | to | ballet | . | They | fit | snugly | around |
the | body | and | stay | in | place | when | worn | with | ballet | shoes | . |
They | are | well | made | and | well | made | , | and | can | last | longer |
than | other | tights | available | . | The | colors | are | beautiful | and |
will | definitely | be | purchasing | again | .
Lemmas: these | soft | , | breathable | tight | be | great | for | transition |
from | tap | to | ballet | . | they | fit | snugly | around | the | body | and |
stay | in | place | when | wear | with | ballet | shoe | . | they | be | well |
make | and | well | make | , | and | can | last | long | than | other | tight |
available | . | the | color | be | beautiful | and | will | definitely | be |
purchase | again | .
Stop Words Removed: soft | , | breathable | tights | great | transitioning | tap
| ballet | . | fit | snugly | body | stay | place | worn | ballet | shoes | . |
, | longer | tights | available | . | colors | beautiful | definitely |
```

purchasing | .
Stemmed Words: these | soft | , | breathabl | tight | are | great | for | transit | from | tap | to | ballet | . | they | fit | snugli | around | the | bodi | and | stay | in | place | when | worn | with | ballet | shoe | . | they | are | well | made | and | well | made | , | and | can | last | longer | than | other | tight | avail | . | the | color | are | beauti | and | will | definit | be | purchas | again | .
===== End of Line =====

===== Line 2 =====
Pass Summary: This purse is a decent size and the material is of high quality. The stones tend to fall off a lot in use, so it's not recommended for carrying. Overall, it's a nice bag and would definitely recommend for any casual or formal outfit.
Tokens: This | purse | is | a | decent | size | and | the | material | is | of | high | quality | . | The | stones | tend | to | fall | off | a | lot | in | use | , | so | it | 's | not | recommended | for | carrying | . | Overall | , | it | 's | a | nice | bag | and | would | definitely | recommend | for | any | casual | or | formal | outfit | .
Lemmas: this | purse | be | a | decent | size | and | the | material | be | of | high | quality | . | the | stone | tend | to | fall | off | a | lot | in | use | , | so | it | be | not | recommend | for | carry | . | overall | , | it | be | a | nice | bag | and | would | definitely | recommend | for | any | casual | or | formal | outfit | .
Stop Words Removed: purse | decent | size | material | high | quality | . | stones | tend | fall | lot | use | , | recommended | carrying | . | Overall | , | nice | bag | definitely | recommend | casual | formal | outfit | .
Stemmed Words: thi | purs | is | a | decent | size | and | the | materi | is | of | high | qualiti | . | the | stone | tend | to | fall | off | a | lot | in | use | , | so | it | 's | not | recommend | for | carri | . | overal | , | it | 's | a | nice | bag | and | would | definit | recommend | for | ani | casual | or | formal | outfit | .
===== End of Line =====

===== Line 3 =====
Pass Summary: These NuGo bars are high quality and they come in a variety of flavors and sizes which make them ideal for serving as snack or as a replacement for processed foods. They are low glycemic and have a smooth, vanilla-like texture which makes them taste great.
Tokens: These | NuGo | bars | are | high | quality | and | they | come | in | a | variety | of | flavors | and | sizes | which | make | them | ideal | for | serving | as | snack | or | as | a | replacement | for | processed | foods | . | They | are | low | glycemic | and | have | a | smooth | , | vanilla | - | like | texture | which | makes | them | taste | great | .
Lemmas: these | NuGo | bar | be | high | quality | and | they | come | in | a | variety | of | flavor | and | size | which | make | they | ideal | for | serve | as | snack | or | as | a | replacement | for | process | food | . | they | be | low | glycemic | and | have | a | smooth | , | vanilla | - | like | texture |

which | make | they | taste | great | .
Stop Words Removed: NuGo | bars | high | quality | come | variety | flavors | sizes | ideal | serving | snack | replacement | processed | foods | . | low | glycemic | smooth | , | vanilla | - | like | texture | makes | taste | great | .
Stemmed Words: these | nugo | bar | are | high | qualiti | and | they | come | in | a | varieti | of | flavor | and | size | which | make | them | ideal | for | serv | as | snack | or | as | a | replac | for | process | food | . | they | are | low | glycem | and | have | a | smooth | , | vanilla | - | like | textur | which | make | them | tast | great | .
