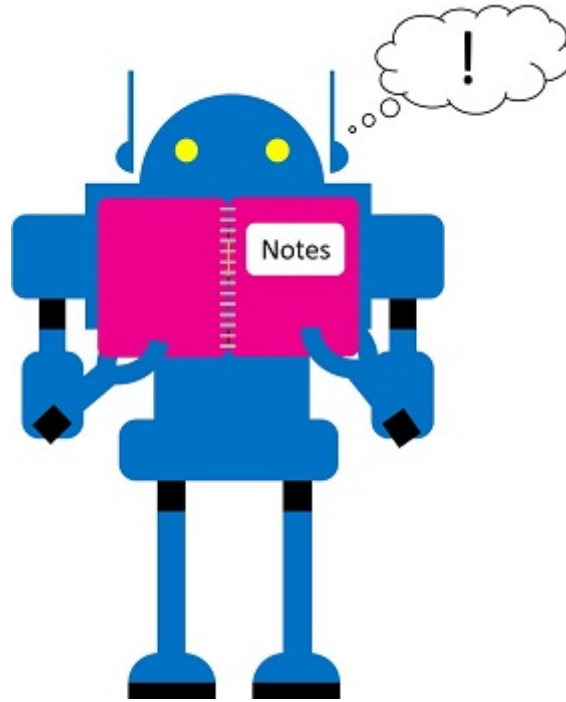


Text Analytics

자연어 처리(Natural Language Processing, NLP)는 문어와 구어를 다루는 인공지능(AI)의 한 분야이다. NLP를 사용하여 텍스트나 음성에서 의미적 의미를 추출하거나 자연 언어로 의미 있는 응답을 공식화하는 솔루션을 구축할 수 있다.

Microsoft Azure *Cognitive* 서비스에는 *Text Analytics* 서비스가 포함되어 있다. 이 서비스는 텍스트의 주요 구문(Key Phrases)을 식별하고 감정에 따라 텍스트를 분류하는 등 기본적으로 제공되는 NLP 기능을 포함한다.



예를 들어 가상의 *Margie's Travel* 조직이 고객에게 호텔 숙박에 대한 리뷰를 제출하도록 요청했다고 가정해보겠다. *Text Analytics* 서비스를 사용하여 주요 구문을 추출하여 리뷰를 요약하고, 어떤 리뷰가 긍정적이고 어떤 리뷰가 부정적인지 확인하거나, 또는 위치나 사람과 같은 알려진 엔티티를 추출하는 등 리뷰 텍스트를 분석할 수 있다.

리뷰 문서 보기

먼저 고객이 남긴 호텔 리뷰를 살펴보자.

리뷰는 텍스트 파일이다. 그 내용들을 보기 위해서는 아래의 셀 왼편에 있는 **Run cell** (▶) 버튼을 클릭하면 된다.

In []:

```
import os

# /data/reviews 폴더에 있는 리뷰들을 읽는다.
reviews_folder = os.path.join('data', 'text', 'reviews')

# ID(파일 이름)과 텍스트 (내용)속성들로 구성된 리뷰들의 컬렉션을 생성한다.
reviews = []
for file_name in os.listdir(reviews_folder):
    review_text = open(os.path.join(reviews_folder, file_name)).read()
    review = {"id": file_name, "text": review_text}
    reviews.append(review)

for review_num in range(len(reviews)):
    # 리뷰 텍스트를 인쇄한다
    print('{ }\n{ }\n'.format(reviews[review_num]['id'], reviews[review_num]['text']))
```

Cognitive 서비스 리소스 만들기

이러한 리뷰의 텍스트를 분석하려면 **Text Analytics** 인지 서비스를 사용하면 된다. 이를 사용하려면 Azure 구독에 **Text Analytics*** 또는 **Cognitive Services***** 리소스를 프로비저닝해야 한다.(이 서비스가 사용하려는 유일한 서비스이거나 사용을 별도로 추적하려는 경우 Text Analytics 리소스를 사용하면 된다. 그렇지 않으면 Cognitive Services 리소스를 사용하여 Text Analytics 서비스를 다른 Cognitive 서비스와 결합할 수 있다. 이렇게 하면 개발자는 하나의 엔드포인트와 키를 사용하여 액세스할 수 있다.)

