Online Data Analyzer 사용 방법

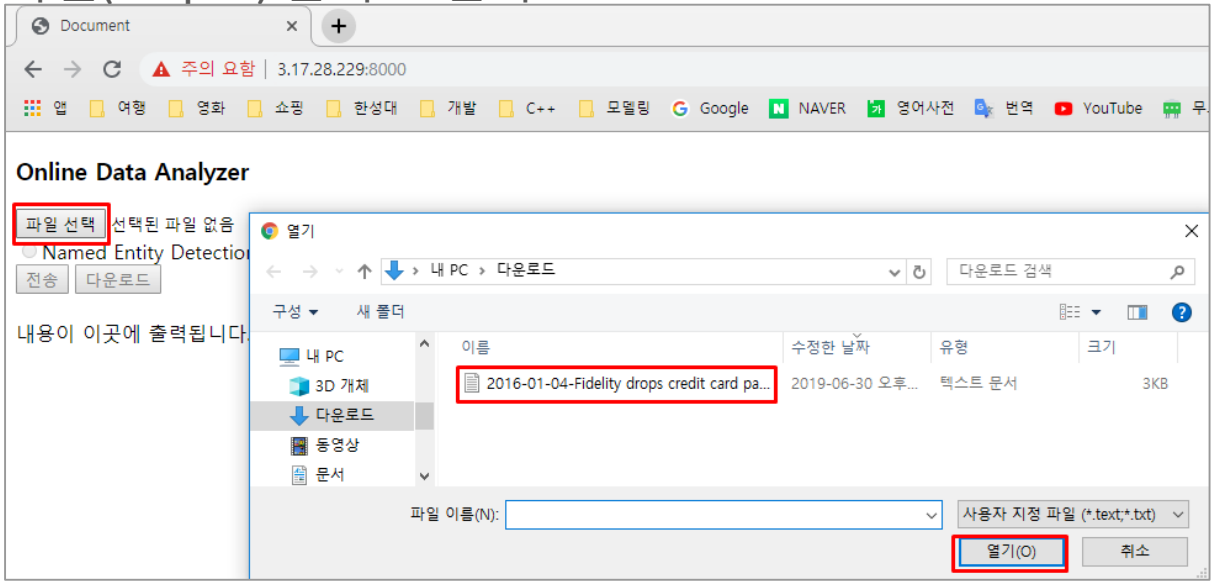
Jin Won,

School of Computer Engineering,

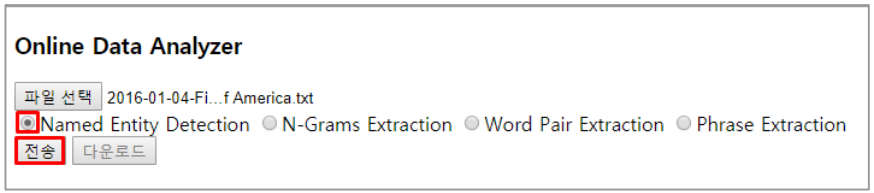
Hansung University

Contents

- 웹 페이지 접속
- 파일(corpus) 선택 -> 열기



Named Entity Detection




```python
from application.nlp import comprehend
from nltk import sent_tokenize

def run(article):
    # 기사를 라인 별로 분할하여 리스트로 저장한다.
    sentences = sent_tokenize(article)

    # Map 객체로써 key는 entity 명, value는 타입이다.
    # ex. key: NEW YORK, value: LOCATION
    dictMap = {}

    # 결과를 저장하는 string이다.
    retVal = ''

    # sentences 내에 있는 모든 sentence에 대해
    for sentence in sentences:
        # entity를 조사한다.
        entities = comprehend.detect_entities(Text=sentence, LanguageCode='en')

        # entities 내에 있는 모든 entitiy에 대해
        for entity in entities["Entities"]:
            # 이름과 type을 저장한다.
            key = entity["Text"]
            value = entity["Type"]

            # dictMap에 해당 이름의 entity가 존재하지 않는 경우에 한해서
            # entity를 추가하고, retVal에 반영한다.
            if key not in dictMap:
                dictMap[key] = value
                row = (key + ': ' + value + '\r\n')
                retVal += row

    return retVal
```

