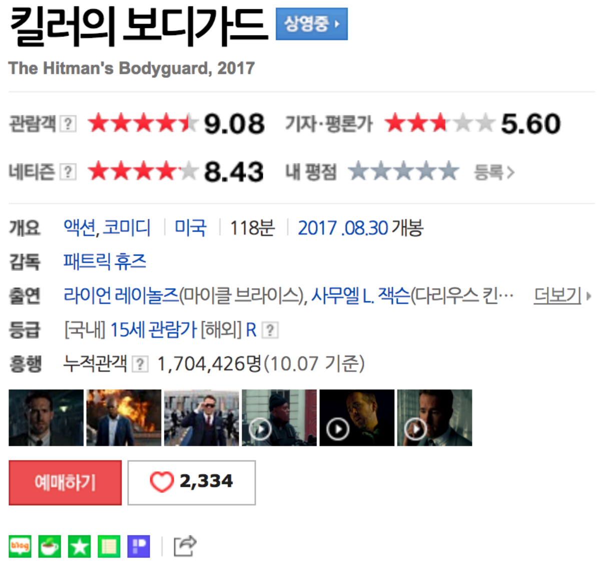
영화 평점 예측 감정 분석

주재걸 교수님 연구실
DAVIAN Lab.

강경필

1. Introduction



킬러의 보디가드 상영중 ›

The Hitman's Bodyguard, 2017

관람객 ? ★★★★★ 9.08 기자·평론가 ★★★★★ 5.60

네티즌 ? ★★★★★ 8.43 내 평점 ★★★★★ 등록 ›

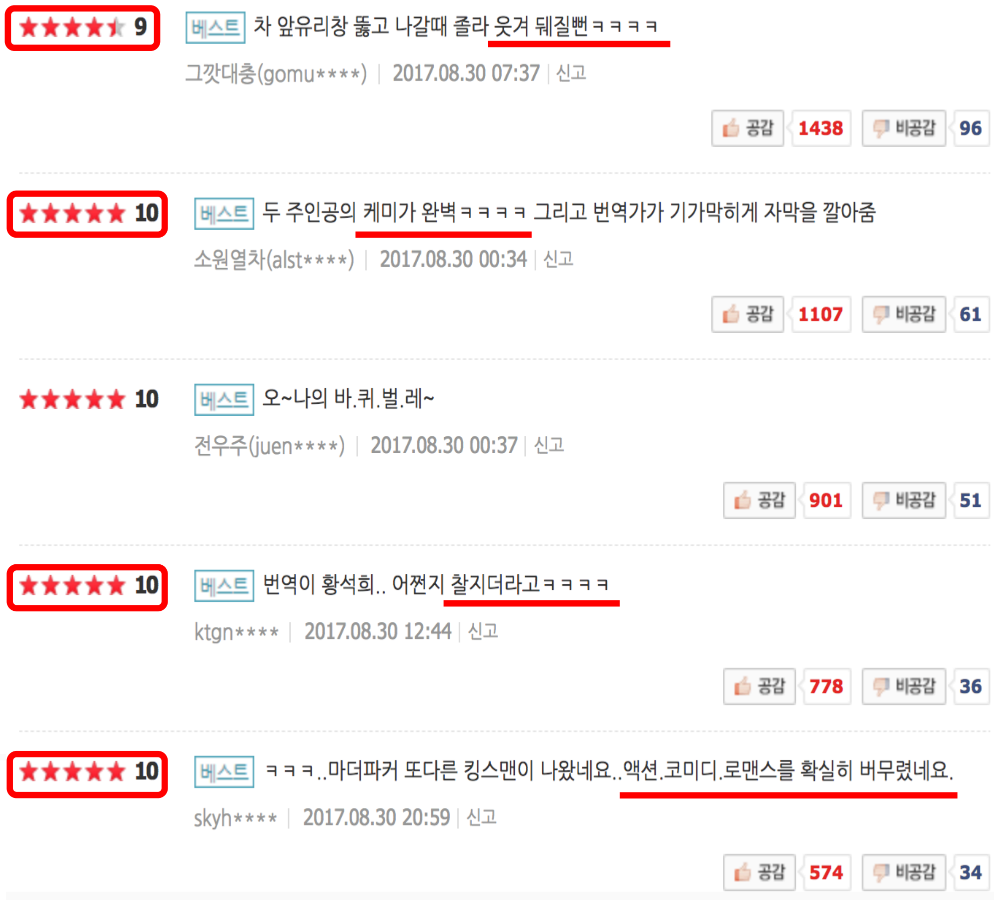
개요 액션, 코미디 | 미국 | 118분 | 2017 .08.30 개봉

감독 패트릭 휴즈

출연 라이언 레이놀즈(마이클 브라이스), 사무엘 L. 잭슨(다리우스 킨… 더보기 ›

등급 [국내] 15세 관람가 [해외] R ?