이미 만들어진 Cognitive 서비스가 없다면 다음과 같은 순서로 Azure 구독에서 **Cognitive Services** 리소스를 생성할 수 있다.

1. 브라우저의 새로운 탭을 열고, Azure portal(<https://portal.azure.com>)을 (<https://portal.azure.com>)을 입력하고, Microsoft계정으로 로그인한다..
2. + 리소스 만들기 버튼을 클릭하고, **Cognitive Services** 서비스를 찾은 다음, **Cognitive Services** 리소스를 다음과 같은 내용으로 생성한다.
 - 이름: 유일한 이름을 입력한다(가능하면 영문과 숫자사용).
 - 구독: Azure 구독선택.
 - 위치: 가능한 위치(한국 중부 추천):
 - 가격책정계층: 표준 S0
 - 리소스 그룹: 원하는 유리한 이름(가능하면 영문과 숫자사용).
3. 배포가 완료될 때까지 기다린다. 그런 다음 Cognitive Services 리소스로 이동하여 *개요 페이지에서 링크를 클릭하여 서비스 키를 관리한다. 클라이언트 응용 프로그램에서 Cognitive Services 리소스에 연결하려면 엔드포인트와 키가 필요하다.

Cognitive Services 리소스에 있는 키와 엔드포인트 가져오기

Cognitive Services 리소스를 사용하기 위해서는, 클라이언트 응용프로그램에서는 엔드포인트와 인증 키가 필요하다.

1. Azure portal에서, Cognitive Services 리소스를 선택하고 **키 및 엔드포인트** 페이지를 선택한 다음 **키1** 을 복사하여 아래의 **YOUR_COG_KEY**.를 붙여 넣는다.
2. 리소스에 있는 **엔드포인트** 를 복사해서 아래의 **YOUR_COG_ENDPOINT**.에 붙여 넣는다.
3. 셀을 선택한 다음 셀 왼쪽에있는 **셀 실행(▶)** 버튼을 클릭하여 아래 코드를 실행한다.

In []:

```
cog_key = 'YOUR_COG_KEY'
cog_endpoint = 'YOUR_COG_ENDPOINT'

print('Ready to use cognitive services at {} using key {}'.format(cog_endpoint, cog_key))
```

만들어진 Cognitive 서비스 리소스에서 Text Analytics 서비스를 사용하기 위해서는 Azure Cognitive Services Text Analytics SDK를 설치해야 한다.

In []:

```
! pip install azure-cognitiveservices-language-textanalytics
```

언어 감지하기

리뷰가 쓰여진 언어를 확인하는 작업부터 시작해보자.

In []:

```
import os
from azure.cognitiveservices.language.textanalytics import TextAnalyticsClient
from msrest.authentication import CognitiveServicesCredentials

# Text Analytics Cognitive 서비스 리소스를 위한 클라이언트 가져오기
text_analytics_client = TextAnalyticsClient(endpoint=cog_endpoint,
                                           credentials=CognitiveServicesCredentials(cog_key))

# 앞에서 언급한 /data/reviews 폴더로부터 읽은 리뷰들을 분석한다
language_analysis = text_analytics_client.detect_language(documents=reviews)

# 각 리뷰에서 감지된 언어 세부 정보들을 프린트한다.
for review_num in range(len(reviews)):
    # print the review id
    print(reviews[review_num]['id'])

    # 이 리뷰에 대한 언어 세부 정보를 가져온다.
    lang = language_analysis.documents[review_num].detected_languages[0]
    print(' - Language: {} \n - Code: {} \n - Score: {} \n'.format(lang.name, lang.iso6391_name,
                                                                lang.score))

# 리뷰 컬렉션에 감지된 언어 코드를 추가한다(이렇게 함으로 더 많은 분석이 가능하다.
reviews[review_num]['language'] = lang.iso6391_name
```

핵심 구(Key Phrases) 추출하기

이제 고객 리뷰의 텍스트를 분석하여 주요 논점을 나타내는 핵심 구를 식별할 수 있다.

