In [28]:
```
import os
import glob
import pandas as pd

os.getcwd()

path_files = glob.glob('data/train/*.txt')
df = pd.DataFrame(columns=['date', 'text'])

for path_file in path_files:

    date, _ = os.path.splitext(os.path.basename(path_file))

    with open(path_file, 'r') as fr:
        raw_data = ''.join(fr.readlines())
        raw_data = raw_data.replace('□', '').replace('\n', '')

    new_row = pd.DataFrame({'date': [date], 'text':[raw_data]})
    df = pd.concat([df, new_row])


df = df.reset_index(drop=True)
df.to_csv('preprocess/pre_input_1.csv')

df = pd.read_csv('preprocess/pre_input_1.csv', index_col=0)
df
```

Out[28]:

	date	text
0	201703	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.25%)...
1	201706	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.25%)...
2	201709	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.25%)...
3	201712	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를 현수준(1.25%)...
4	201803	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.50%)...
5	201806	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.50%)...
6	201809	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.50%)...
7	201812	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.50%)...
8	201903	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.75%)...
9	201906	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.75%)...
10	201909	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현재의 1.75%에서...
11	201912	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현재의 1.50%에서...
12	202003	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.25%)...
13	202006	(붙임)통화정책방향 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를...
14	202009	(붙임)통화정책방향 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를...
15	202012	(붙임)통화정책방향 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를...
16	202103	(붙임)통화정책방향 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를...
17	202106	(붙임)통화정책방향 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를...
18	202109	(붙임)통화정책방향 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를...
19	202112	(붙임)통화정책방향 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를...

In [35]:
```
import os
import glob
import pandas as pd

df = pd.read_csv('preprocess/pre_input_6.csv')

train_df = df.dropna(subset=['text','Y'])
test_df = df[df['text'].isna()]
```

In [36]:
```
train_df
```

Out[36]:

	회사명	회계년도	분기	profit	text	Y
1	DB손해보험(주)	2018	1	210697471	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.50%)...	1.0
2	DB손해보험(주)	2019	1	232419677	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현재의 1.50%에서...	1.0
3	DB손해보험(주)	2020	1	226549867	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.25%)...	0.0
4	DB손해보험(주)	2021	1	221215499	(붙임)통화정책방향 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를...	0.0
5	DB손해보험(주)	2017	2	192880168	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를 현수준(1.25%)...	0.0
...	...	...	...	...	...	...
195	흥국화재해상보험(주)	2017	4	71589000	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를 현수준(1.25%)...	0.0
196	흥국화재해상보험(주)	2018	4	73599000	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현 수준(1.50%)...	1.0
197	흥국화재해상보험(주)	2019	4	73006000	금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를현재의 1.50%에서...	0.0
198	흥국화재해상보험(주)	2020	4	69680000	(붙임)통화정책방향 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를...	0.0
199	흥국화재해상보험(주)	2021	4	77094000	(붙임)통화정책방향 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를...	1.0

190 rows × 6 columns

In [4]:
```
test_df
```

Out[4]:

	회사명	회계년도	분기	profit	text
3	DB손해보험(주)	2020	1	226549867	NaN
4	DB손해보험(주)	2021	1	221215499	NaN
8	동양생명보험(주)	2020	1	204702889	NaN
9	동양생명보험(주)	2021	1	201575178	NaN
13	롯데손해보험(주)	2020	1	39024137	NaN
...	...	...	...	...	...
189	한화손해보험(주)	2021	4	91170364	NaN
193	현대해상화재보험(주)	2020	4	222947000	NaN
194	현대해상화재보험(주)	2021	4	232555905	NaN
198	흥국화재해상보험(주)	2020	4	69680000	NaN
199	흥국화재해상보험(주)	2021	4	77094000	NaN