흥행 누적관객 ? 1,704,426명(10.07 기준)



★★★★★ 9 베스트 차 앞유리창 뚫고 나갈때 졸라 웃겨 뒤질뻔ㅋㅋㅋㅋ

그깟대충(gomu****) | 2017.08.30 07:37 | 신고

공감 1438 비공감 96

★★★★★ 10 베스트 두 주인공의 케미가 완벽ㅋㅋㅋㅋ 그리고 번역가가 기가막히게 자막을 깔아줌

소원열차(alst****) | 2017.08.30 00:34 | 신고

공감 1107 비공감 61

★★★★★ 10 베스트 오~나의 바.퀴.벌.레~

전우주(juen****) | 2017.08.30 00:37 | 신고

공감 901 비공감 51

★★★★★ 10 베스트 번역이 황석희.. 어쩐지 찰지더라고ㅋㅋㅋㅋ

ktgn**** | 2017.08.30 12:44 | 신고

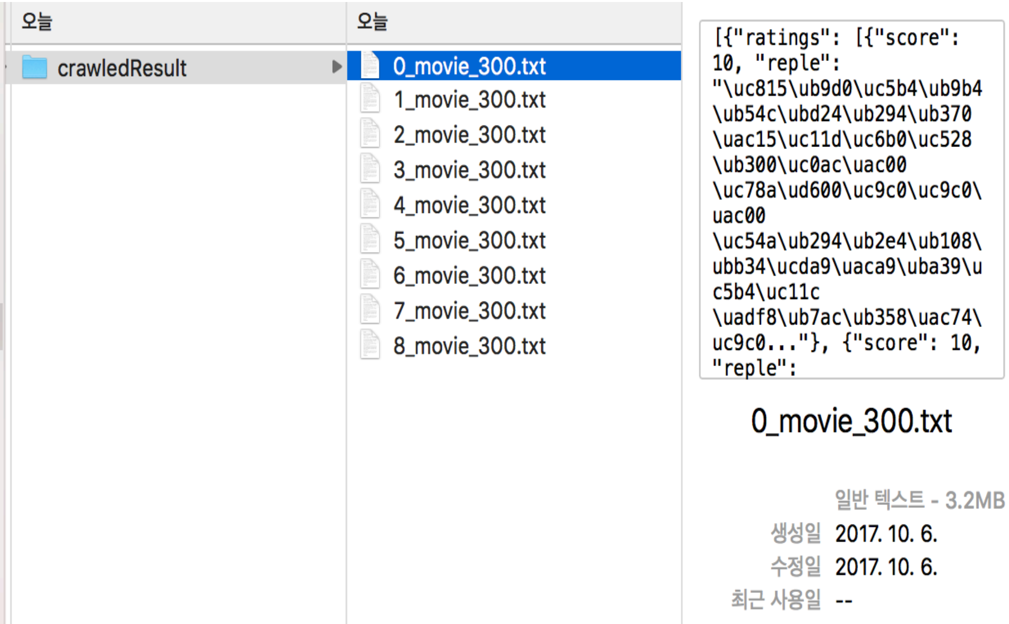
공감 778 비공감 36

★★★★★ 10 베스트 ㅋㅋㅋ..마더파커 또다른 킹스맨이 나왔네요..액션.코미디.로맨스를 확실히 버무렸네요.

skyh**** | 2017.08.30 20:59 | 신고

공감 574 비공감 34

2. Data

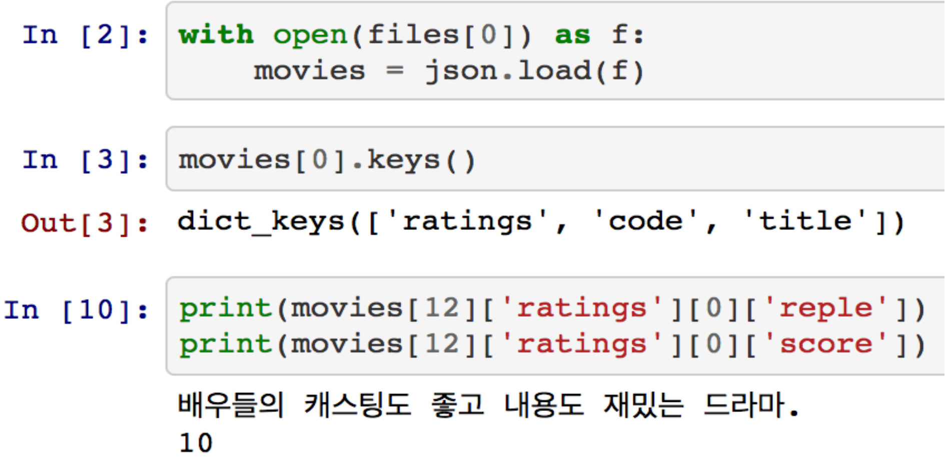


0_movie_300.txt
1_movie_300.txt
2_movie_300.txt
3_movie_300.txt
4_movie_300.txt
5_movie_300.txt
6_movie_300.txt
7_movie_300.txt
8_movie_300.txt

```
[{"ratings": [{"score":
10, "reple":
"\uc815\ub9d0\uc5b4\ub9b4
\ub54c\ubd24\ub294\ub370
\uac15\uc11d\uc6b0\uc528
\ub300\uc0ac\uac00
\uc78a\ud600\uc9c0\uc9c0\
uac00
\uc54a\ub294\ub2e4\ub108\
ubb34\ucda9\uaca9\uba39\u
c5b4\uc11c
\uadf8\ub7ac\ub358\uac74\
uc9c0..."}, {"score": 10,
"reple":
```