Named Entity Detection

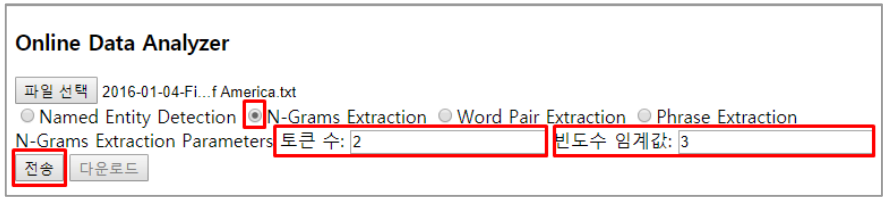
- input

BOSTON (Reuters) - Fidelity Investments said on Monday it is dropping long-time credit card partners American Express Co and Bank of America Corp, ending a 12-year partnership that has generated billions of dollars in fees.
Boston-based Fidelity, which has 24 million customers, said its new partners will be U.S.
Bancorp and Visa Inc, effective Monday.
The exclusive alliance will provide Visa branded credit-card products to U.S. consumers, including Fidelity customers.
The switch is another setback for American Express, already reeling from its lost deal with warehouse club retailer Costco Wholesale Corp.
AmEx said earlier this year the loss of the Costco contract would hurt profit for two years.
AmEx shares are off 25 percent over the past year and were down 2.9 percent Monday morning.
Many other major U.S. financial stocks were also lower, with shares of Bank of America, Visa Inc and U.S.
Bancorp all down by more than 2 percent.
Ram Subramaniam, president of Fidelity's retail brokerage business, did not give any specific reason for ending the partnership with American Express and Bank of America.
"It's been a long, good partnership," he said.
A spokeswoman for American Express said the Fidelity portfolio accounts for less than 1 percent of billings.
A Bank of America spokeswoman said the agreement not to continue the relationship with American Express was a mutual decision between the two companies.
"Over the past several years, Bank of America has been exiting from our financial institutions card business where Bank of America has limited opportunity to deepen customer relationships, and this move is consistent with that strategy," she wrote via email.
Since 2003, Fidelity has offered 2 percent cash back credit cards with American Express and Bank of America's FIA Card Services. During that time Fidelity customers have earned $1.1 billion cash rewards.
The new alliance will feature cards with chip security technology, with access to digital wallets that include Apple Pay, Samsung Pay and Android pay.
The new card program will issue the Fidelity Rewards Visa Signature Card and the Fidelity Investments 529 College Rewards Visa Signature Card, where card members can earn unlimited 2 percent cash back with no annual fees, caps or categories when directing rewards into eligible Fidelity accounts.
U.S.
Bank also has agreed to acquire Fidelity's existing co-brand credit card portfolio with about $1.7 billion in associated balances.
Additional reporting by Richa Naidu in Bangalore and Dan Freed in New York; Editing by Alan Crosby and David Gregorio

output

BOSTON: LOCATION
Reuters: ORGANIZATION
Fidelity Investments: ORGANIZATION
Monday: DATE
American Express Co: ORGANIZATION
Bank of America Corp: ORGANIZATION
12-year: QUANTITY
billions of dollars: QUANTITY
Boston: LOCATION
Fidelity: ORGANIZATION
24 million customers: QUANTITY
U.S.
Bancorp: ORGANIZATION
Visa Inc: ORGANIZATION
Visa: ORGANIZATION
U.S.: LOCATION
American Express: ORGANIZATION
Costco Wholesale Corp.: ORGANIZATION
AmEx: ORGANIZATION
earlier this year: DATE
Costco: ORGANIZATION
two years: QUANTITY
25 percent: QUANTITY
past year: DATE
2.9 percent: QUANTITY
Monday morning: DATE
Bank of America: ORGANIZATION
more than 2 percent: QUANTITY
Ram Subramaniam: PERSON
less than 1 percent: QUANTITY
two companies: QUANTITY
past several years: DATE
2003: DATE

