3조

한글 (뉴스) 자동으로 분류 프로그램 구현

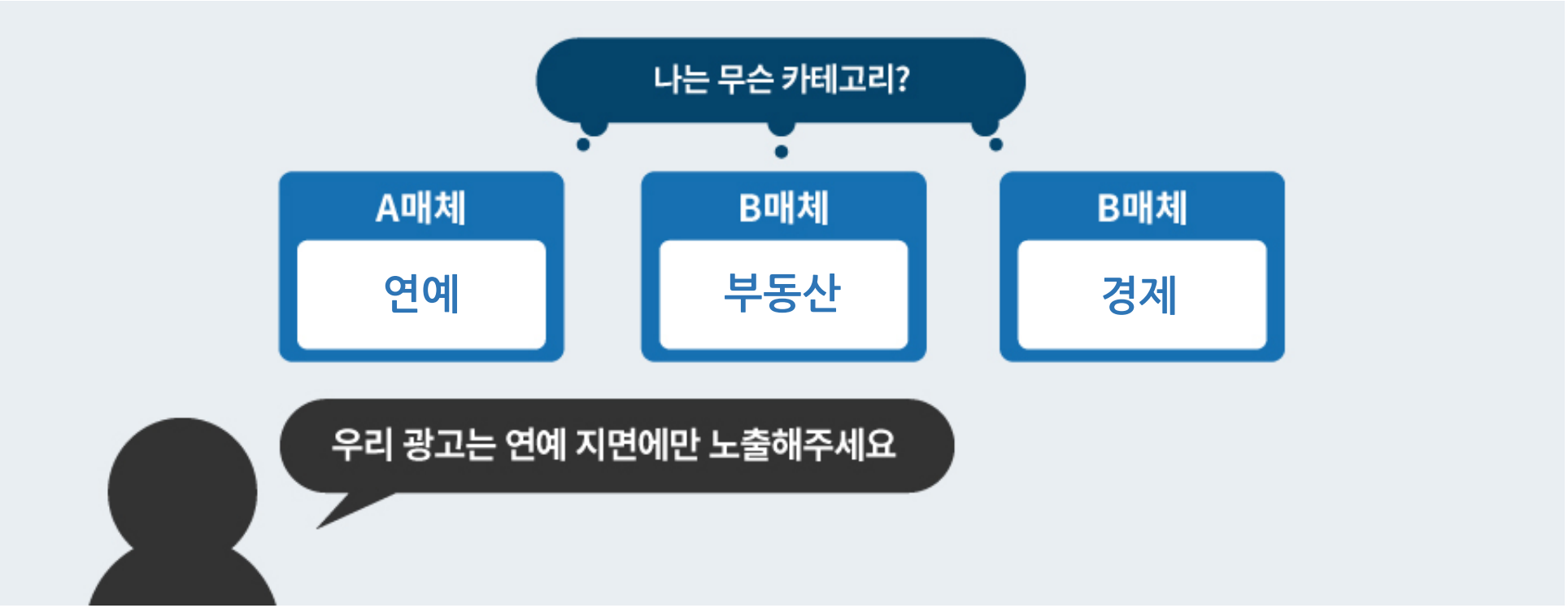
임채명, 한대건, 이정서

Team Members

3

CONTENTS

프로젝트 목적



목적

뉴스 카테고리 분류해주는 프로그램 구현하여
새로운 뉴스가 들어왔을 때 가장 밀접한 관련이 있는 카테고리를 알려준다

활용 데이터

네이버 부동산

"분양 기다리자"…'전세 버티기'에 서울 전셋값 더 오르나
서울 아파트 전세가격이 오름세를 보이고 있는 가운데 내주 초 발표될 민간
전셋값 상방 압력으로 작용할 것이란 업계의 전망이 나오고 있다.9일 국토
2019.08.09