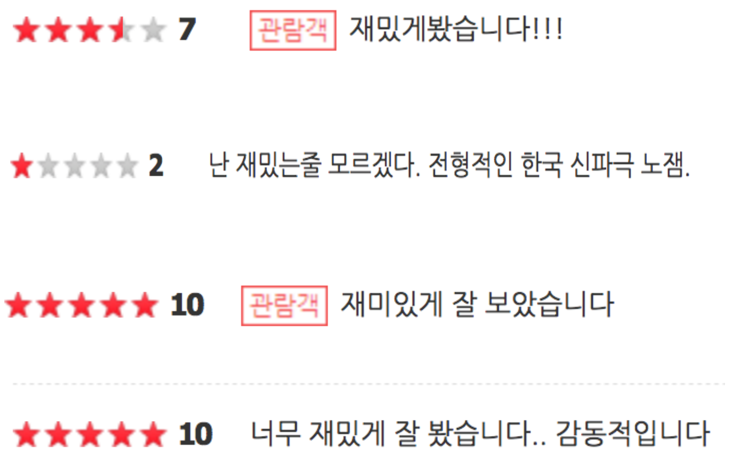
0_movie_300.txt

일반 텍스트 - 3.2MB
생성일 2017. 10. 6.
수정일 2017. 10. 6.
최근 사용일 --



```
In [2]: with open(files[0]) as f:
            movies = json.load(f)

In [3]: movies[0].keys()

Out[3]: dict_keys(['ratings', 'code', 'title'])

In [10]: print(movies[12]['ratings'][0]['reple'])
         print(movies[12]['ratings'][0]['score'])

         배우들의 캐스팅도 좋고 내용도 재밌는 드라마.
         10
```

3. Dictionary based model



7 관람객 재밌게봤습니다!!!

2 난 재밌는줄 모르겠다. 전형적인 한국 신파극 노잼.

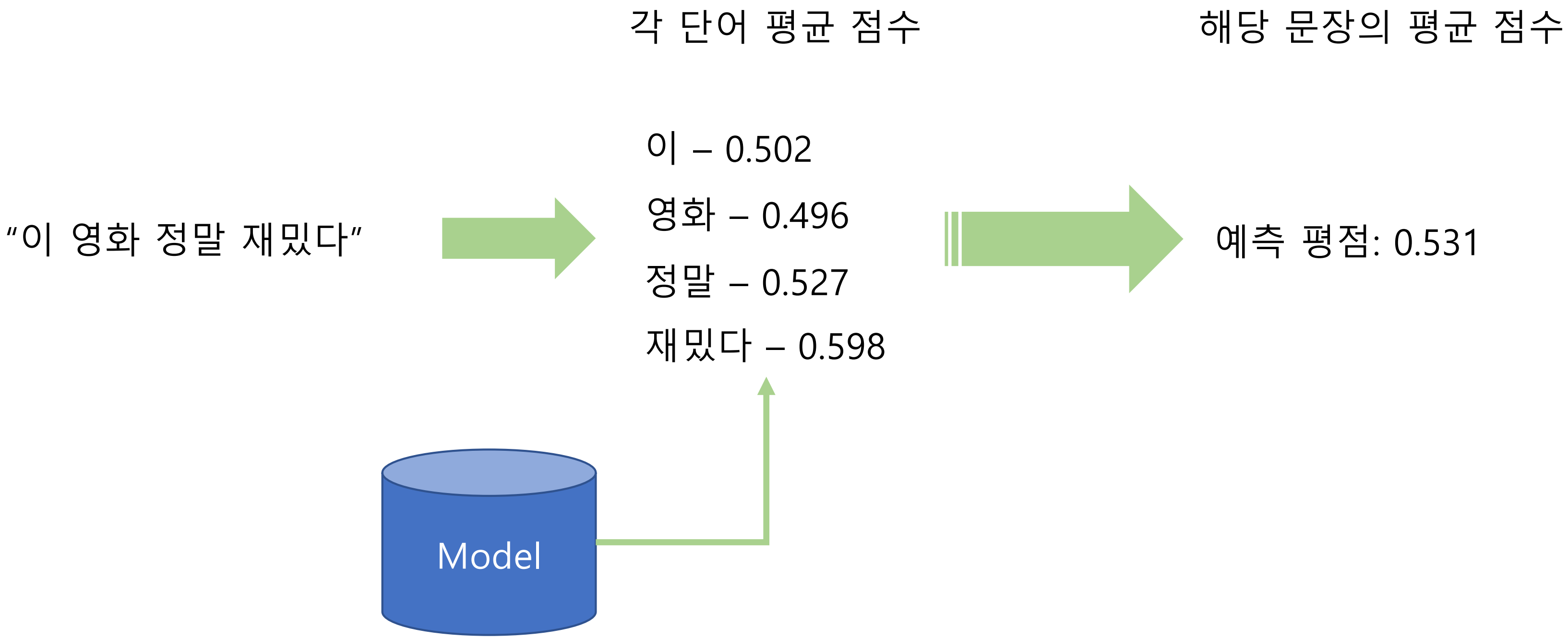
10 관람객 재미있게 잘 보았습니다

10 너무 재밌게 잘 봤습니다.. 감동적입니다

각 단어마다 평균 감정 점수를 계산하자!

“재밌다” => 7점, 2점, 10점, 10점 => **7.25점**

3. Dictionary based model



각 단어 평균 점수

해당 문장의 평균 점수

“이 영화 정말 재밌다”

