In [1]:

```
import dask.dataframe as dd
from konlpy.tag import Mecab, Okt
from collections import Counter
from wordcloud import WordCloud
import matplotlib.pyplot as plt
import numpy as np
from PIL import Image
import re
import time
import pandas as pd # pandas import는 유지

# matplotlib 한글 폰트 설정 (워드클라우드 제목 등에 한글 표시 위함)
# 이 부분은 wordcloud 라이브러리의 font_path와 별개로, matplotlib의 설정입니다.
# 시스템에 설치된 한글 폰트 경로로 변경해주세요. (예: Malgun Gothic, AppleGothic,
plt.rcParams['font.family'] = 'Malgun Gothic' # Windows 예시
# plt.rcParams['font.family'] = 'AppleGothic' # Mac 예시
plt.rcParams['axes.unicode_minus'] = False # 마이너스 기호 깨짐 방지


# --- 설정 (이 부분을 사용자 환경에 맞게 수정하세요) ---
csv_file_path = "C:/Users/HomePc/Desktop/review.csv"
text_column_name = 'review_text'
sentiment_column_name = 'sentiment'
font_path = "C:/Windows/Fonts/NotoSansKR-VF.ttf" # WordCloud용 폰트 경로 (시스템
mask_image_path = "C:/Users/HomePc/Desktop/aaa.jpg"

# 불용어 리스트 (초기 테스트를 위해 일부만 남기거나 잠시 비워두세요)
# 만약 단어가 전혀 추출되지 않는다면, 이 리스트를 비워놓고 테스트해보세요.
stopwords = ['영화', '정말', '이런', '그냥','은', '는', '이', '가', '을', '를',
# test_stopwords = [] # 디버깅용: 불용어를 일시적으로 비워두고 테스트
# stopwords = test_stopwords # 불용어 테스트 시 이 줄 활성화


# --- 코드 시작 ---
start_time = time.time()
print("데이터 로딩 중...")

df = dd.read_csv(csv_file_path, encoding='utf-8', on_bad_lines='skip')

print(f"**Dask가 인식한 컬럼:** {df.columns.tolist()}")

print(f"**'{sentiment_column_name}' 컬럼의 고유 값 및 빈도:**")
sentiment_counts = df[sentiment_column_name].value_counts().compute()
print(sentiment_counts)
print("-" * 30)

try:
    tagger = Mecab()
    print("Mecab 형태소 분석기 사용.")
except Exception as e:
    print(f"Mecab 초기화 실패: {e}. Okt 형태소 분석기로 대체합니다.")
    tagger = Okt()

def preprocess_text_for_wc(text):
    if pd.isna(text):
        return []
    text = str(text)
    text = re.sub(r'[^가-힣\s]', '', text) # 한글과 공백만 남기고 제거
```

```
    # --- 디버깅을 위해 이 부분을 잠시 활성화하여 tagger.pos() 결과를 확인해보세요
    # print(f"원본 텍스트: {text[:50]}...") # 텍스트 앞부분 확인
    words_and_pos = tagger.pos(text)
    # print(f"형태소 분석 결과: {words_and_pos[:20]}") # 형태소 분석 결과 앞부분
    # ----------------------------------------------------------------------

    extracted_words = []
    for word, pos in words_and_pos:
        # 명사 (N), 동사 (V), 형용사 (A)를 포함
        # len(word) > 0 으로 변경하여 한 글자 단어도 포함
        if (pos.startswith('N') or pos.startswith('V') or pos.startswith('A')) a
            extracted_words.append(word)

    return extracted_words

print("긍정/부정 데이터 필터링 및 텍스트 전처리 중...")

# sentiment 컬럼의 값을 모두 소문자로 변환하여 비교
positive_reviews_processed = df[df[sentiment_column_name].str.lower() == 'positi
    preprocess_text_for_wc, meta=('words', 'object')
).compute()

positive_words_list = [item for sublist in positive_reviews_processed if sublist

negative_reviews_processed = df[df[sentiment_column_name].str.lower() == 'negati
    preprocess_text_for_wc, meta=('words', 'object')
).compute()

negative_words_list = [item for sublist in negative_reviews_processed if sublist

print("단어 빈도 계산 중...")
positive_word_counts = Counter(positive_words_list)
negative_word_counts = Counter(negative_words_list)

print(f"**긍정 단어 총 개수 (전처리 후):** {len(positive_words_list)}")
print(f"**부정 단어 총 개수 (전처리 후):** {len(negative_words_list)}")
print(f"**긍정 단어 종류 개수 (Counter):** {len(positive_word_counts)}")
print(f"**부정 단어 종류 개수 (Counter):** {len(negative_word_counts)}")

print(f"**긍정 단어 상위 10개:** {positive_word_counts.most_common(10)}")
print(f"**부정 단어 상위 10개:** {negative_word_counts.most_common(10)}")
print("-" * 30)

print("마스크 이미지 로드 중...")
try:
    mask_image = np.array(Image.open(mask_image_path))
    if mask_image.ndim == 3 and mask_image.shape[2] == 4:
        mask = mask_image[:, :, 3]
    else:
        mask = mask_image
except FileNotFoundError:
    print(f"오류: 마스크 이미지 파일 '{mask_image_path}'을 찾을 수 없습니다. 마스
    mask = None
except Exception as e:
    print(f"마스크 이미지 로드 중 오류 발생: {e}. 마스크 없이 워드클라우드를 생성
    mask = None

print("워드클라우드 생성 및 시각화 중...")
```

```
wordcloud_pos = None
wordcloud_neg = None

if positive_word_counts:
    wordcloud_pos = WordCloud(
        font_path=font_path,
        background_color='white',
        width=800,
        height=600,
        max_words=100,
        prefer_horizontal=0.9,
        min_font_size=10,
        mask=mask,
        colormap='Greens'
    ).generate_from_frequencies(positive_word_counts)
else:
    print("경고: 긍정 워드클라우드를 생성할 단어가 없습니다.")

if negative_word_counts:
    wordcloud_neg = WordCloud(
        font_path=font_path,
        background_color='white',
        width=800,
        height=600,
        max_words=100,
        prefer_horizontal=0.9,
        min_font_size=10,
        mask=mask,
        colormap='Reds'
    ).generate_from_frequencies(negative_word_counts)
else:
    print("경고: 부정 워드클라우드를 생성할 단어가 없습니다.")

fig, axes = plt.subplots(1, 2, figsize=(20, 10))

if wordcloud_pos:
    axes[0].imshow(wordcloud_pos, interpolation='bilinear')
    axes[0].set_title('긍정 리뷰 워드클라우드 (명사 + 동사)', fontsize=16)
else:
    axes[0].text(0.5, 0.5, '긍정 단어 없음', horizontalalignment='center', verti
    axes[0].set_title('긍정 리뷰 워드클라우드 (단어 없음)', fontsize=16, color='
axes[0].axis('off')

if wordcloud_neg:
    axes[1].imshow(wordcloud_neg, interpolation='bilinear')
    axes[1].set_title('부정 리뷰 워드클라우드 (명사 + 동사)', fontsize=16)
else:
    axes[1].text(0.5, 0.5, '부정 단어 없음', horizontalalignment='center', verti
    axes[1].set_title('부정 리뷰 워드클라우드 (단어 없음)', fontsize=16, color='
axes[1].axis('off')

plt.tight_layout()
plt.show()

end_time = time.time()
print(f"\n총 소요 시간: {end_time - start_time:.2f} 초")
print("워드클라우드 생성이 완료되었습니다!")
```