주택거래량 절반 '뚝'…금리인하에 관망 수요 움직일까?
서울 주택 거래량이 급격하게 줄어드는 가운데 최근 금리인하로 인해 거래
일 서울부동산정보광장에 따르면 서울 상반기 아파트 매매거래는 총 4만2
2019.08.09

"강남권·중간값 이상 아파트가 서울 주택시장 흔들어"
KB국민은행 부동산 통계에 따르면 서울 아파트 매매가는 지난해 9·13 부동
승폭이 줄어들면서 올해 1월 0.01% 하락으로 전환한 이후 최근 기준으로는
2019.08.09

서울 아파트값 8주 연속 상승… 서초·마포·동대문 강세
정부의 민간택지 분양가상한제 도입 발표가 임박한 가운데 서울 아파트값이
났다.9일 KB부동산 리브온의 주간 KB주택시장동향 자료(5일 기준)에 따르

`상한제 포비아`…강남 대표 재건축 아파트값 1억뚝↓
12일 민간택지 분양가상한제 발표를 앞두고 서울 아파트 시장이 숨죽이고
받는 강남 재건축 단지들의 약세가 지속된 가운데 매수 심리도 꺾이면서 관
2019.08.09

자사고+재개발 효과…강남·목동·노원 '주춤' 성동·서대문 '쑥'
전통의 교육1번지인 '강남, 목동, 노원'의 집값 상승률이 2010년대 들어 주
정과 대규모 재개발·재건축 단지 조성이 맞물리면서 이들 지역으로 수요가

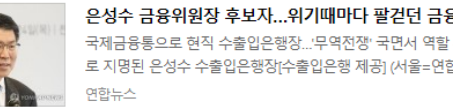
네이버 경제

50 윤동한 한국콜마 회장 • 직원행사서 정부비판 영상 틀어

한국콜마 "국민께 사과…편향된 내용에 현혹되면 안된다
[서울경제] 지난 7일 전 직원이 모인 월례조회에서 한일관계와 문
막말이 담긴 유튜브 동영상을 상영한 것으로 논란을 빚은 한국콜
서울경제 | 300+

한국콜마 "리섭TV 유튜브 인용, 女 부적절 언급 無"(전문) 이데일리 | 10+
직원조회서 "아베는 대단한 지도자" 막말 영상 튼 한국콜마, 결국 사과문 경향
한국콜마 '막말 유튜브' 시청 해명… "일부 인용, 여성 비하 X" 머니투데이 | 50

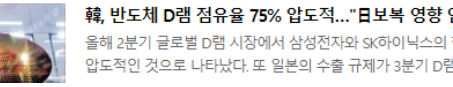
21 은성수 수출입은행장 • 금융위원장 후보자…국제금융 전문가



은성수 금융위원장 후보자…위기때마다 팔걷던 금융전
국제금융통으로 현직 수출입은행장…'무역전쟁' 국면서 역할 주목
로 지명된 은성수 수출입은행장[수출입은행 제공] (서울=연합뉴스
연합뉴스

文대통령, 금융위원장에 은성수 수은 행장 지명 헤럴드경제
은성수 금융위원장 후보자는…"국제금융 전문가" 이데일리
은성수 금융위원장 후보자…"경제 위기때마다 활로 뚫어" 서울경제