이 – 0.502

영화 – 0.496

정말 – 0.527

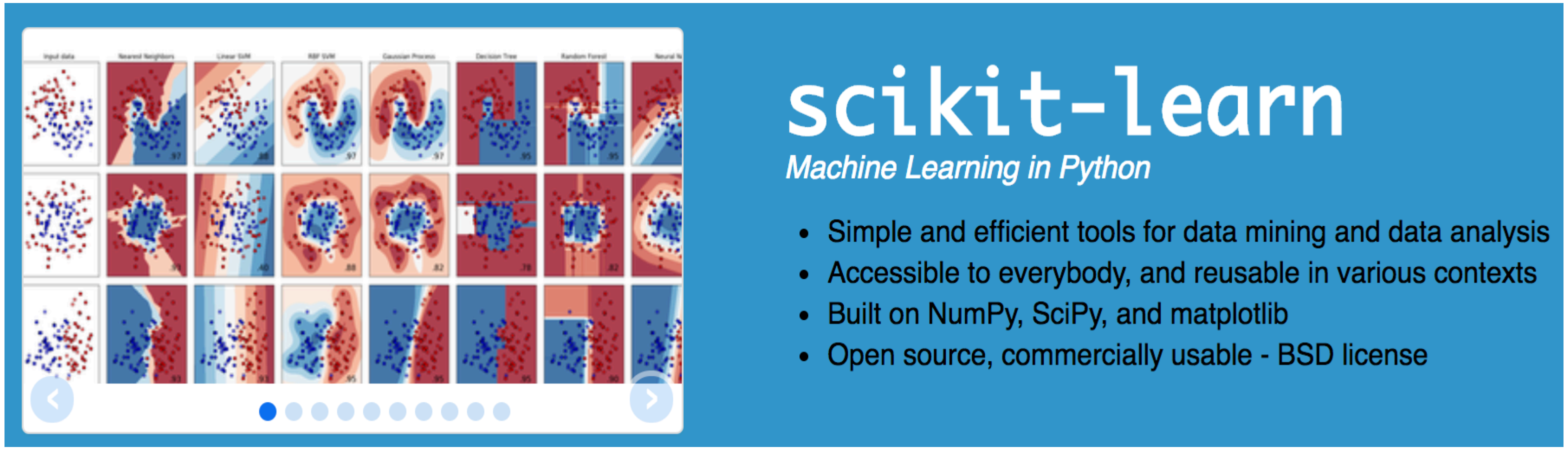
재밌다 – 0.598

예측 평점: 0.531

Model

4. ML based models

- Scikit-Learn



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

- 기존 기계학습(Classification, Regression, Clustering 등) 모델들
- 매우 빠름(C++ 등 구현됨, multiprocessing 지원)
- 다양한 utility 지원
- 쉽고 직관적인 API

```
model = Model()
model.fit(train_X, train_y)
model.predict(X)
```

4. ML based models

- Linear regression

Linear regression equation

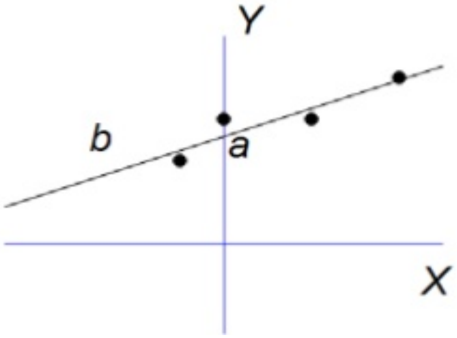
(without error)

$$\hat{Y} = bX + a$$

predicted values of *Y*

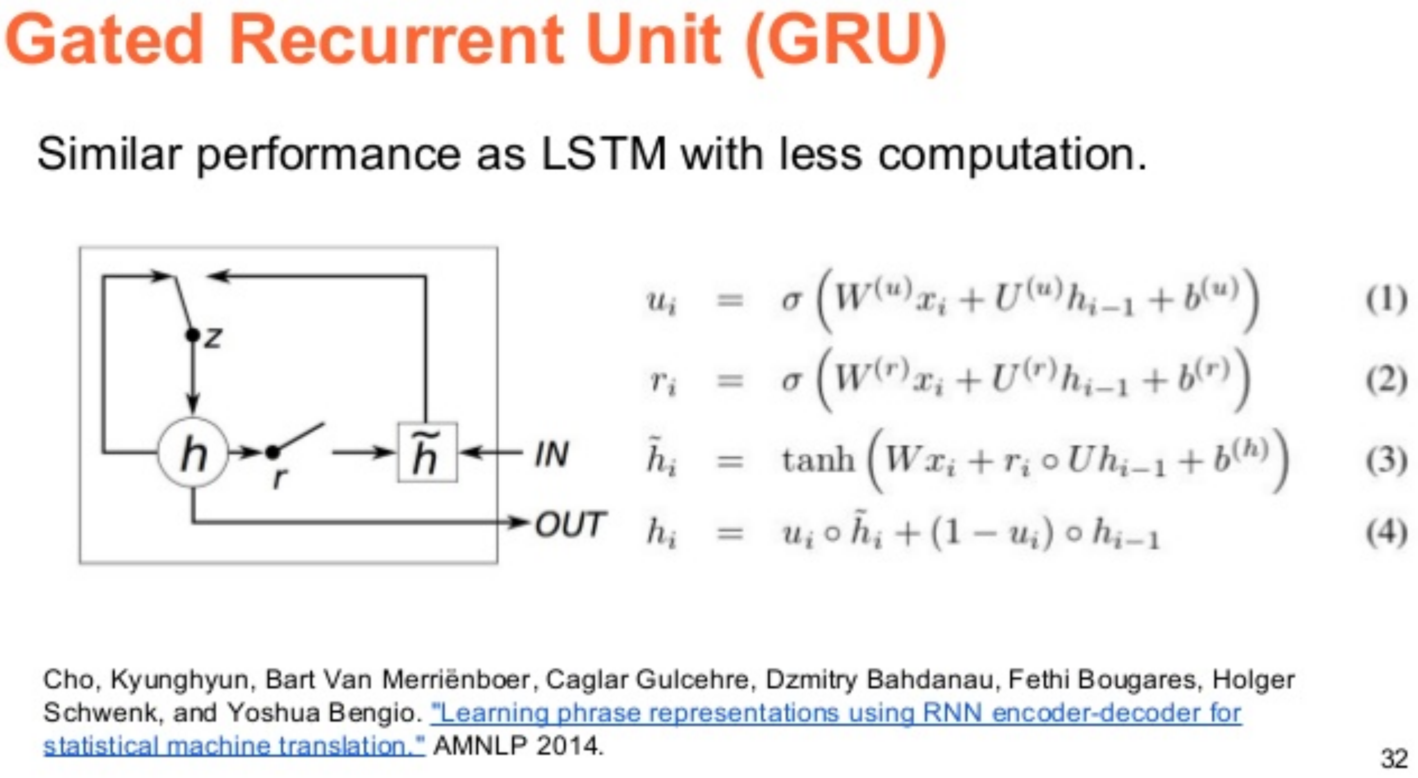
b = slope = rate of predicted ↑/↓ for Y scores for each unit increase in *X*