===== End of Line =====

===== Line 4 =====
Pass Summary: This is a high quality product that comes with 120 tablets. The tablets don't break open, so they need to be stored in their original containers. This is a great product for those who have been struggling with depression.
Tokens: This | is | a | high | quality | product | that | comes | with | 120 | tablets | . | The | tablets | do | n't | break | open | , | so | they | need | to | be | stored | in | their | original | containers | . | This | is | a | great | product | for | those | who | have | been | struggling | with | depression | .
Lemmas: this | be | a | high | quality | product | that | come | with | 120 | tablet | . | the | tablet | do | not | break | open | , | so | they | need | to | be | store | in | their | original | container | . | this | be | a | great | product | for | those | who | have | be | struggle | with | depression | .
Stop Words Removed: high | quality | product | comes | 120 | tablets | . | tablets | break | open | , | need | stored | original | containers | . | great | product | struggling | depression | .
Stemmed Words: thi | is | a | high | qualiti | product | that | come | with | 120 | tablet | . | the | tablet | do | n't | break | open | , | so | they | need | to | be | store | in | their | origin | contain | . | thi | is | a | great | product | for | those | who | have | been | struggl | with | depress | .
===== End of Line =====

===== Line 5 =====
Pass Summary: This is a very comfortable and well-designed dining set. It's easy to assemble and comes with a variety of safety features. The only drawback may be that it's fragile and requires some repair before it's ready for shipping. Overall, this dining set is recommended.
Tokens: This | is | a | very | comfortable | and | well | - | designed | dining | set | . | It | 's | easy | to | assemble | and | comes | with | a | variety | of | safety | features | . | The | only | drawback | may | be | that | it | 's | fragile | and | requires | some | repair | before | it | 's | ready | for | shipping | . | Overall | , | this | dining | set | is | recommended | .
Lemmas: this | be | a | very | comfortable | and | well | - | design | dining | set | . | it | be | easy | to | assemble | and | come | with | a | variety | of | safety | feature | . | the | only | drawback | may | be | that | it | be | fragile | and | require | some | repair | before | it | be | ready | for |

shipping | . | overall | , | this | dining | set | be | recommend | .
Stop Words Removed: comfortable | - | designed | dining | set | . | easy | assemble | comes | variety | safety | features | . | drawback | fragile | requires | repair | ready | shipping | . | Overall | , | dining | set | recommended | .
Stemmed Words: thi | is | a | veri | comfort | and | well | - | design | dine | set | . | it | 's | easi | to | assembl | and | come | with | a | varieti | of | safeti | featur | . | the | onli | drawback | may | be | that | it | 's | fragil | and | requir | some | repair | befor | it | 's | readi | for | ship | . | overal | , | thi | dine | set | is | recommend | .
===== End of Line =====

===== Line 6 =====
Pass Summary: This Yuasa battery is of high quality. It holds a charge well and is of high enough quality to be used as a replacement for a dead battery on arrival. The shipping speed is great and the customer service is superb. Would definitely buy from this vendor again.
Tokens: This | Yuasa | battery | is | of | high | quality | . | It | holds | a | charge | well | and | is | of | high | enough | quality | to | be | used | as | a | replacement | for | a | dead | battery | on | arrival | . | The | shipping | speed | is | great | and | the | customer | service | is | superb | . | Would | definitely | buy | from | this | vendor | again | .
Lemmas: this | Yuasa | battery | be | of | high | quality | . | it | hold | a | charge | well | and | be | of | high | enough | quality | to | be | use | as | a | replacement | for | a | dead | battery | on | arrival | . | the | shipping | speed | be | great | and | the | customer | service | be | superb | . | would | definitely | buy | from | this | vendor | again | .
Stop Words Removed: Yuasa | battery | high | quality | . | holds | charge | high | quality | replacement | dead | battery | arrival | . | shipping | speed | great | customer | service | superb | . | definitely | buy | vendor | .
Stemmed Words: thi | yuasa | batteri | is | of | high | qualiti | . | it | hold | a | charg | well | and | is | of | high | enough | qualiti | to | be | use | as | a | replac | for | a | dead | batteri | on | arriv | . | the | ship | speed | is | great | and | the | custom | servic | is | superb | . | would | definit | buy | from | thi | vendor | again | .