N-Grams Extraction




```python
from nltk import *

# line 단위의 sentence를 입력받아 빈도 분포 객체를 반환한다.
# @param sentence: 빈도 분포를 분석할 문장
# @param numTokens: 빈도 분포를 생성할 기준 토큰 수
def getFreqDist(sentence, numTokens):
    tokenList = word_tokenize(sentence)
    ngramList = ngrams(tokenList, numTokens)

    return FreqDist(ngramList)

def run(article, numTokens, freqThreshold):
    # 텍스트 파일을 라인 별로 분할하여 리스트로 저장
    sentences = article.splitlines()

    # sentences의 모든 빈도 분포를 분석한 결과가 저장된다.
    # 이 객체는 Map 객체이며, key는 복합어, value는 빈도 수이다.
    globalFreqDist = {}

    # sentences 내에 있는 모든 sentence에 대하여
    for sentence in sentences:
        # 현재 sentence에 대한 빈도 분포 분석
        localDist = getFreqDist(sentence, numTokens)

        # localDist에 존재하는 빈도 분포를 조사
        # key: 복합어 / value: 빈도 수
        for key, value in localDist.items():

            # 만약 globalFreqDist에 이미 존재하는 복합어라면
            if key in globalFreqDist:
                # 해당 복합어의 빈도 수에 value를 누적
                globalFreqDist[key] = globalFreqDist[key] + value
            # 처음 나타난 복합어라면
            else:
                # globalFreqDist 맵에 새로 추가
                globalFreqDist[key] = value

    retVal = ''
```

N-Grams Extraction

- input

BOSTON (Reuters) - Fidelity Investments said on Monday it is dropping long-time credit card partners American Express Co and Bank of America Corp, ending a 12-year partnership that has generated billions of dollars in fees.
Boston-based Fidelity, which has 24 million customers, said its new partners will be U.S.
Bancorp and Visa Inc, effective Monday.
The exclusive alliance will provide Visa branded credit-card products to U.S. consumers, including Fidelity customers.
The switch is another setback for American Express, already reeling from its lost deal with warehouse club retailer Costco Wholesale Corp.
AmEx said earlier this year the loss of the Costco contract would hurt profit for two years.
AmEx shares are off 25 percent over the past year and were down 2.9 percent Monday morning.
Many other major U.S. financial stocks were also lower, with shares of Bank of America, Visa Inc and U.S.
Bancorp all down by more than 2 percent.
Ram Subramaniam, president of Fidelity's retail brokerage business, did not give any specific reason for ending the partnership with American Express and Bank of America.
"It's been a long, good partnership," he said.
A spokeswoman for American Express said the Fidelity portfolio accounts for less than 1 percent of billings.
A Bank of America spokeswoman said the agreement not to continue the relationship with American Express was a mutual decision between the two companies.
"Over the past several years, Bank of America has been exiting from our financial institutions card business where Bank of America has limited opportunity to deepen customer relationships, and this move is consistent with that strategy," she wrote via email.
Since 2003, Fidelity has offered 2 percent cash back credit cards with American Express and Bank of America's FIA Card Services.
During that time Fidelity customers have earned $1.1 billion cash rewards.
The new alliance will feature cards with chip security technology, with access to digital wallets that include Apple Pay, Samsung Pay and Android pay.
The new card program will issue the Fidelity Rewards Visa Signature Card and the Fidelity Investments 529 College Rewards Visa Signature Card, where card members can earn unlimited 2 percent cash back with no annual fees, caps or categories when directing rewards into eligible Fidelity accounts.
U.S.
Bank also has agreed to acquire Fidelity's existing co-brand credit card portfolio with about $1.7 billion in associated balances.
Additional reporting by Richa Naidu in Bangalore and Dan Freed in New York; Editing by Alan Crosby and David Gregorio