Y-intercept = level of *Y* when X is 0



5. Deep learning based models

- Gated Recurrent Unit



Gated Recurrent Unit (GRU)

Similar performance as LSTM with less computation.

$$u_i = \sigma\left(W^{(u)}x_i + U^{(u)}h_{i-1} + b^{(u)}\right) \quad (1)$$

$$r_i = \sigma\left(W^{(r)}x_i + U^{(r)}h_{i-1} + b^{(r)}\right) \quad (2)$$

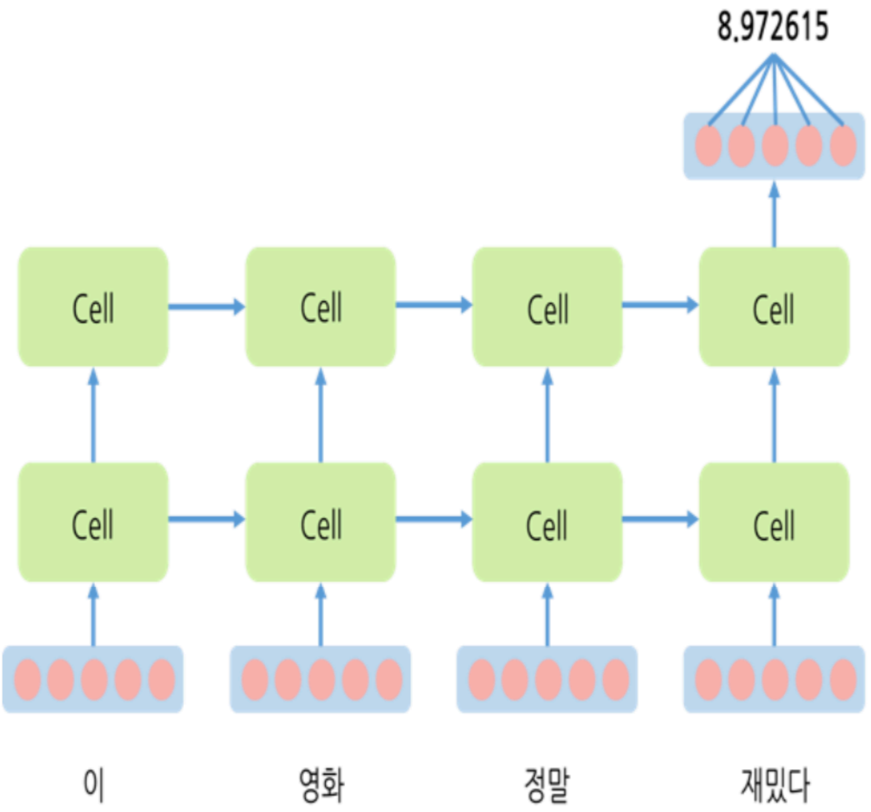
$$\tilde{h}_i = \tanh\left(Wx_i + r_i \circ Uh_{i-1} + b^{(h)}\right) \quad (3)$$

$$h_i = u_i \circ \tilde{h}_i + (1 - u_i) \circ h_{i-1} \quad (4)$$

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." AMNLP 2014.