80 rows × 5 columns

In [37]:
```
from konlpy.tag import Okt
from konlpy.tag import Kkma
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectori

tokenizer = Okt()
# kkma = Kkma()

tokenized_sentences = []

for _, row in train_df.iterrows():
    sentence = row['text']
    tokens = tokenizer.pos(sentence)
    tokenized_sentence = ' '.join([word+'/'+pos for word, pos in tokens])
    tokenized_sentences.append(tokenized_sentence)

# vectorizer = TfidfVectorizer()
vectorizer = CountVectorizer()

tfidf_matrix = vectorizer.fit_transform(tokenized_sentences)
```

In [38]:
```
display(
    vectorizer.get_feature_names_out(),
    tfidf_matrix.toarray(),
    pd.DataFrame(tfidf_matrix.toarray(), columns=list(vectorizer.get_feat
)
```

```
array(['00', '10년', '10월', '11월', '17일', '19', '1년', '1일', '2013', '202
0',
       '2020년', '25', '2월', '3년', '3월', '4분', '50', '5월', '75', '7월',
       '8월', 'aa', 'adjective', 'adverb', 'alpha', 'bp', 'determiner',
       'exclamation', 'foreign', 'gdp', 'it', 'josa', 'mbso', 'modifier',
       'noun', 'number', 'punctuation', 'rp', 'suffix', 'verb', '가가',
       '가격', '가견', '가계', '가계부채', '가능성', '가운데', '각각', '각국', '감소',
'갔다',
       '갔으나', '갔으며', '강세', '강화', '개발', '개선', '개시', '개지', '거리', '거
시경제',
       '건설', '겠지만', '격변', '견실', '견조', '결정', '경감', '경계', '경기', '경
로', '경제',
       '경제지표', '경제활동', '계속', '고용', '고조', '공개시장', '공공', '공급', '공
사', '공업',
       '과의', '과정', '관계', '관련', '교역', '국고', '국내', '국내외', '국제', '국
채', '군사',
       '규모', '그간', '그치는', '근원', '근접한', '근접할', '글로벌', '금년', '금률',
'금리',
       '금번', '금융', '금융기관', '금융시장', '금융채', '금융통화위원회', '급등', '급
락', '기간',
       '기관', '기대', '기록', '기별', '기상', '기업', '기여', '기적', '기조', '기
준', '기준금리',
       '기중', '긴장', '깊게', '까지', '까지는', '꾸준한', '나가되', '나갈', '나타났으
나',
       '나타내겠지만', '나타내고', '나타내는', '나타내다가', '나타내었다', '나타내었으나',
'나타내었으며',
       '나타낸', '나타낼', '나타냈다', '낮아져', '낮아졌다', '낮아질', '낮은', '내년',
'내외',
       '농산물', '농업', '농축', '높고', '높아져', '높아졌다', '높아졌으며', '높아지고',
'높아지면서',
       '높아질', '높은', '높은만큼', '누적', '누증', '늘어나는', '다만', '다소', '다
시', '다음',
       '달러', '달러화', '담보', '당분간', '당초', '대로', '대를', '대상', '대응', '대
의', '대중',
       '대책', '대체로', '대출', '대한', '더딘', '더딜', '도기', '도모', '돌다가', '돌
아서는',
       '되겠으나', '되고', '되나', '되는', '되며', '되면서', '되므로', '되었다', '되었
다가', '되었던',
       '되었으나', '되었으며', '된다', '됨에', '두기', '둔화', '들어', '등락', '따라',
'따른',
       '또한', '리스크', '마이너스', '망치', '매매', '매우', '머무를', '머무를것으로',
'머물',
       '머물다', '머물렀으며', '면밀히', '모두', '모습', '목표', '무역', '물가', '물가
상승률', '물은',
       '물을', '물이', '미국', '미약', '미치는', '미치면서', '민간', '바이러스', '반도
체', '반등',
       '반락', '반면', '반영', '반의', '받아', '받으며', '받을', '받을것으로', '발생',
'발표',
       '발행', '방안', '방역', '방향', '백신', '범위', '벗어나는', '벗어나지', '변동
성', '변수',
       '변화', '보급', '보다', '보였다', '보였으나', '보였으며', '보이', '보이다가',
