

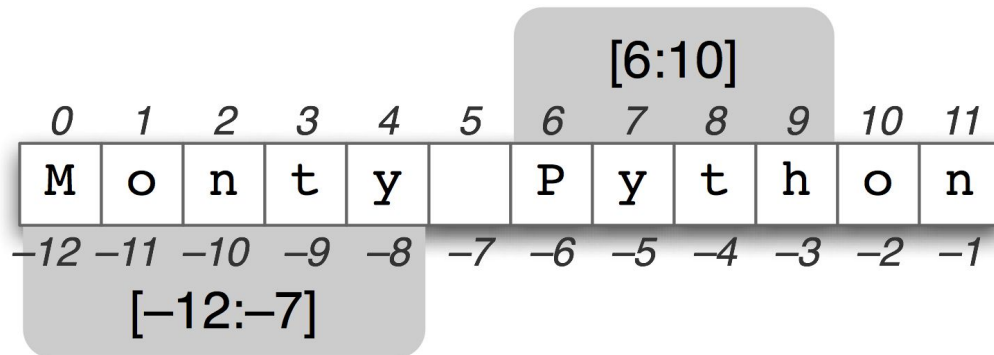
Python for Text Representation

성균관대학교
정윤경

파이썬 복습

Strings

```
>>> name = 'Monty Python'
>>> name[0]
'M'
>>> name[1:4]
'ont'
>>> name[-2]
'o'
>>> ' '.join(['Monty', 'Python'])
'Monty Python'
>>> 'Monty Python'.split()
['Monty', 'Python']
>>> 'onty' in name
>>> True
```



List

A variable can refer to a 250,000-word book

```
>>> my_sent = ['Bravely', 'bold', 'Sir', 'Robin'] + [',', 'rode', 'forth', 'from', 'Camelot', '.']
>>> words = my_sent[1:4]                # ['bold', 'Sir', 'Robin']
>>> words.sort()                        # ['Robin', 'Sir', 'bold']
>>> words.append('bold')                 # ['Robin', 'Sir', 'bold', 'bold']
>>> words.index('Sir')                   # 1
>>> words[-1]                           # 'bold'
>>> vocab = set(words)                   # {'Robin', 'Sir', 'bold'}
>>> vocab_size = len(vocab)               # 3
>>> 'Sir' in words
True
```

Exercise

```
>>> saying = ['After', 'all', 'is', 'said', 'and', 'done', 'all']
```

```
>>> tokens = set(saying)
```

```
>>> tokens = sorted(tokens)
```

```
>>> tokens[2:5]
```

what output do you expect here?

Dictionary

- **key**와 **value** 값으로 이루어진 자료구조로 원하는 값을 쉽게 찾을 수 있음
- **key**는 중복될 수 없음

```
>>> tel = {'jack': 4098, 'sape': 4139}
```

```
>>> tel['guido'] = 4127
```

```
>>> tel
```

```
{'guido': 4127, 'jack': 4098, 'sape': 4139}
```

```
>>> tel['jack']
```

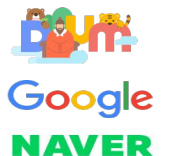
```
4098
```

```
>>> tel['jack'] = 1234
```

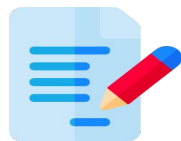
```
>>> del tel['sape']
```

Text as a List of Words

한눈에 보는 자연어 처리 과정


웹사이트

웹 크롤링



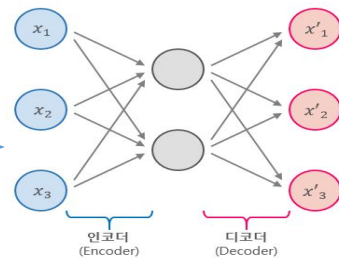
텍스트

문서 전처리

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots \\ a_{2,1} & a_{2,2} & a_{2,3} & \dots \\ a_{3,1} & a_{3,2} & a_{3,3} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