5. Deep learning based models

- RNN Model



Linear Layer

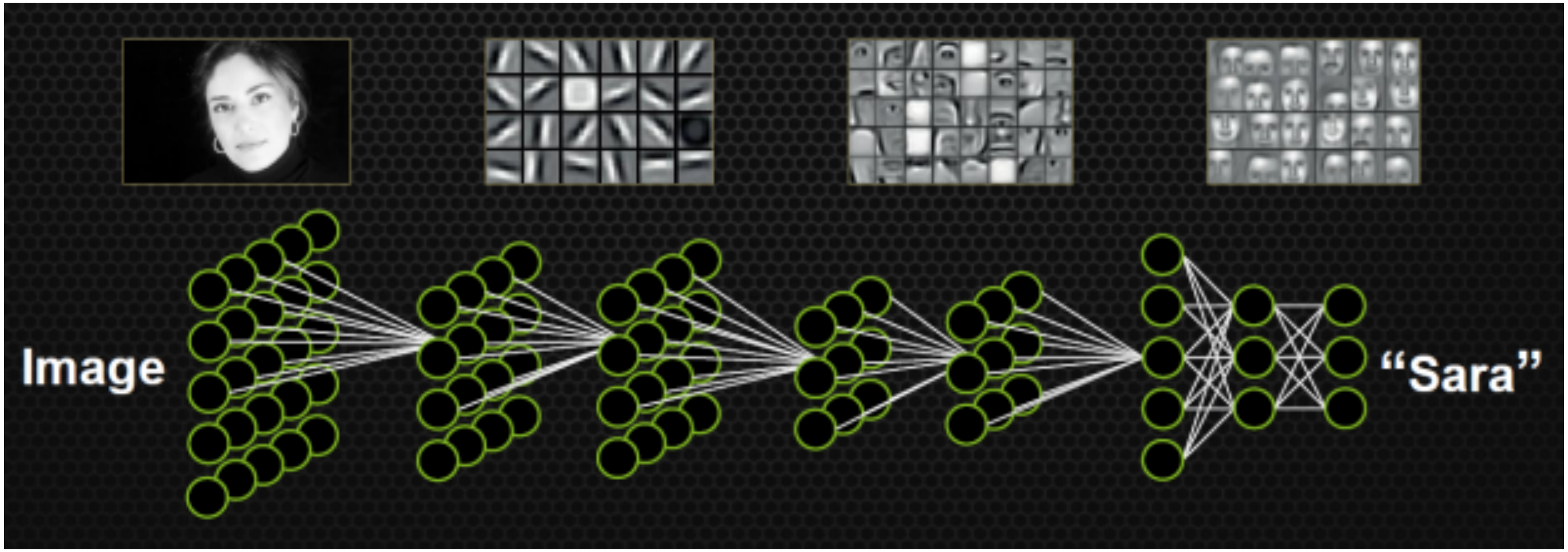
GRU Layer

GRU Layer

Embedding

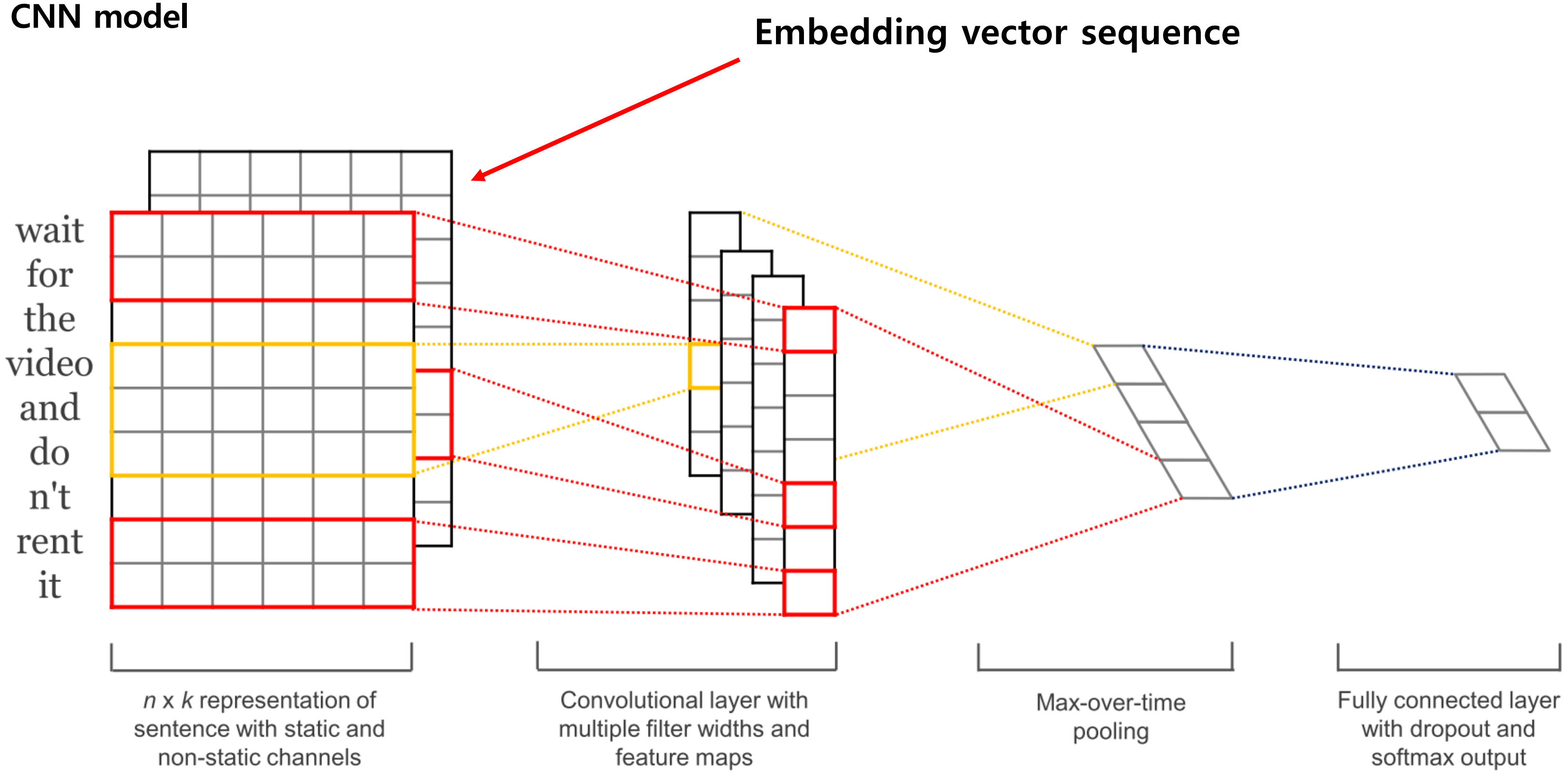
5. Deep learning based models

- Convolutional neural networks