output

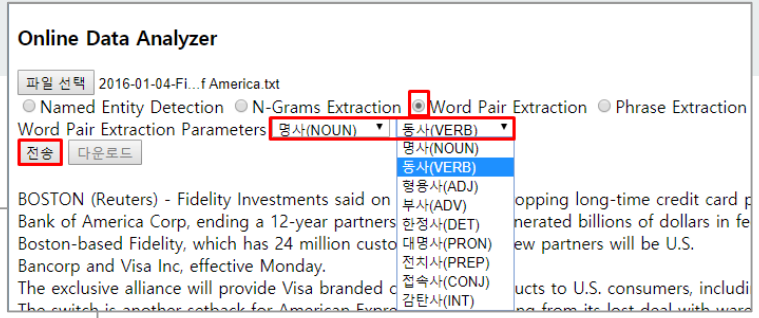
('American', 'Express'), 6
('and', 'Bank'), 3
('Bank', 'of'), 7
('of', 'America'), 7
('U.S', '.'), 3
('2', 'percent'), 3
('with', 'American'), 3
('the', 'Fidelity'), 3

Word Pair Extraction

- code

```python
from nltk import sent_tokenize
from nltk import RegexpParser
from nltk import word_tokenize
from nltk import pos_tag
from nltk import tree
from re import search

def run(article, pos1, pos2):
    # 텍스트 파일을 문장 단위로 분할하여 sentenceList로 저장
    sentenceList = sent_tokenize(article)

    # 재정의 품사 패턴
    noun = 'FW|NN'      # 명사
    verb = 'MD|VB'      # 동사
    adj = 'JJ|VBN'      # 형용사
    adv = 'W?RB|^RP'    # 부사
    det = 'DT?'         # 한정사
    pron = 'EX|PRP|^WP' # 대명사
    prep = '^TO|^IN'    # 전치사
    conj = 'CC'         # 접속사
    int_ = 'UH'          # 감탄사

    # word pair 패턴 정의
    patternList = ('pattern1: {<%s><%s>}' % (pos1, pos2))

    # 결과를 저장하는 string이다.
    retVal = ''

    # 파서 생성
    parser = RegexpParser(patternList)
```

```python
    # 리스트 내 모든 문장에 대하여
    for sentence in sentenceList:

        # 공백을 기준으로 tokenizing
        tokenList = word_tokenize(sentence)

        # 리스트 내 모든 토큰들을 품사별로 매핑
        taggedList = pos_tag(tokenList)

        # (토큰, 재정의된 품사) 튜플을 저장할 리스트
        retaggedList = []

        # POS 재정의
        for tagged in taggedList:

            if search(noun, tagged[1]):
                retaggedList.append((tagged[0], 'NOUN'))
            elif search(verb, tagged[1]):
                retaggedList.append((tagged[0], 'VERB'))
            elif search(adj, tagged[1]):
                retaggedList.append((tagged[0], 'ADJ'))
            elif search(adv, tagged[1]):
                retaggedList.append((tagged[0], 'ADV'))
            elif search(det, tagged[1]):
                retaggedList.append((tagged[0], 'DET'))
            elif search(pron, tagged[1]):
                retaggedList.append((tagged[0], 'PRON'))
            elif search(prep, tagged[1]):
                retaggedList.append((tagged[0], 'PREP'))
            elif search(conj, tagged[1]):
                retaggedList.append((tagged[0], 'CONJ'))
            elif search(int_, tagged[1]):
                retaggedList.append((tagged[0], 'INT'))
            else:
                retaggedList.append((tagged[0], tagged[1]))
```




```python
        # 파싱
        parsedTree = parser.parse(retaggedList)

        # 파일에 demo1 출력
        for subtree in parsedTree:
            if isinstance(subtree, tree.Tree):
                retVal += (subtree[0][0] + ' [' + subtree[0][1] + '] ')
                retVal += (subtree[1][0] + ' [' + subtree[1][1] + ']\n')

    return retVal
```

