

## Gathering Data

- 여러가지 프로젝트에 사용될 수 있는 데이터는 온라인에서 구할 수 있다
  - o 이 챕터는 Bear Detector 모델을 만드는 프로젝트를 다룬다
  - o 이것 말고 다른 형태의 image recognition application 도 가능함
- 이 책이 만들어질 당시 Bing Image Search 가 데이터 수집에 용이한 플랫폼이었음
  - o Microsoft Azure Cognitive Service 기반의 API 사용

- Bing Image Search 사용법

## 1. 마이크로소프트 어카운트를 만들어서 key 를 받는다.

아래 스크립트를 주피터의 cell 에 넣는다

```
key = 'XXX'
```

또는 커멘드 터미널에 아래를 지정한 후 주피터 노트북을 실행한다

```
export AZURE_SEARCH_KEY=your_key_here
```

그리고 아래 내용을 cell 에 넣어서 환경 설정을 한다

```
key = os.environ['AZURE_SEARCH_KEY']
```

## 2. 기본적인 사진을 다운로드 하는방법 (search\_images\_bing 함수 사용)

검색한 내용의 URL 을 받는 방법

```
In [ ]: search_images_bing
```

```
Out[ ]: <function utils.search_images_bing(key, term, min_sz=128)>
```

```
In [ ]: results = search_images_bing(key, 'grizzly bear')
ims = results.attrgot('content_url')
len(ims)
```

```
Out[ ]: 150
```

검색된 URL 을 이용하여 사진을 지정 폴더에 다운로드 하는방법

```
In [ ]: dest = 'images/grizzly.jpg'
download_url(ims[0], dest)
```

```
In [ ]: im = Image.open(dest)
im.to_thumb(128,128)
```

```
Out[ ]:
```



여러가지 검색어를 사용하여 관련 폴더를 만들고 해당 파일들을 저장하는 법

```
In [ ]: bear_types = 'grizzly','black','teddy'
path = Path('bears')
```

```
In [ ]: if not path.exists():
    path.mkdir()
    for o in bear_types:
        dest = (path/o)
        dest.mkdir(exist_ok=True)
        results = search_images_bing(key, f'{o} bear')
        download_images(dest, urls=results.attrgot('content_url'))
```

Our folder has image files, as we'd expect:

```
In [ ]: fns = get_image_files(path)
fns
```

```
Out[ ]: (#421) [Path('bears/black/00000095.jpg'),Path('bears/black/00000133.jpg'),Path('bears/black/00000062.
g'),Path('bears/black/00000029.jpg'),Path('bears/black/00000094.jpg'),Path('bears/black/00000124.jpg',
Path('bears/black/00000046.jpg'),Path('bears/black/00000045.jpg')...]
```

다운로드한 파일이 corrupt 됐는지 체크하는 법

```
In [ ]: failed = verify_images(fns)
failed
```

```
Out[ ]: (#0) []
```

그리고 unlink 함수를 사용해서 corrupt 된 파일들을 지운다

```
In [ ]: failed.map(Path.unlink);
```

- 주피터 노트북에서 함수의 사용방법을 잘 모를 때  
??verify\_images

- 아래와 같이 해당 함수의 소스코드와 코드의 저장위치를 보여준다

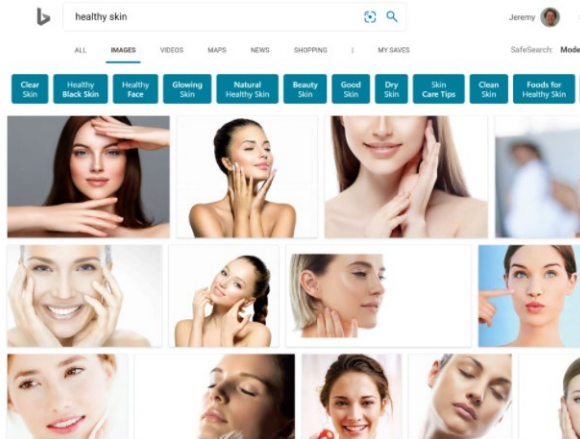
Signature: verify\_images(fns)

Source:

```
def verify_images(fns):
    "Find images in `fns` that can't be opened"
    return L(fns[i] for i,o in
              enumerate(parallel(verify_image, fns)) if not o)
File:      ~/git/fastai/fastai/vision/utils.py
Type:      function
```

- 이러한 서치 엔진을 이용한 데이터 수집에서는 biased 된 데이터를 수집할 수 있는 점을 유의하자

- 건강한 피부를 검색했을 때의 예제



- 데이터가 준비되면, 모델을 학습시키기 위한 포맷으로 변형시킬 것

- Fastai 에서는 DataLoaders 라는 object 를 형성