===== End of Line =====

===== Line 7 =====
Pass Summary: This is a great alternative to other medications for motion sickness. It works well and doesn't cause any side effects. The tablets are chewable and not irritable, so they can be taken along on a cruise or on a cruise ship.
Tokens: This | is | a | great | alternative | to | other | medications | for | motion | sickness | . | It | works | well | and | does | n't | cause | any | side | effects | . | The | tablets | are | chewable | and | not | irritable | , | so | they | can | be | taken | along | on | a | cruise | or | on | a | cruise | ship | .
Lemmas: this | be | a | great | alternative | to | other | medication | for |

motion | sickness | . | it | work | well | and | do | not | cause | any | side | effect | . | the | tablet | be | chewable | and | not | irritable | , | so | they | can | be | take | along | on | a | cruise | or | on | a | cruise | ship | .
Stop Words Removed: great | alternative | medications | motion | sickness | . | works | cause | effects | . | tablets | chewable | irritable | , | taken | cruise | cruise | ship | .
Stemmed Words: thi | is | a | great | altern | to | other | medic | for | motion | sick | . | it | work | well | and | doe | n't | caus | ani | side | effect | . | the | tablet | are | chewabl | and | not | irrit | , | so | they | can | be | taken | along | on | a | cruis | or | on | a | cruis | ship | .
===== End of Line =====

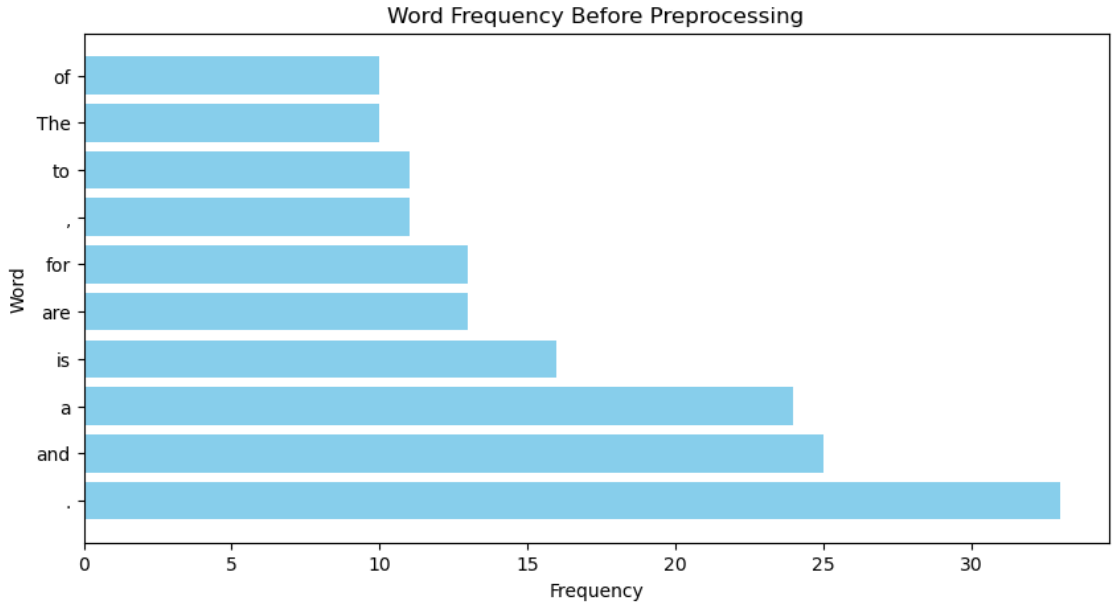
===== Line 8 =====
Pass Summary: This Apple laptop case has a nice, sturdy design and fits laptops with up to a MacBook Air well. The color of the case is great and the design is appealing. The case has retractable feet which help to keep the device secure. The case is also very durable and will last a long time.
Tokens: This | Apple | laptop | case | has | a | nice | , | sturdy | design | and | fits | laptops | with | up | to | a | MacBook | Air | well | . | The | color | of | the | case | is | great | and | the | design | is | appealing | . | The | case | has | retractable | feet | which | help | to | keep | the | device | secure | . | The | case | is | also | very | durable | and | will | last | a | long | time | .