5. Deep learning based models

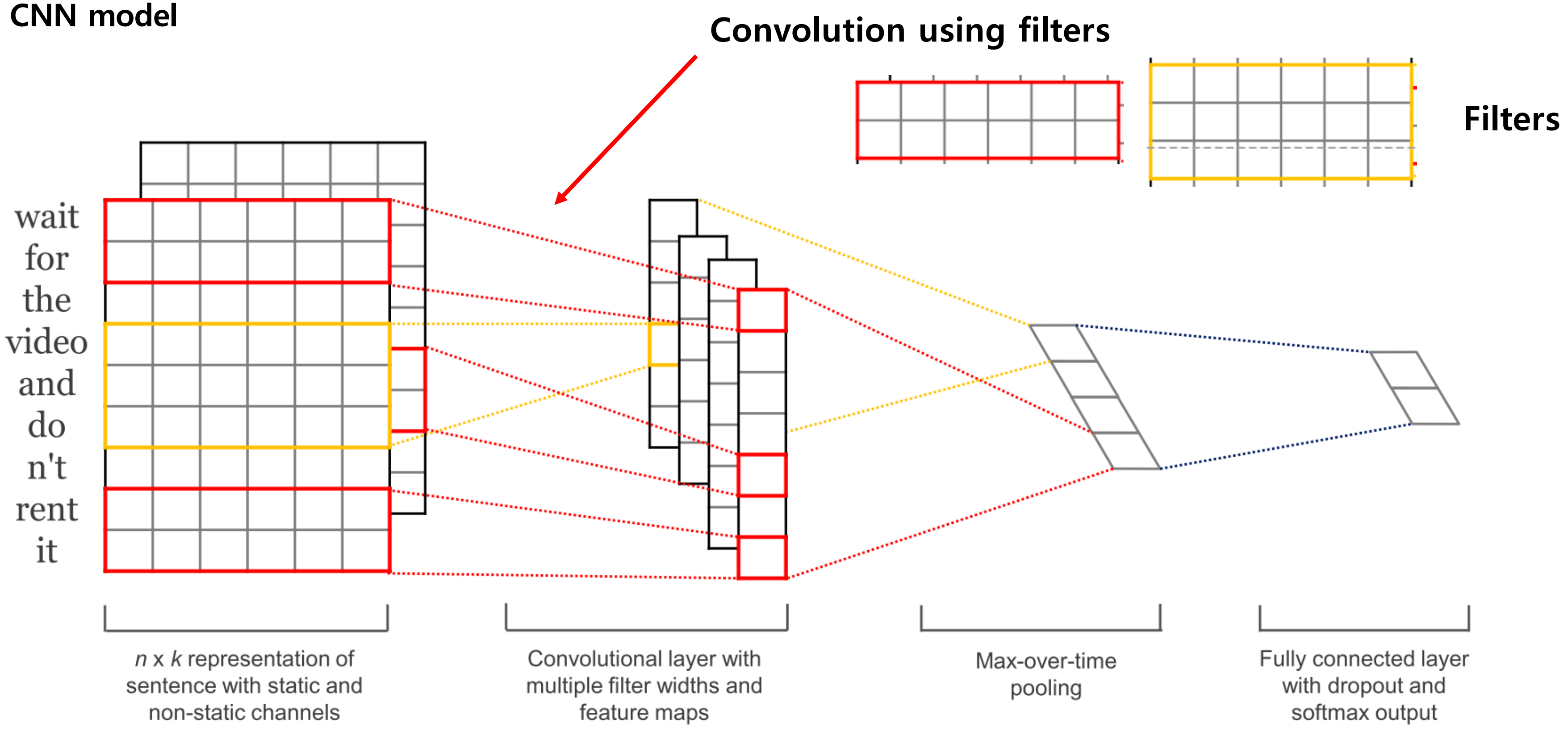
- CNN model


Embedding vector sequence
wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Convolutional Neural Networks for Sentence Classification, Yoon Kim, 2014