Word Pair Extraction

- input

BOSTON (Reuters) - Fidelity Investments said on Monday it is dropping long-time credit card partners American Express Co and Bank of America Corp, ending a 12-year partnership that has generated billions of dollars in fees.
Boston-based Fidelity, which has 24 million customers, said its new partners will be U.S.
Bancorp and Visa Inc, effective Monday.
The exclusive alliance will provide Visa branded credit-card products to U.S. consumers, including Fidelity customers.
The switch is another setback for American Express, already reeling from its lost deal with warehouse club retailer Costco Wholesale Corp.
AmEx said earlier this year the loss of the Costco contract would hurt profit for two years.
AmEx shares are off 25 percent over the past year and were down 2.9 percent Monday morning.
Many other major U.S. financial stocks were also lower, with shares of Bank of America, Visa Inc and U.S.
Bancorp all down by more than 2 percent.
Ram Subramaniam, president of Fidelity's retail brokerage business, did not give any specific reason for ending the partnership with American Express and Bank of America.
"It's been a long, good partnership," he said.
A spokeswoman for American Express said the Fidelity portfolio accounts for less than 1 percent of billings.
A Bank of America spokeswoman said the agreement not to continue the relationship with American Express was a mutual decision between the two companies.
"Over the past several years, Bank of America has been exiting from our financial institutions card business where Bank of America has limited opportunity to deepen customer relationships, and this move is consistent with that strategy," she wrote via email.
Since 2003, Fidelity has offered 2 percent cash back credit cards with American Express and Bank of America's FIA Card Services. During that time Fidelity customers have earned $1.1 billion cash rewards.
The new alliance will feature cards with chip security technology, with access to digital wallets that include Apple Pay, Samsung Pay and Android pay.
The new card program will issue the Fidelity Rewards Visa Signature Card and the Fidelity Investments 529 College Rewards Visa Signature Card, where card members can earn unlimited 2 percent cash back with no annual fees, caps or categories when directing rewards into eligible Fidelity accounts.
U.S.
Bank also has agreed to acquire Fidelity's existing co-brand credit card portfolio with about $1.7 billion in associated balances.
Additional reporting by Richa Naidu in Bangalore and Dan Freed in New York; Editing by Alan Crosby and David Gregorio

output

Investments [NOUN] said [VERB]
partners [NOUN] will [VERB]
alliance [NOUN] will [VERB]
Visa [NOUN] branded [VERB]
switch [NOUN] is [VERB]
AmEx [NOUN] said [VERB]
contract [NOUN] would [VERB]
shares [NOUN] are [VERB]
stocks [NOUN] were [VERB]
Express [NOUN] said [VERB]
spokeswoman [NOUN] said [VERB]
Express [NOUN] was [VERB]
America [NOUN] has [VERB]
America [NOUN] has [VERB]
move [NOUN] is [VERB]
Fidelity [NOUN] has [VERB]
customers [NOUN] have [VERB]
alliance [NOUN] will [VERB]
program [NOUN] will [VERB]
members [NOUN] can [VERB]

Phrase Extraction

*현재는 명사구만 추출되며
chunk 정규식을 어떻게 입력 받을지에 따라
개선할 수 있습니다.

```python
from nltk import word_tokenize
from nltk import pos_tag
from nltk import RegexpParser
from nltk import tree

def run(article):
    # 컨텐츠 내에 있는 모든 단어들을 공백 문자열 기준으로 Tokenizing
    rawTokens = word_tokenize(article)

    # Part of Speech(POS) tagging
    # rawTokens(list)내에 저장된 모든 Token들을 POS 별로 매핑
    taggedTokens = pos_tag(rawTokens)

    # Chunk 정규식 정의
    # Chunk 정규식은 절대적인 기준이 없으며, 개발자가 직접 원문의 내용을 분석하고
    # 상황에 따라 별도로 정의해주어야 한다.
    regexes = """
        NP: {<DT|PP\$>?<JJ>*<NN.*>+} # noun phrase
        PP: {<IN><NP>}               # prepositional phrase
        VP: {<MD>?<VB.*><NP|PP>}     # verb phrase
    """

    # 파서 생성
    parser = RegexpParser(regexes)

    # 파싱
    parsedTree = parser.parse(taggedTokens)

    # 결과를 저장하는 string이다.
    retVal = ''

    for subtree in parsedTree:
        if isinstance(subtree, tree.Tree) and subtree.label() == "NP":
            retVal += (str(subtree) + '\r\n')

    return retVal
```

