[숙제13] 악성 트윗 분류기 개발

언어와 컴퓨터

2018년 12월 17일 월요일 13시까지

- 소스코드 스크립트 `automate_group0.py`, 분류기 `classifier_group0.pkl` 및 보고서 `hw13_group0.pdf` 파일을 `hw13_group0.zip` 파일로 압축하여 제출하라. `0`는 조의 번호이다.
- 보고서는 문제 해결 방법, 코드 설명, 실행 결과 등을 포함하여 작성하라.

1 악성 트윗 분류기 개발

이 과제의 목표는 영어 트윗이 혐오발언이거나 공격적인지를 자동으로 식별하는 모형을 만드는 것이다.

1.1 데이터

데이터 파일 `train.txt` [1]는 혐오 발언이 될 수 있는 단어를 포함한 트윗 22,000개에 사람의 분류 결과가 추가된 코퍼스이다. 각 행은 분류 결과(`'hate', 'offensive', 'non-offensive'`)와 본문 2개 필드로 이루어져 있고, 각 필드는 탭(`\t`)으로 구분되어 있다. Excel 등 스프레드시트를 처리하는 프로그램에서 불러오면 아래와 같은 표의 형태로 표현된다.

offensive	RT @shayconnnor: u ain't sailor moon bitch
offensive	@RAEPaelMunturo @WorldStarFunny aye sae en feeld kawkin dyke
offensive	RT @EHSSecretAdmire: "Can @CarterMarteeny murder my pussy?"
non-offensive	I have to dress up tomorrow for mock interviews and I still don't know how to tie the tie 😒
hate	RT @WhitesOnly_1: #niggers! http://t.co/Hb3uJaLky2
offensive	@JoeTheMailman @emilyrs But that bitch could care less about Benghazi victims!

이 데이터의 트윗은 이미 무작위로 배열되어 있으므로, 학습 과정에서 순서를 다시 섞을 필요는 없다.

1.2 분류기 학습

위의 데이터를 사용하여 각 트윗의 범주를 예측하는 분류기를 학습시키라. 분류기로는 단순 베이즈 분류기와 최대 엔트로피 분류기(=로지스틱 회귀분석)를, 기계학습 도구로는 NLTK와 scikit-learn을 사용할 수 있다.

분류기의 성능을 개선하기 위해 아래의 세 가지 내용을 중심으로 다양한 방법을 탐색하고 시도하라.

- 데이터 전처리
- 특성 선택
- 모형 매개변수(add-k 평탄화의 k 등) 조절

이러한 기법을 NLTK나 scikit-learn에서 구현하는 함수에 대해서는 아래의 웹 문서를 참조할 수 있다.

- NLTK
  - `http://www.nltk.org/howto/classify.html`
  - `http://www.nltk.org/book/ch06.htm`

[1] `https://github.com/t-davidson/hate-speech-and-offensive-language`에 공개된 자료를 가공하였다. 이 자료는 아래의 논문에서 발표된 것이다.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *International AAAI Conference on Web and Social Media*. Retrieved from `https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665`

- scikit-learn
  - `https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html`

데이터와 함께 공개된 원저자의 코드를 참조해도 괜찮다.

- `https://github.com/t-davidson/hate-speech-and-offensive-language`

적절한 기법을 찾기 위해서는 데이터의 성질과 기존 연구 내용을 잘 이해하는 것이 중요하다. 기존 연구로 대표적인 것은 아래와 같은 논문이 있으며, 이외에도 많은 연구가 존재한다.

- Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop* (pp. 88–93). Retrieved from `https://aclanthology.coli.uni-saarland.de/papers/N16-2013/n16-2013`

선택한 기법을 평가할 개발 집합은 훈련 집합 내에서 자유롭게 선택하되, 아래의 내용도 고려해 보라.

- Cross-validation이라는 기술을 조사해서 사용하면 점수를 더 받을 수 있다.

최종적으로 결정한 기법에 대하여

1. 데이터를 어떻게 처리했는지
2. 무슨 분류기를 사용했는지
3. 어떤 특성을 선택했는지
4. 개발 집합에서 정확도·정밀도·재현율·F_1 값이 얼마였는지

등을 포함하여 서술하고, 이 분류기를 피클 파일 `classifier_group0.pkl`로 저장하여 함께 제출하라.

1.3 평가

필요한 내용을 충분히 탐색하고 보고서를 알차게 작성하면 좋은 점수를 받을 수 있다. 과제에서 공개되지 않은 실험 집합에 대한 분류 결과도 성적에 반영된다.