SVM 실험 보고서

월 7, 8, 9교시

실용자연어처리

프랑스어학부FATI전공

201901562 박시연

**1. 성능 표**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CountVectorizer 수정** | **N-gram** | **TF-IDF** | **형태소**  **분석 입력** | **불용어**  **처리** | **성능** | **모델 이름** |
| O | (1,2) | X | O | X | 84.03 | 모델 8 |
| O | (1,3) | X | O | X | 85.11 | 모델 9 |
| O | (1,1) | O | X | X | 80.88 | 모델 10 |
| O | (1,3) | O | X | X | 81.85 | 모델 12 |
| O | (1,1) | O | O | X | 83.76 | 모델 13 |
| O | (1,2) | O | O | X | 85.98 | 모델 14 |
| O | (1,3) | O | O | X | 86.20 | 모델 15 |
| X | (1,2) | O | O | O | 79.83 | 모델 16 |

**2. 실험 분석**

(1) 모델 8과 9을 비교했을 때, 같은 조건일 때 N-gram을 uni-gram만을 사용했을 때보다 uni-gram, bi-gram을 함께 사용했을 때 성능이 1.08%p 향상되었다. 모델 13과 14을 비교했을 때도, uni-gram만을 사용했을 때보다 uni-gram, bi-gram을 함께 사용했을 때 성능이 2.22%p 향상되었다. 또한 모델 14와 15를 비교하면, uni-gram, bi-gram을 사용했을 때 보다, tri-gram도 함께 사용했을 때 성능이 0.22%p 향상되었다. 이는 bi-gram과 tri-gram을 사용함으로써, 자주 함께 사용되는 단어들의 조합이나, 단어 간의 순서를 고려하게 됨으로써 성능에 긍정적 영향을 주었다고 예측할 수 있다.

(2) 모델 12와 15를 비교했을 때, 같은 조건일 때 형태소 분석을 했을 때가 그렇지 않은 모델에 비해 성능이 4.35%p 향상된 것을 알 수 있다. 이는 형태소 분석을 통해, 문장 구성 성분과 품사를 파악하게 됨으로써 모델의 성능의 긍정적 영향을 준 것으로 볼 수 있다.

(3) 모델 14와 16을 비교했을 때, 같은 조건일 때 CountVectorizer를 사용했을 때가, 불용어 처리를 했을 때보다 성능이 6.15%p 향상되었다. 이는 불용어 처리를 통해 처리하지 않도록 한 단어들 중 일부가 실제로는 분석 상 필요한 어휘였기 때문에 성능 저하를 이끌었다고 예상된다.

(4) 모델 9와 모델 15를 비교했을 때, 같은 조건일 때 TF-IDF 가중치를 사용했을 때가 그렇지 않을 때보다 성능이 1.09%p 향상되었다. 이는 TF-IDF 가중치를 통해 단어의 해당 문서 내 사용빈도와, 타 문서에서의 사용 빈도 등을 고려하여 단어의 중요도를 파악했기 때문이 이것이 성능 향상과 연결되었음을 보여준다.

**3. 최종 모델 선택**

모델 14를 선택.

실험 분석 결과, 불용어 처리는 하지 않고, CountVectorizer와 tf-idf 가중치, 형태소 분석은 모두 하는 것이 성능 향상에 도움이 된다는 결과를 도출해냈다. 또한 uni-gram만 사용했을 때보다 bi-gram을 함께 사용했을 때 더 높은 성능이 나온다. 위 실험에서는 tri-gram까지 함께 사용했을 경우 성능이 가장 높게 계산되었으나, tri-gram까지 함께 사용할 때 길이가 긴 문서를 처리할 경우 피쳐 수가 너무 많아져 작동 시간이 길어지는 등의 효율성 문제가 생길 수 있을 것으로 예상된다. 또한 tri-gram도 함께 사용했을 때와 그렇지 않을 때의 성능 차이가 0.22%p로 근소하므로 모델 14를 선택하고자 한다.

**4. 또다른 실험 계획**

모델 13, 14, 15에 이어 TF-IDF, 형태소 분석 등의 피쳐가 동일하게 유지된 상황에서, n-gram의 피쳐를 더 추가해 실험했을 때 성능이 어떻게 변화하는 지 실험해보고 싶다. 위 실험에서는 tri-gram까지 사용했을 때 성능이 향상되었으나, 어디까지 사용했을 때 성능이 저하되는지, 또는 계속 성능이 향상될 것인지 기록하고 싶다.