In []:

```
# 이전 코드 셀에서 생성한 클라이언트와 리뷰들을 이용하여 핵심구를 추출한다
key_phrase_analysis = text_analytics_client.key_phrases(documents=reviews)

# 각 리뷰별로 핵심 구를 인쇄한다
for review_num in range(len(reviews)):
    # 리뷰 ID를 인쇄한다.
    print(reviews[review_num]['id'])

    # 이 리뷰에 있는 핵심 구를 가져온다.
    print('\nKey Phrases:')
    key_phrases = key_phrase_analysis.documents[review_num].key_phrases
    # 각 핵심 구를 인쇄한다.
    for key_phrase in key_phrases:
        print('Wt ', key_phrase)
    print('\n')
```

핵심 구는 리뷰 단위로 가장 중요한 논점을 이해하는 데 도움이 될 수 있다. 예를 들어, "도움이 되는 직원 (helpful staff)" 또는 "서비스 불량(poor service)"이라는 문구가 포함된 리뷰는 주요 관심사를 나타낼 수 있다.

감정분석(정서분석)

감정 점수를 기준으로 리뷰를 긍정적 또는 부정적으로 분류하는 것이 유용할 수 있다. 이것도 Text Analytics 서비스를 사용하여 수행할 수 있다.

In []:

```
# 이전 코드 셀에서 생성한 클라이언트와 리뷰들을 이용하여 감정 점수를 가져온다
sentiment_analysis = text_analytics_client.sentiment(documents=reviews)

# 리뷰들에 대한 결과를 프린트한다.
for review_num in range(len(reviews)):

    # 이 리뷰에 대한 감정점수를 가져온다.
    sentiment_score = sentiment_analysis.documents[review_num].score

    # 0.5보가 크면 '긍정적'으로 분류한다.
    if sentiment_score < 0.5:
        sentiment = 'negative'
    else:
        sentiment = 'positive'

    # 파일 이름과 감정을 프린트한다.
    print('{} : {} ({}).format(reviews[review_num]['id'], sentiment, sentiment_score))
```

알려진 엔터티 추출하기

Entities 는 일반적으로 널리 이해되는 유형의 항목들을 의미하며 텍스트안에 언급되어 있는 것들의 목록이다. 예를 들면 장소, 사람 또는 날짜 등이다. 리뷰에서 언급되어 있는 날짜와 장소에 관심이 있다고 가정하면 다음 코드를 실행하여 그런 것들을 찾을 수 있다.

In []:

```
# 이전 코드 셀에서 생성한 클라이언트와 리뷰들을 이용하여 정한 엔티티들을 가져온다
entity_analysis = text_analytics_client.entities(documents=reviews)

# 리뷰들에 대한 결과들을 프린트한다
for review_num in range(len(reviews)):
    print(reviews[review_num]['id'])
    # 이 리뷰에 포함되어 있는 엔티티들을 가져온다
    entities = entity_analysis.documents[review_num].entities
    for entity in entities:
        # 장소 날짜와 엔티티들을 가져온다
        if entity.type in ['DateTime', 'Location']:
            link = '(' + entity.wikipedia_url + ')' if entity.wikipedia_id is not None else ''
            print(' - {}: {} {}'.format(entity.type, entity.name, link))
```

일부 엔티티는 연결된 Wikipedia 페이지가 있을 정도로 잘 알려져 있으며, 이 경우 Text Analytics 서비스는 해당 페이지의 URL을 반환한다.

심화학습

Text Analytics 서비스에 대한 더 많은 정보는 다음을 참조하기 바란다. [Text Analytics service documentation](https://docs.microsoft.com/azure/cognitive-services/text-analytics/) (<https://docs.microsoft.com/azure/cognitive-services/text-analytics/>).