5. Deep learning based models

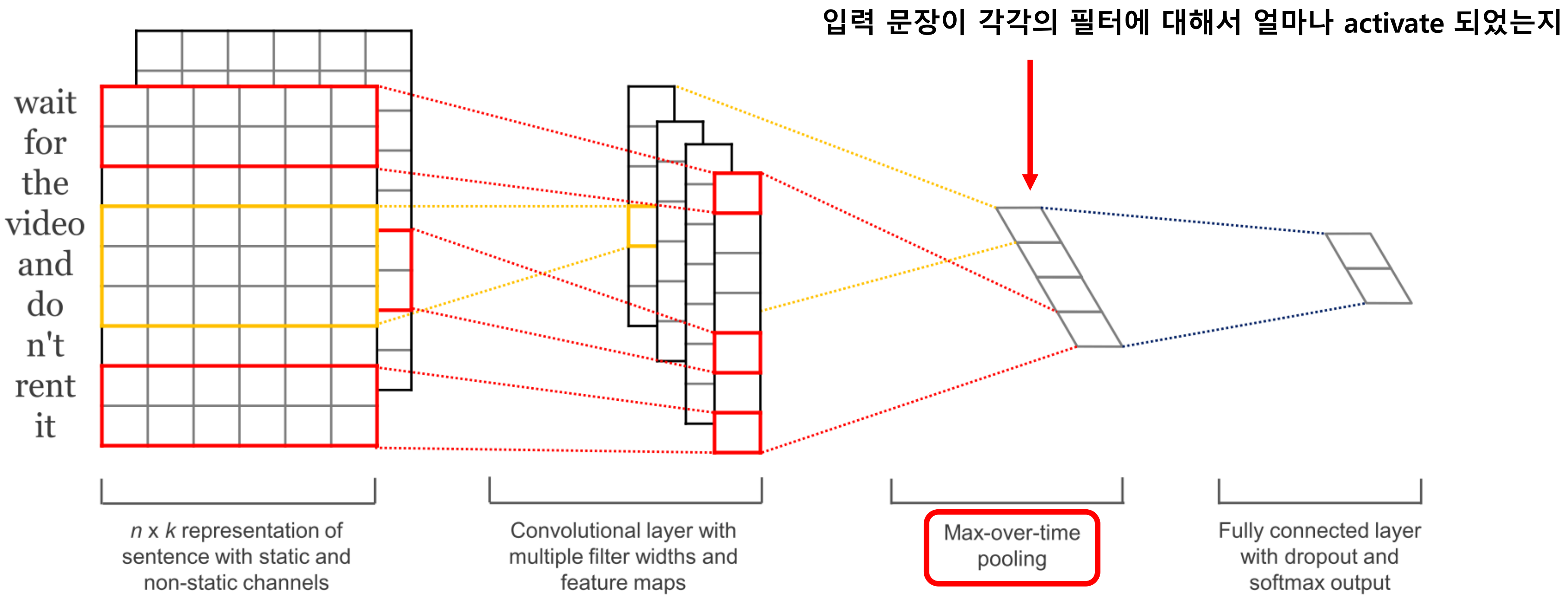
- CNN model


Convolution using filters
Filters
wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Convolutional Neural Networks for Sentence Classification, Yoon Kim, 2014

5. Deep learning based models

- CNN model


입력 문장이 각각의 필터에 대해서 얼마나 activate 되었는지
wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Convolutional Neural Networks for Sentence Classification, Yoon Kim, 2014

6. 모델 비교

Dictionary based model: 제일 간단, 성능은 낮음

ML models : 각 단어에 가중치 부여, 성능이 나쁘지 않음

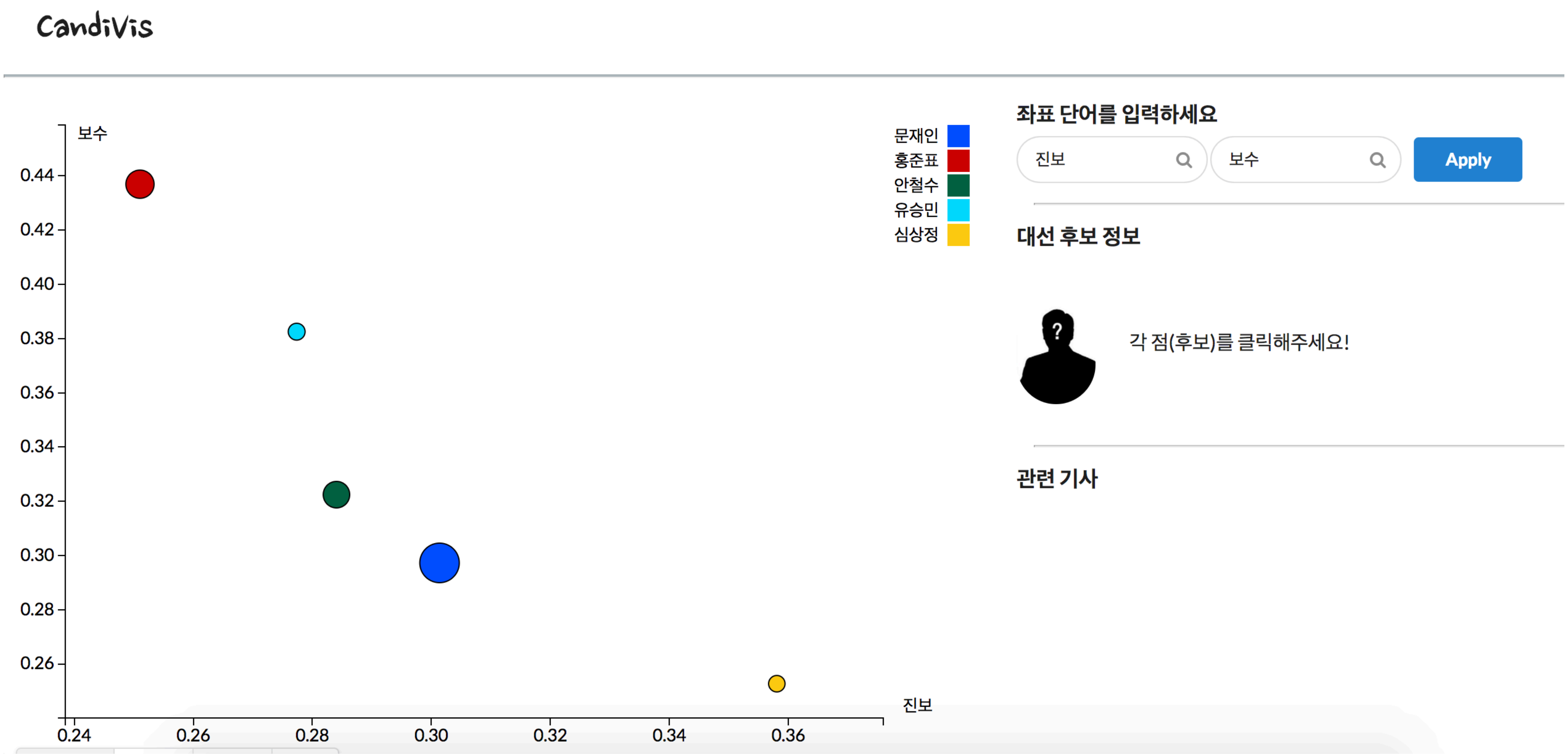
- Linear Regression
- Ridge Regression
- GradientBoostingRegression

Deep learning models: 맥락 고려, Word embedding 사용, 최적화 필요

- CNN model
- RNN model

모델 복잡,
성능 좋음

Visualization – Example



좌표 단어를 입력하세요

진보

보수

대선 후보 정보

각 점(후보)를 클릭해주세요!

관련 기사

감사합니다

Any Questions?

rudvlf0413@naver.com