벡터

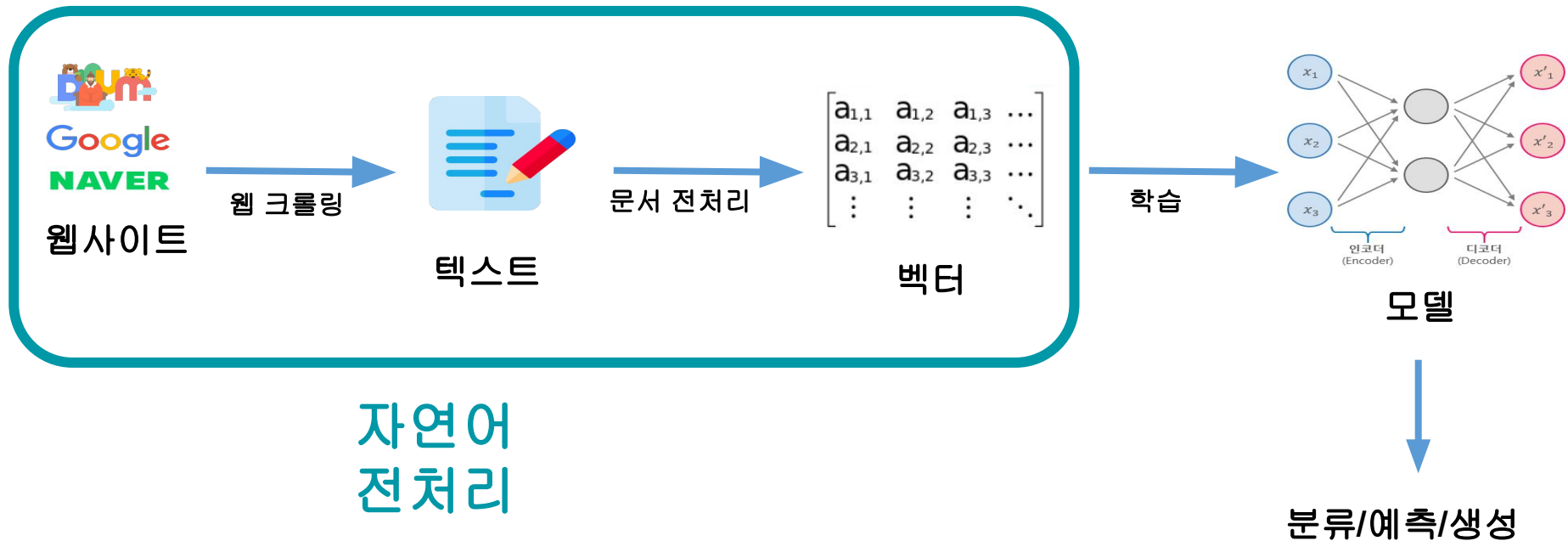
학습



모델

↓
분류/예측/생성

한눈에 보는 자연어 처리 과정



텍스트 표현

- `text = ""내 인생이 비극인줄 알았는데, 코미디였어.
고담시의 광대 아서 플렉은 코미디언을 꿈꾸는 남자.""`
- `text = ['내 인생이 비극인줄 알았는데, 코미디였어.', '고담시의
광대 아서 플렉은 코미디언을 꿈꾸는 남자.']`
- `text = ['내', '인생이', '비극인줄', '알았는데', ',', '코미디였어', '.',
'고담시의', '광대', '아서', '플렉은', '코미디언을', '꿈꾸는', '남자',
'']`

텍스트 표현

- `text = ""내 인생이 비극인줄 알았는데, 코미디였어.
고담시의 광대 아서 플렉은 코미디언을 꿈꾸는 남자.""`
- `text = ['내 인생이 비극인줄 알았는데, 코미디였어.', '고담시의
광대 아서 플렉은 코미디언을 꿈꾸는 남자.']`
- `text = ['내', '인생이', '비극인줄', '알았는데', ',', '코미디였어', '.',
'고담시의', '광대', '아서', '플렉은', '코미디언을', '꿈꾸는', '남자',
'']`