Lemmas: this | Apple | laptop | case | have | a | nice | , | sturdy | design | and | fit | laptop | with | up | to | a | MacBook | Air | well | . | the | color | of | the | case | be | great | and | the | design | be | appeal | . | the | case | have | retractable | foot | which | help | to | keep | the | device | secure | . | the | case | be | also | very | durable | and | will | last | a | long | time | .
Stop Words Removed: Apple | laptop | case | nice | , | sturdy | design | fits | laptops | MacBook | Air | . | color | case | great | design | appealing | . | case | retractable | feet | help | device | secure | . | case | durable | long | time | .
Stemmed Words: thi | appl | laptop | case | ha | a | nice | , | sturdi | design | and | fit | laptop | with | up | to | a | macbook | air | well | . | the | color | of | the | case | is | great | and | the | design | is | appeal | . | the | case | ha | retract | feet | which | help | to | keep | the | devic | secur | . | the | case | is | also | veri | durabl | and | will | last | a | long | time | .
===== End of Line =====

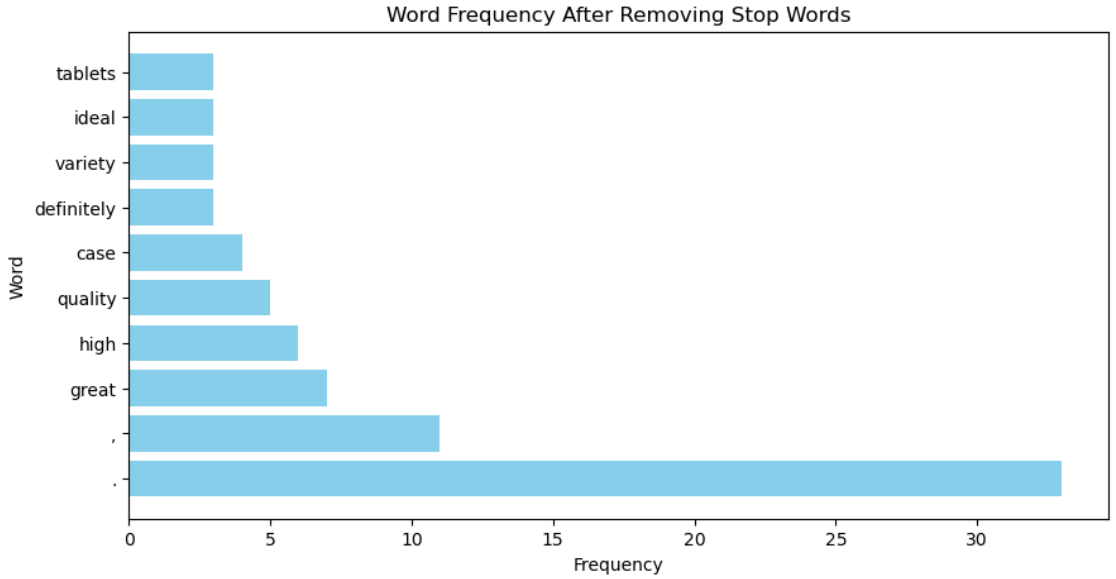
===== Line 9 =====
Pass Summary: These Reeboks are great for supporting a high arch and are lightweight and comfortable. They come in a variety of colors and sizes, and are ideal for walking or biking. They are also flexible and well made.