데이터 로딩 중...
Dask가 인식한 컬럼: ['review_id', 'reviewer', 'gender', 'review_date', 'year', 'score', 'sentiment', 'review_text']
'sentiment' 컬럼의 고유 값 및 빈도:
sentiment
Negative   399843
Neutral   300275
Positive   299882
Name: count, dtype: int64[pyarrow]

Mecab 초기화 실패: Install MeCab in order to use it: http://konlpy.org/en/latest/install/. Okt 형태소 분석기로 대체합니다.
긍정/부정 데이터 필터링 및 텍스트 전처리 중...
단어 빈도 계산 중...
긍정 단어 총 개수 (전처리 후): 3587112
부정 단어 총 개수 (전처리 후): 4908145
긍정 단어 종류 개수 (Counter): 42517
부정 단어 종류 개수 (Counter): 42795
긍정 단어 상위 10개: [('감동', 44838), ('최고', 29728), ('잘', 28090), ('좋다', 27765), ('재밌다', 25807), ('연기', 25236), ('생각', 24279), ('그저', 23269), ('훌륭하다', 23099), ('완벽하다', 22979)]
부정 단어 상위 10개: [('별로', 60622), ('실망', 53972), ('지루하다', 51672), ('스럽다', 48748), ('어색하다', 48456), ('산만하다', 48409), ('점', 35653), ('평점', 34879), ('스토리', 31186), ('왜', 29832)]

마스크 이미지 로드 중...
워드클라우드 생성 및 시각화 중...

총 소요 시간: 2202.10 초
워드클라우드 생성이 완료되었습니다!