20 삼성 • 글로벌D램 점유율 또 최고치



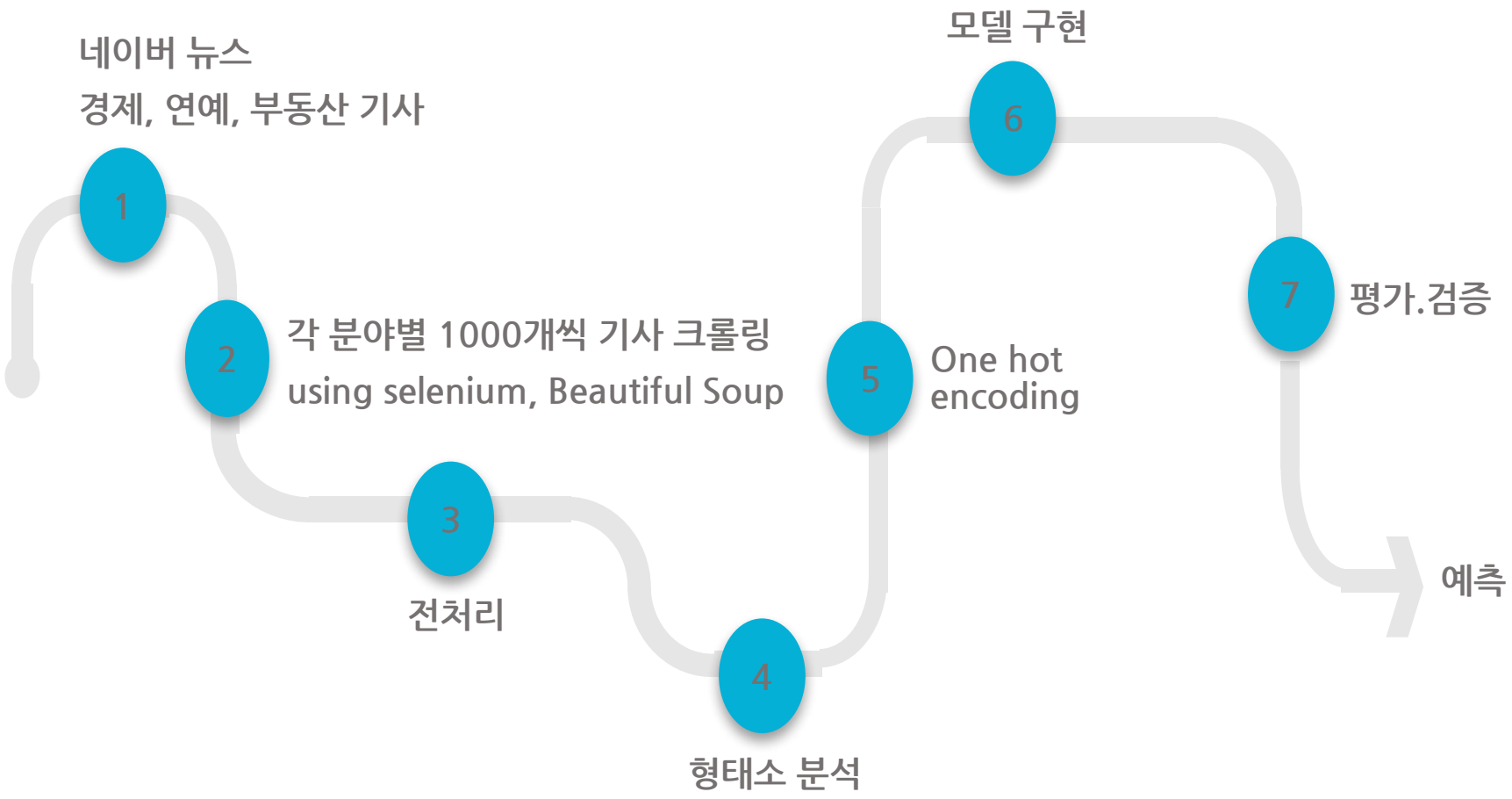
韓, 반도체 D램 점유율 75% 압도적…"日보복 영향 안 클
올해 2분기 글로벌 D램 시장에서 삼성전자와 SK하이닉스의 합산
압도적인 것으로 나타났다. 또 일본의 수출 규제가 3분기 D램 생