Tokens: These | Reeboks | are | great | for | supporting | a | high | arch | and | are | lightweight | and | comfortable | . | They | come | in | a | variety |

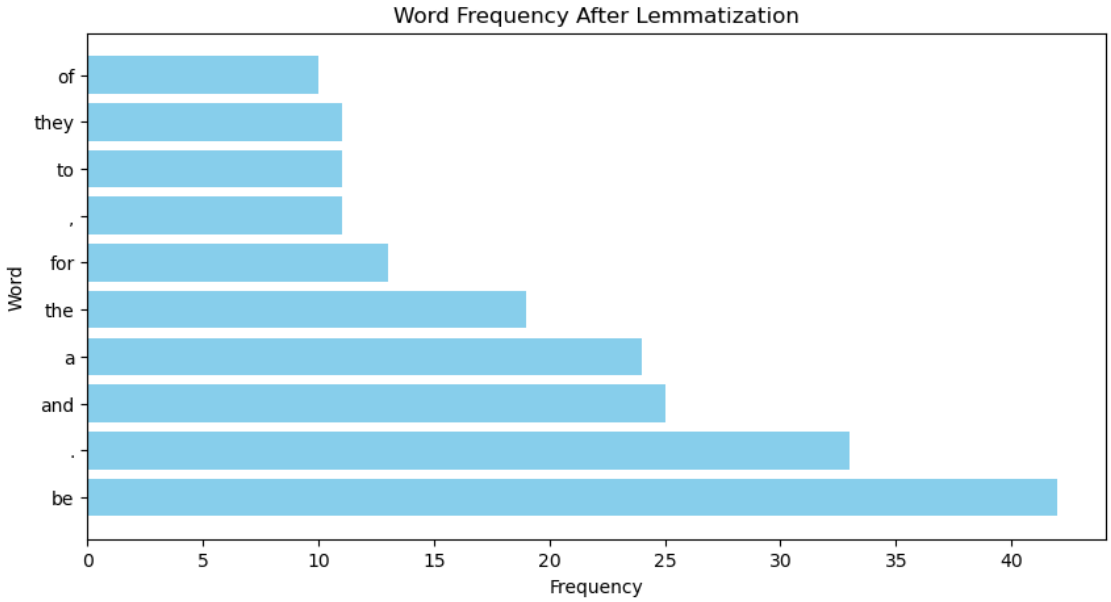
of | colors | and | sizes | , | and | are | ideal | for | walking | or | biking | . | They | are | also | flexible | and | well | made | .
Lemmas: these | Reeboks | be | great | for | support | a | high | arch | and | be | lightweight | and | comfortable | . | they | come | in | a | variety | of | color | and | size | , | and | be | ideal | for | walk | or | biking | . | they | be | also | flexible | and | well | make | .
Stop Words Removed: Reeboks | great | supporting | high | arch | lightweight | comfortable | . | come | variety | colors | sizes | , | ideal | walking | biking | . | flexible | .
Stemmed Words: these | reebok | are | great | for | support | a | high | arch | and | are | lightweight | and | comfort | . | they | come | in | a | varieti | of | color | and | size | , | and | are | ideal | for | walk | or | bike | . | they | are | also | flexibl | and | well | made | .
===== End of Line =====

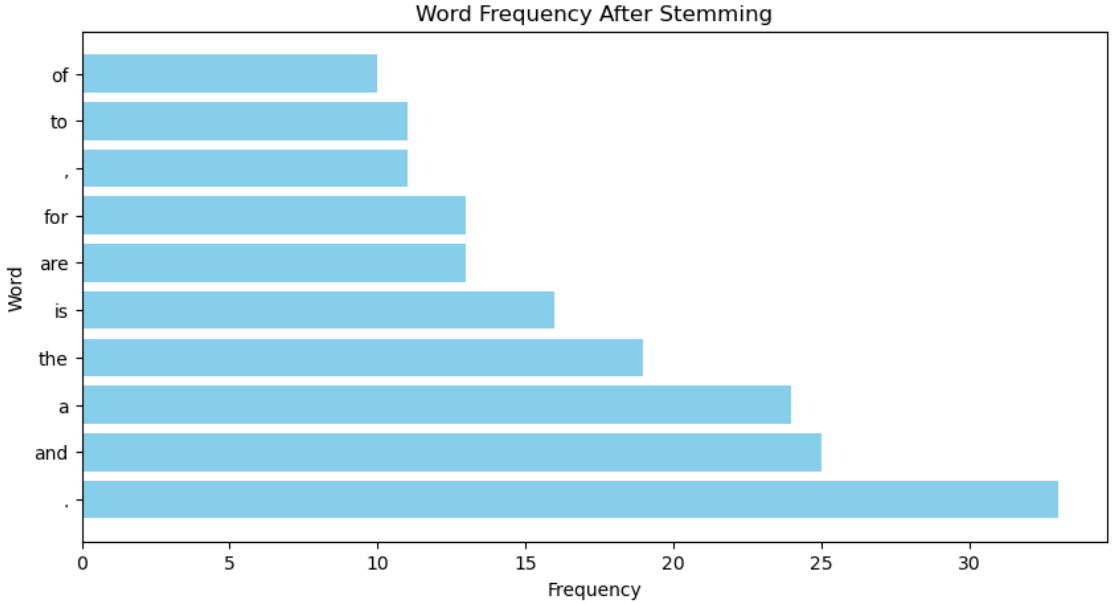
===== Line 10 =====
Pass Summary: These magnets are not strong enough to hold a lot of weight and aren't ideal for hanging a large number of items. They are however too small to be useful for holding onto a metal surface. The size of the hook is dependent on the object being hung, so gentle handling is required to ensure that the magnet stays in place.