문자열

문장의
리스트

단어의
리스트

텍스트 표현

- `text = ""내 인생이 비극인줄 알았는데, 코미디였어.
고담시의 광대 아서 플렉은 코미디언을 꿈꾸는 남자.""`

문자열

- `text = ['내 인생이 비극인줄 알았는데, 코미디였어.', '고담시의
광대 아서 플렉은 코미디언을 꿈꾸는 남자.']`

문장의
리스트

- `text = ['내', '인생이', '비극인줄', '알았는데', ',', '코미디였어', '.',
'고담시의', '광대', '아서', '플렉은', '코미디언을', '꿈꾸는', '남자',
'']`

단어의
리스트

제일 유용한 방식

텍스트를 단어 리스트로 변환하려면?

워너브러더스에서는 DCEU(DC 확장 유니버스)와 상관없는 조커 영화가 하나 더 제작된다. 조커 역으로는 호아킨 피닉스가 발탁되었으며, 이 외에도 로버트 드니로, 재지 비츠 등이 캐스팅 되어 현재 촬영이 진행 중이다.

앞으로 조커를 연기할 모든 배우가 언제나 히스 레저를 뛰어넘고 싶다는 압박에 시달리겠지만, [수어사이드 스쿼드]에 출연했던 자레드 레토는 그 정도가 심한 편이었다. 그는 카메라가 돌아가지 않을 때도 조커처럼 행동했으며, 각종 기행으로 같이 출연하는 배우들을 괴롭히기까지 했다. 그러나 자레드 레토의 별난 노력과는 상관없이 [수어사이드 스쿼드]는 혹평에 시달렸으며, 조커 캐릭터 역시 모든 관객을 만족시키지 못했다. 오히려 너무 ‘과잉’된 캐릭터 해석이 불편했다는 의견이 많았다.

과연 호아킨 피닉스는 조커를 어떻게 해석할지 이목이 쏠리는 가운데, 지난 9월 21일 워너 브러더스가 조커의 카메라 테스트 영상을 공개했다. 조커가 되기 전, 아서 플렉 (Arthur Fleck)이라는 평범한 남자와 조커 이미지가 겹치는 형식의 독특한 영상이었다.

출처: <https://movie.v.daum.net/v/gmNkyLP4Zm>

표준 파이썬의 `split()` 함수 사용

```
text = "" 워너브러더스에서는 DCEU(DC 확장 유니버스)와 상관없는 조커  
영화가 하나 더 제작된다. 내 인생이 비극인줄 알았는데, 코미디였어.""  
print(text.split())
```

```
[' 워너브러더스에서는', ' DCEU (DC', ' 확장', ' 유니버스) 와',  
' 상관없는', ' 조커', ' 영화가', ' 하나', ' 더', ' 제작된다.',  
' 내', ' 인생이', ' 비극인줄', ' 알았는데,', ' 코미디였어.']
```

N-gram

- contiguous sequence of n items from a given sample of text or speech
- **text:** 워너브러더스에서는 DCEU(DC 확장 유니버스)와 상관없는 조커 영화가 하나 더 제작된다.

- **Unigram ($n=1$):**

['워너브러더스에서는', 'DCEU(DC', '확장', '유니버스)와', '상관없는', '조커', '영화가', '하나', '더', '제작된다.']

- **Bigram ($n=2$):**

['워너브러더스에서는 DCEU(DC', 'DCEU(DC 확장', '확장 유니버스)와', '유니버스)와 상관없는', '상관없는 조커', '조커 영화가', '영화가 하나', '하나 더', '더 제작된다.']

- **Trigram ($n=3$):**

['워너브러더스에서는 DCEU(DC 확장', 'DCEU(DC 확장 유니버스)와', '확장 유니버스)와 상관없는', '유니버스)와 상관없는 조커', '상관없는 조커 영화가', '조커 영화가 하나', '영화가 하나 더', '하나 더 제작된다.']