네이버 연예

이 본 TV연예 뉴스 더

1 '연애의 맛' 오창석♥이채은, 사귄 지 한 달 만에 위기…'
2 주상욱 아내 차예련, 이렇게 예쁜 엄마 봤어?
3 [종합] "어쩔 수 없이 참아" 이대 백반집 태도 돌변…골목
4 "잘할게 여보 고마워"…가희, 성공적 복귀에 남편♥ 뚝뚝
5 '뉴스룸' "양현석, 美카지노 VIP룸 11번 확인…승리는 13
5 [단독] "불법 환치기 포착됐다"…양현석, 라스베가스 바
7 '나혼자산다' 박나래, 이성 잃은 절정의 식욕파티
8 '엑시트' 하루만에 박스오피스 1위 재탈환 "장기흥행 본
9 "톱모델 비율이란 이런것"…한혜진, 긴 팔·다리 완벽 피
10 뚱뚱한 토르X배 나온 '나홀로집에' 맥컬리 컬킨 "폭소"

분석과정



분야별 1000개씩 기사 크롤링 e.g. 부동산

In [1]:

```
# 부동산 기사 크롤링

import time
from selenium import webdriver
from bs4 import BeautifulSoup

news_land = {'title' : [], 'body' : []}
driver = webdriver.Firefox()
start = 430000
counter = 0

while counter < 1000:
    try:
        driver.get("https://land.naver.com/news/newsRead.nhn?type=headline&bss_ymd=&prsco_id=366&arti_id={:010d}".format(start))

        html = driver.page_source
        soup = BeautifulSoup(html, 'html.parser')

        title = soup.find('div', class_='article_header')
        title_text = title.get_text(strip=True)

        body = soup.find('div', class_='article_body size4')
        body_text = body.get_text(strip=True)

        news_land['body'].append(body_text)
        news_land['title'].append(title_text)
        counter += 1
    except:
        pass

    start += 1
    print(counter)
    time.sleep(1)
```

분야별 1000개씩 기사 크롤링 e.g. 부동산

In [9]:
```
import pandas as pd

land = pd.DataFrame(news_land)
land.to_csv('부동산_수정.csv', index=False)
enter = pd.DataFrame(news_entertain)
enter.to_csv('연예_수정.csv', index=False)
eco = pd.DataFrame(news_eco)
eco.to_csv('경제_수정.csv', index=False)
```

In [10]:
```
land
```

Out[10]:

	title	body
0	삼성전자 "1분기 실적, 시장 기대 밑돌 것"조선비즈|2019.03.26 08:53	삼성전자(005930)는 "당초 예상보다 디스플레이와 메모리 사업의 환경이 좋지 않...
1	[특징주] 아시아나항공, 재감사 적정 의견에도 급락세조선비즈|2019.03.26 0...	아시아나항공 주가가 재감사에서 적정 의견을 받았다는 공시에도 26일 장 초반 크게 ...
2	금호산업 "아시아나항공 '적정'에 따라 금호산업도 '적정' 감사"조선비즈|2019....	금호산업은 자회사인 아시아나항공이 재감사를 통해 '적정의견'의 감사보고서를 받아 모...
3	카카오뱅크, 앱 안에서 신청하는 주식계좌개설 서비스 출시조선비즈|2019.03.26...	카카오은행은 26일 주식계좌개설 신청 서비스를 출시했다고 밝혔다. 이 서비스를 통해...
4	홍남기 부총리 "최저임금법 등 주요 법안 국회 통과해야"조선비즈|2019.03.26...	홍남기 경제부총리 겸 기획재정부 장관은 26일 최저임금법과 서비스발전기본법 등 주요...
5	오비맥주, 카스 병맥주 가격 56원 올린다...평균 5.3% 인상조선비즈|2019....	국내 맥주업계 점유율 1위인 오비맥주가 가격인상을 단행했다.오비맥주는 다음달 4일부...
6	[단독] '씰리침대-리앤산업' 이메일 입수...라돈검출 메모리폼 업체 선정 논란조선...	최근 1급 발암물질 '라돈'이 검출되며 논란을 일으킨 씰리침대가 책임 소재를 놓고 ...
7	부천 영상문화사업단지 복합개발사업 공모제안서 최종 마감조선비즈|2019.03.26 ...	부천시가 '영상문화산업단지 복합개발사업' 관련 25일 공모제안서 제출 마감을 시작으...
8	삼성전자, 동남아서 'QLED TV' 라인업 공개…"프리미엄 시장 잡을 것"조선비즈...	25일부터 이틀간 싱가포르에서 열린 '삼성 동남아 포럼 2019'에 참석한 미디어 ...
9	현대로템 "2022년 매출 4조원, 영업이익률 5% 달성"조선비즈|2019.03.2...	현대로템(064350)은 오는 2022년 매출액 4조원, 영업이익률 5%를 달성하겠...
10	코스피, 외인·기관 쌍끌이 매도에 불안한 출발조선비즈|2019.03.26 09:41	전날 크게 추락했던 코스피지수가 26일 오전에는 보합권에서 등락을 거듭하고 있다. ...
11	코이카, 국제질병퇴치기금 국민참여단 'V-CREATORS' 발대식 개최조선비즈|20...	대한민국 정부 무상원조 전담기관 코이카 (KOICA, 한국국제협력단)는 25일 성남...
12	"청약 열기는 뜨겁지만…" 후분양이 선분양을 대체할 수 있을까?조선비즈|2019.0...	최근 후분양 아파트가 높은 청약 경쟁률을 기록하면서 후분양에 대한 관심이 덩달아 높...
13	KT, 데이터로밍 가입하면 음성로밍 무료 제공 이벤트 실시조선비즈|2019.03.2...	KT가 음성로밍 통화 무료 이벤트를 시행한다.KT는 데이터로밍 요금제를 가입하면 음...
14	한숨 돌렸지만 적자폭 커진 아시아나항공, 유증 나서나조선비즈|2019.03.26 0...	최대 3조원의 회사채와 차입금, 자산유동화증권(ABS)이 동반 디폴트(채무불이행)에...
15	"휴대폰 쓰지 말입니다?"…SK텔레콤, 군 병사 전용 요금제 출시조선비즈|2019....	SK텔레콤이 4월부터 휴대폰 사용이 가능해지는 군 병사를 위한 요금제를 내놓는다.S...
16	[인터뷰] 피터 샤프 "韓 마트·복합몰 강제휴무, 해외선 본 적 없는 규제"조선비즈...	"현대 소비자에게 엔터테인먼트는 일종의 생활방식입니다. 사람들은 제한된 여가 시간을...
17	NHN페이코, '카드조회' 서비스 탑재…"금융 서비스 강화"조선비즈|2019.03....	NHN페이코가 '카드조회' 서비스를 신규 탑재하고 '페이코(PAYCO)' 금융 서비...

전처리

In [87]:
```
import pandas as pd

land.to_csv('부동산_수정.csv', index=False)
enter.to_csv('연예_수정.csv', index=False)
eco.to_csv('경제_수정.csv', index=False)
```

In [86]:
```
# 부동산 : 0, 연예 : 1, 경제 : 2
land['label'] = 0
enter['label'] = 1
eco['label'] = 2
```

In [88]:
```
join = land.append([enter, eco])
join.reset_index(inplace=True)
join.drop('index', axis=1, inplace=True)
```

In [90]:
```
join
```

Out[90]:

	title	body	label
0	삼성전자 "1분기 실적, 시장 기대 밑돌 것"조선비즈|2019.03.26 08:53	삼성전자(005930)는 "당초 예상보다 디스플레이와 메모리 사업의 환경이 좋지 않...	0
1	[특징주] 아시아나항공, 재감사 적정 의견에도 급락세조선비즈|2019.03.26 0...	아시아나항공 주가가 재감사에서 적정 의견을 받았다는 공시에도 26일 장 초반 크게 ...	0
2	금호산업 "아시아나항공 '적정'에 따라 금호산업도 '적정' 감사"조선비즈|2019....	금호산업은 자회사인 아시아나항공이 재감사를 통해 '적정의견'의 감사보고서를 받아 모...	0
3	카카오뱅크, 앱 안에서 신청하는 주식계좌개설 서비스 출시조선비즈|2019.03.26...	카카오은행은 26일 주식계좌개설 신청 서비스를 출시했다고 밝혔다. 이 서비스를 통해...	0
4	홍남기 부총리 "최저임금법 등 주요 법안 국회 통과해야"조선비즈|2019.03.26...	홍남기 경제부총리 겸 기획재정부 장관은 26일 최저임금법과 서비스발전기본법 등 주요...	0
5	오비맥주, 카스 병맥주 가격 56원 올린다...평균 5.3% 인상조선비즈|2019....	국내 맥주업계 점유율 1위인 오비맥주가 가격인상을 단행했다.오비맥주는 다음달 4일부...	0
6	[단독] '씰리침대-리앤산업' 이메일 입수...라돈검출 메모리폼 업체 선정 논란조선...	최근 1급 발암물질 '라돈'이 검출되며 논란을 일으킨 씰리침대가 책임 소재를 놓고 ...	0
7	부천 영상문화사업단지 복합개발사업 공모제안서 최종 마감조선비즈|2019.03.26 ...	부천시가 '영상문화산업단지 복합개발사업' 관련 25일 공모제안서 제출 마감을 시작으...	0
8	삼성전자, 동남아서 'QLED TV' 라인업 공개…"프리미엄 시장 잡을 것"조선비즈...	25일부터 이틀간 싱가포르에서 열린 '삼성 동남아 포럼 2019'에 참석한 미디어 ...	0
9	현대로템 "2022년 매출 4조원, 영업이익률 5% 달성"조선비즈|2019.03.2...	현대로템(064350)은 오는 2022년 매출액 4조원, 영업이익률 5%를 달성하겠...	0

예측 모델 만들기: 기사 내용

In [211]:
```
from sklearn.model_selection import train_test_split

train_x, test_x, train_y, test_y = train_test_split(join['body'], join['label'], test_size=0.5)
```

In [212]:
```
from keras.preprocessing.text import Tokenizer

token = Tokenizer(num_words=10000)
token.fit_on_texts(train_x)
train_x = token.texts_to_matrix(train_x)
test_x = token.texts_to_matrix(test_x)
```

In [213]:
```
from keras.utils import to_categorical

train_y = to_categorical(train_y)
test_y = to_categorical(test_y)
```

In [214]:
```
len(train_x)
```

Out[214]: 1500

In [215]:
```
from keras import models
from keras import layers

model = models.Sequential()
model.add(layers.Dense(64, activation='relu', input_shape=(10000,)))
model.add(layers.Dense(64, activation='relu'))
model.add(layers.Dense(3, activation='softmax'))
```

In [216]:
```
model.compile(optimizer='adam',
              loss='categorical_crossentropy',
              metrics=['accuracy'])
```

In [217]:
```
history = model.fit(train_x,
                    train_y,
                    epochs=10,
                    batch_size=512,
                     validation_split=0.2)
```

```
Train on 1200 samples, validate on 300 samples
Epoch 1/10
1200/1200 [==============================] - 1s 806us/step - loss: 1.0287 - acc: 0.5525 - val_loss: 0.7810 - val_acc: 0.9533
Epoch 2/10
1200/1200 [==============================] - 0s 139us/step - loss: 0.6810 - acc: 0.9858 - val_loss: 0.4905 - val_acc: 0.9667
Epoch 3/10
```

In [218]:
```
model.evaluate(test_x, test_y)
```

```
1500/1500 [==============================] - 0s 126us/step
```

Out[218]: [0.011815074673543374, 0.9966666666666667]

예측 모델 만들기: 기사 제목

In [192]:
```
from sklearn.model_selection import train_test_split

train_x, test_x, train_y, test_y = train_test_split(join['title'], join['label'], test_size=0.5)
```