Phrase Extraction

- input

BOSTON (Reuters) - Fidelity Investments said on Monday it is dropping long-time credit card partners American Express Co and Bank of America Corp, ending a 12-year partnership that has generated billions of dollars in fees.
Boston-based Fidelity, which has 24 million customers, said its new partners will be U.S.
Bancorp and Visa Inc, effective Monday.
The exclusive alliance will provide Visa branded credit-card products to U.S. consumers, including Fidelity customers.
The switch is another setback for American Express, already reeling from its lost deal with warehouse club retailer Costco Wholesale Corp.
AmEx said earlier this year the loss of the Costco contract would hurt profit for two years.
AmEx shares are off 25 percent over the past year and were down 2.9 percent Monday morning.
Many other major U.S. financial stocks were also lower, with shares of Bank of America, Visa Inc and U.S.
Bancorp all down by more than 2 percent.
Ram Subramaniam, president of Fidelity's retail brokerage business, did not give any specific reason for ending the partnership with American Express and Bank of America.
"It's been a long, good partnership," he said.
A spokeswoman for American Express said the Fidelity portfolio accounts for less than 1 percent of billings.
A Bank of America spokeswoman said the agreement not to continue the relationship with American Express was a mutual decision between the two companies.
"Over the past several years, Bank of America has been exiting from our financial institutions card business where Bank of America has limited opportunity to deepen customer relationships, and this move is consistent with that strategy," she wrote via email.
Since 2003, Fidelity has offered 2 percent cash back credit cards with American Express and Bank of America's FIA Card Services. During that time Fidelity customers have earned $1.1 billion cash rewards.
The new alliance will feature cards with chip security technology, with access to digital wallets that include Apple Pay, Samsung Pay and Android pay.
The new card program will issue the Fidelity Rewards Visa Signature Card and the Fidelity Investments 529 College Rewards Visa Signature Card, where card members can earn unlimited 2 percent cash back with no annual fees, caps or categories when directing rewards into eligible Fidelity accounts.
U.S.
Bank also has agreed to acquire Fidelity's existing co-brand credit card portfolio with about $1.7 billion in associated balances.
Additional reporting by Richa Naidu in Bangalore and Dan Freed in New York; Editing by Alan Crosby and David Gregorio

output

(NP BOSTON/NNP)
(NP Reuters/NNP)
(NP Fidelity/NN Investments/NNPS)
(NP Bank/NNP)
(NP Boston-based/JJ Fidelity/NNP)
(NP customers/NNS)
(NP new/JJ partners/NNS)
(NP Visa/NNP Inc/NNP)
(NP effective/JJ Monday/NNP)
(NP The/DT exclusive/JJ alliance/NN)
(NP U.S./NNP consumers/NNS)
(NP The/DT switch/NN)
(NP this/DT year/NN)
(NP the/DT loss/NN)
(NP years/NNS)
(NP AmEx/NNP shares/NNS)
(NP percent/NN)
(NP percent/NN Monday/NNP morning/NN)
(NP other/JJ major/JJ U.S./NNP)
(NP financial/JJ stocks/NNS)
(NP Visa/NNP Inc/NNP)
(NP U.S./NNP Bancorp/NNP)
(NP percent/NN)
(NP Ram/NNP Subramaniam/NNP)
(NP president/NN)
(NP retail/JJ brokerage/NN business/NN)
(NP Bank/NNP)
(NP good/JJ partnership/NN)