Tokens: These | magnets | are | not | strong | enough | to | hold | a | lot | of | weight | and | are | n't | ideal | for | hanging | a | large | number | of | items | . | They | are | however | too | small | to | be | useful | for | holding | onto | a | metal | surface | . | The | size | of | the | hook | is | dependent | on | the | object | being | hung | , | so | gentle | handling | is | required | to | ensure | that | the | magnet | stays | in | place | .
Lemmas: these | magnet | be | not | strong | enough | to | hold | a | lot | of | weight | and | be | not | ideal | for | hang | a | large | number | of | item | . | they | be | however | too | small | to | be | useful | for | hold | onto | a | metal | surface | . | the | size | of | the | hook | be | dependent | on | the | object | be | hang | , | so | gentle | handling | be | require | to | ensure | that | the | magnet | stay | in | place | .
Stop Words Removed: magnets | strong | hold | lot | weight | ideal | hanging | large | number | items | . | small | useful | holding | metal | surface | . | size | hook | dependent | object | hung | , | gentle | handling | required | ensure | magnet | stays | place | .
Stemmed Words: these | magnet | are | not | strong | enough | to | hold | a | lot | of | weight | and | are | n't | ideal | for | hang | a | larg | number | of | item | . | they | are | howev | too | small | to | be | use | for | hold | onto | a | metal | surfac | . | the | size | of | the | hook | is | depend | on | the | object | be | hung | , | so | gentl | handl | is | requir | to | ensur | that | the | magnet | stay | in | place | .
===== End of Line =====

[23]:
```
def plot_word_frequencies(word_list, title):
    word_freq = Counter(word_list)
    common_words = word_freq.most_common(10)

    words = [word[0] for word in common_words]
    counts = [word[1] for word in common_words]

    plt.figure(figsize=(10,5))
    plt.barh(words, counts, color='skyblue')
    plt.xlabel('Frequency')
    plt.ylabel('Word')
    plt.title(title)
    plt.show()

# Initialize counters
original_text_counter = Counter()
stopwords_removed_counter = Counter()
lemmatized_counter = Counter()
stemmed_counter = Counter()

with open('amazon_product_pass_summary.jsonl', 'r') as f:
    for line in f:
        data = json.loads(line.strip())
        pass_summary = data.get('pass_summary', 'Key not found')
        doc = nlp(pass_summary)

        tokens = [token.text for token in doc]
        lemmas = [token.lemma_ for token in doc]
        stop_words_removed = [token.text for token in doc if not token.is_stop]
        stemmed_words = [stemmer.stem(token.text) for token in doc]

        original_text_counter.update(tokens)
        stopwords_removed_counter.update(stop_words_removed)
        lemmatized_counter.update(lemmas)
        stemmed_counter.update(stemmed_words)

# Plot the top 10 most common words for each counter
plot_word_frequencies(original_text_counter.elements(), 'Word Frequency Before␣
 ↪Preprocessing')
plot_word_frequencies(stopwords_removed_counter.elements(), 'Word Frequency␣
 ↪After Removing Stop Words')
plot_word_frequencies(lemmatized_counter.elements(), 'Word Frequency After␣
 ↪Lemmatization')
plot_word_frequencies(stemmed_counter.elements(), 'Word Frequency After␣
 ↪Stemming')
```


Word Frequency Before Preprocessing
of
The
to
,
for
are
is
a
and
.
Word
0
5
10
15
20
25
30
Frequency


Word Frequency After Removing Stop Words
tablets
ideal
variety
definitely
case
quality
high
great
,
.
Word
0
5
10
15
20
25
30
Frequency


Word Frequency After Lemmatization
of
they
to
,
for
the
a
and
.
be
Word
0
5
10
15
20
25
30
35
40
Frequency


Word Frequency After Stemming
of
to
,
for
are
is
the
a
and
.
Word
0
5
10
15
20
25
30
Frequency