'보이면서',
       '보인다', '보합', '보호무역', '부담', '부문', '부양책', '부진', '부진하였다',
'부진한',
       '부진할', '부터', '부합', '분쟁', '불구', '불균형', '불안', '불확실', '붙임',
'비교',
       '비도', '비해', '빠르게', '사정', '사회', '산업', '살펴보면서', '살펴볼', '살펴
볼것이다',
       '상당', '상반기', '상승', '상승세', '상존', '상향', '상황', '상회', '서비스',
'서비스업',
```

```
        '서울', '선진국', '선호', '설비', '성도', '성은', '성장', '성향', '세계', '세
는', '세도',
        '세자', '소멸', '소비', '소비자', '속도', '수가', '수도', '수도권', '수산', '수
산물',
        '수요', '수의', '수준', '수출', '수출입', '승률', '시계', '시기', '시장', '시키
는',
        '시행일', '식료품', '신속한', '신용', '신용등급', '신중히', '신흥', '신흥시장',
'실물', '실시',
        '실적', '심리', '심화', '쏠림', '아울러', '악화', '안정', '안정화', '않는', '않
을', '압력',
        '약세', '약화', '양호', '어가', '어간', '어려움', '에너지', '에는', '에도', '에
서',
        '에서는', '에서의', '에전망한', '여건', '여력', '여부', '여전히', '연간', '연
내', '영업',
        '영향', '예년', '예상', '오름세', '오름세가', '오름세를', '완만', '완화', '우
려', '운영',
        '운용', '움직임', '원연', '원활', '위축', '위해', '위험', '위험관리', '유가',
'유동성',
        '유로', '유류', '유의', '유인', '유지', '으로', '으로가', '으로는', '으로도',
'으로써',
        '으로의', '은행', '의성', '의한', '이다', '이번', '이어', '이어가다가', '이어졌
다',
        '이어졌으나', '이어졌으며', '이어지겠지만', '이어지고', '이어지고있으나', '이어지
는', '이어지면서',
        '이어진', '이외', '이자', '이하', '이후', '인플레이션', '인플레이션율', '인하',
'일반인',
        '일부', '일시', '일자리', '입어', '있고', '있는', '있다', '있다고', '있도록',
'있으나',
        '있으므로', '자금', '자금지원', '자기', '자산', '자수', '잔존', '잠재', '장기',
'장기시',
        '재개', '재원', '재정', '재정정책', '적극', '적요', '적용', '적절히', '전개',
'전기',
        '전기요금', '전망', '전망치', '전반', '전세계', '전월', '전환', '점검', '점진',
'점차',
        '접종', '정도', '정부', '정부보증채', '정상화', '정책', '정치', '제약', '제외',
'제품',
        '제한', '조건', '조달', '조정', '조치', '좁은', '종료', '주가', '주요', '주
의', '주춤',
        '주택', '줄어들었으나', '줄여', '중개', '중동', '중반', '중소기업', '중순', '중
심', '중앙은행',
        '중후', '증가', '증거', '증권', '증대', '지난', '지난해', '지방', '지속', '지
수', '지역',
        '지원', '지정학적', '지출', '지켜보면서', '진전', '집행', '차등', '차례', '차
입', '차질',
        '채권', '채널', '체결', '초반', '최근', '최대', '추가', '추경', '추구', '추
진', '축소',
        '충격', '취약', '취업', '측면', '침체', '커지는', '커지면서', '커질', '코로나',
'크게',
        '크지', '테이퍼링', '통화', '통화스왑', '통화정책', '투자', '투자가', '특히',
'파급', '판단',
        '포함', '프로그램', '피해', '필요', '필요성', '하게', '하겠으나', '하고', '하고
있으나',
        '하기로', '하는', '하다', '하다가', '하락', '하면서', '하방', '하여', '하였고',
'하였다',
        '하였다가', '하였으나', '하였으며', '하향', '한국', '한국은행', '한기대', '한시',
'한편',
        '할것으로', '함께', '해졌으며', '했던', '향후', '현재', '현행', '협상', '호조',
'확대',
        '확산', '확충', '환매', '환율', '회복', '회사', '회피', '회하', '효과', '후
```

```
반', '휴직',
       '흐름'], dtype=object)
array([[0, 0, 2, ..., 2, 0, 9],
       [0, 0, 0, ..., 3, 0, 4],
       [0, 0, 0, ..., 6, 0, 0],
       ...,
       [0, 0, 0, ..., 3, 0, 4],
       [0, 0, 0, ..., 1, 0, 5],
       [1, 0, 0, ..., 3, 0, 6]])
```