In [193]:
```
from keras.preprocessing.text import Tokenizer

token = Tokenizer(num_words=10000)
token.fit_on_texts(train_x)
train_x = token.texts_to_matrix(train_x)
test_x = token.texts_to_matrix(test_x)
```

In [194]:
```
from keras.utils import to_categorical

train_y = to_categorical(train_y)
test_y = to_categorical(test_y)
```

In [195]:
```
from keras import models
from keras import layers

model = models.Sequential()
model.add(layers.Dense(64, activation='relu', input_shape=(10000,)))
model.add(layers.Dense(64, activation='relu'))
model.add(layers.Dense(3, activation='softmax'))
```

In [196]:
```
model.compile(optimizer='adam',
              loss='categorical_crossentropy',
              metrics=['accuracy'])
```

In [197]:
```
history = model.fit(train_x,
                    train_y,
                    epochs=20,
                    batch_size=512,
                     validation_split=0.2)
```

```
Train on 1200 samples, validate on 300 samples
Epoch 1/20
1200/1200 [==============================] - 1s 793us/step - loss: 1.0939 - acc: 0.3792 - val_loss: 1.0792 - val_acc: 0.5900
Epoch 2/20
1200/1200 [==============================] - 0s 148us/step - loss: 1.0487 - acc: 0.8792 - val_loss: 1.0508 - val_acc: 0.7967
Epoch 3/20
```

In [198]:
```
model.evaluate(test_x, test_y)
```

```
1500/1500 [==============================] - 0s 148us/step
```

Out[198]:
```
[0.2411182912985484, 0.9399999995231628]
```

새로운 기사로 예측하기

In [189]:
```
import requests
from bs4 import BeautifulSoup

land_new = requests.get("https://land.naver.com/news/newsRead.nhn?type=headline&bss_ymd=&prsco_id=366&arti_id=0000400000")
enter_new = requests.get("https://entertain.naver.com/ranking/read?oid=109&aid=0004000000")
#eco_new = requests.get("https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=101&oid=421&aid=0004000121")

land_html = land_new.text
enter_html = enter_new.text
#eco_html = eco_new.text

land_soup = BeautifulSoup(land_html, 'html.parser')
enter_soup = BeautifulSoup(enter_html, 'html.parser')
#eco_soup = BeautifulSoup(eco_html, 'html.parser')

land_title = land_soup.find('div', class_='article_header').get_text(strip=True)
enter_title = enter_soup.find('h2', class_='end_tit').get_text(strip=True)
#eco_title = eco_soup.find('h3', class_='tts_head').get_text(strip=True)

land_body = land_soup.find('div', class_='article_body size4').get_text(strip=True)
enter_body = enter_soup.find('div', class_='end_body_wrp').get_text(strip=True)
#eco_body = eco_soup.find('div', class_='_article_body_contents').get_text(strip=True)

# 경제 기사는 에러나서 pass
```

새로운 기사로 예측하기: 예측 결과

부동산 : 0, 연예 : 1, 경제 : 2

```
In [209]: model.predict(token.texts_to_matrix([land_title]))
Out[209]: array([[0.9459676 , 0.01787367, 0.03615881]], dtype=float32)

In [210]: model.predict(token.texts_to_matrix([enter_title]))
Out[210]: array([[0.03899919, 0.8271603 , 0.13384044]], dtype=float32)

In [219]: model.predict(token.texts_to_matrix([land_body]))
Out[219]: array([[0.9238843 , 0.01786363, 0.05825203]], dtype=float32)

In [220]: model.predict(token.texts_to_matrix([enter_body]))
Out[220]: array([[2.1687846e-03, 9.9690920e-01, 9.2208211e-04]], dtype=float32)

In [1]: 9.9690920e-01
Out[1]: 0.9969092

In [2]: 2.1687846e-03
Out[2]: 0.0021687846
```

THANK YOU