	00	10 년	10 월	11 월	17 일	19	1 년	1 일	2013	2020	...	환 매	환 율	회 복	회 사	회 피	회 하	효 과	후 반	휴 직
0	0	0	2	1	0	0	0	0	0	0	...	0	2	2	0	0	1	0	2	0
1	0	0	0	0	0	0	0	0	0	0	...	0	2	2	0	1	1	1	3	0
2	0	0	0	3	1	10	1	1	0	1	...	2	3	3	1	0	0	3	6	0
3	0	0	0	3	0	8	0	0	0	0	...	0	2	15	0	0	0	4	1	0
4	0	0	1	0	0	0	0	0	0	0	...	0	2	10	0	0	0	2	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	..
185	0	0	1	0	0	0	0	0	0	0	...	0	2	10	0	0	0	2	2	0
186	0	0	2	1	0	0	0	0	0	0	...	0	2	2	0	0	1	0	2	0
187	0	0	0	0	0	0	0	0	0	0	...	0	2	2	0	1	1	1	3	0
188	0	0	0	0	0	10	0	0	1	0	...	0	2	12	0	1	0	4	1	0
189	1	0	0	0	0	7	0	0	0	0	...	0	2	10	0	0	0	1	3	0

In [39]:
```
import matplotlib as mpl
import matplotlib.pyplot as plt

import matplotlib.font_manager as fm
from wordcloud import WordCloud

fontpath = '/usr/share/fonts/truetype/nanum/NanumBarunGothic.ttf'
font = fm.FontProperties(fname=fontpath, size=9)
plt.rc('font', family='NanumBarunGothic')
# mpl.font_manager._rebuild()
```

In [52]:
```
text = ' '.join(train_df['text'].tolist())

# Generate word cloud
wordcloud = WordCloud(width=800, height=400, font_path=fontpath).generate

# Display the generated word cloud using matplotlib
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```

In [55]:
```
pd.DataFrame(tfidf_matrix.toarray(), columns=list(vectorizer.get_feature_
```

Out[55]:
```
noun           107380
josa            39450
verb            22540
punctuation     13330
으로               6080
modifier         4490
number           4300
adjective        2930
통화정책             2340
성장               2180
에서               2180
경제               2100
수준               1940
하였다              1770
지속               1640
상승               1530
suffix          1520
증가               1400
물가               1310
금융시장             1300
dtype: int64
```

In [42]:
```
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn.metrics import r2_score

X = tfidf_matrix.toarray()
y = train_df['Y']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,

model = LinearRegression()

model = LogisticRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
```

```
mse = mean_squared_error(y_test, y_pred)
print("MSE:", mse)

mae = mean_absolute_error(y_test, y_pred)
print("MAE:", mae)

r2 = r2_score(y_test, y_pred)
print("R-Squared:", r2)
```

```
MSE: 0.15789473684210525
MAE: 0.15789473684210525
R-Squared: 0.361344537815126
```