n-gram extraction

Hyunjoong Kim

soy.lovit@gmail.com

github.com/lovit

unigram

```
tokenize('라라랜드는 재미있는 영화입니다')
```

```
라라랜드, 는, 재미, 있는, 영화, 입니다
```

- 독립된 하나의 단어를 unigram이라 합니다.
- 위 문장은 6개의 단어로 이뤄져 있습니다.

bigram

```
tokenize('라라랜드는 재미있는 영화입니다')
```

```
라라랜드, 는, 재미, 있는, 영화, 입니다
```

- bigram은 두 개의 단어 조합을 하나의 단어로 취급합니다
  - 연결 부분을 표현하기 위하여 '-'을 이용합니다.
    - "재미 – 있는"
  - 두 단어는 반드시 연속될 필요는 없습니다
    - "재미 – 영화 "

bigram

```
tokenize('라라랜드는 재미있는 영화입니다')
```

```
라라랜드, 는, [재미, 있는], 영화, 입니다
```

- bigram은 문맥의 표현력이 좋습니다
  - '재미'라는 단어 만으로는 이 문장의 긍/부정을 알기 어렵습니다
  - '있는' 만으로는 어떤 의미인지 알기 어렵습니다
  - '재미 – 있는'은 긍정적인 문맥을 표현합니다.

bigram

- document classification은 bigram + linear model 이면 충분합니다
  - 많은 연구들에서도 sentiment/category classification에서는 bigram features 이면 logistic regression과 같은 모델이어도 분류가 잘된다 알려졌습니다 [1,2]
  - 하지만 unigram 보다 bigram을 이용하는 것은 큰 도움이 됩니다

[1] Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics
[2] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

n-gram

```
tokenize('라라랜드는 재미있는 영화입니다')
```

```
라라랜드, 는, [재미, 있는, 영화], 입니다
```

- n-gram은 세 개 이상의 단어 조합을 하나의 단어로 취급합니다
  - Phrase 혹은 "바람 – 의 – 나라"와 같은 단어가 되기도 합니다

n-gram

- n-gram의 추출 방법은 다양합니다
  - 가장 좋은 것은 없습니다
  - 하지만, 계산 과정에서 많은 메모리가 필요할 수 있습니다

n-gram

- By counting
  - 가장 간단한 방법은 모든 n-gram에 대하여 빈도수를 계산하는 것입니다
  - ‘Josa + Verb + Noun’과 같은 형태의 ngram이 추출될 수도 있습니다

```
[(('열린/Verb', '영화/Noun'), 476),
 (('에서/Noun', '열린/Verb', '영화/Noun'), 288),
 (('코미디/Noun', '영화/Noun'), 204),
 (('국제/Noun', '영화제/Noun'), 200),
 (('에서/Josa', '열린/Verb', '영화/Noun'), 185)]
```

< 마지막 단어가 영화인 bi/trigram >

```
[(('재/Noun', '배포/Noun', '금지/Noun'), 20436),
 (('및/Noun', '재/Noun', '배포/Noun'), 14687),
 (('전재/Noun', '및/Noun', '재/Noun'), 14340),
 (('무단/Noun', '전재/Noun', '및/Noun'), 14340),
 (('무단/Noun', '전재/Noun', '재/Noun'), 5178)]
```

< 뉴스기사의 빈도수 기준 상위 5개의 trigram >

n-gram

- By Point Mutual Information (PMI) – like
  - Mikolov는 PMI를 조금 바꾼 간단한 bigram 점수를 만들었습니다[1]

$$score(w_i, w_j) = \frac{count(w_i, w_j) - \delta}{count(w_i) \times count(w_j)}$$

('허심/Noun', '탄회/Noun'), ('무라카미/Noun', '하루키/Noun'), ('로웰/Noun', '패독/Noun')

...

('가습기/Noun', '살균제/Noun'), ('자유로이/Adverb', '접근할/Verb'), ('새판/Noun', '짜기/Verb')

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems

n-gram

- Extending Point Mutual Information (PMI)
  - PMI는 2개의 items에 대하여 정의되어 있습니다
  - n 개의 items에 대한 확장방법은 다양하며, 절대적인 정답은 없습니다
  - bigram이 더 정확한 문맥을 나타내므로, 이를 이용하여 PMI를 확장합니다

$$score(w_i, w_j, w_k) = \frac{count(w_i, w_j, w_k) - \delta}{count(w_i, w_j) \times count(w_j, w_k)}$$

Korean n-gram

```
tokenize('라라랜드는 재미있는 영화입니다')
```

```
라라랜드, 는, 재미, 있는, 영화, 입니다
```

- 한국어에서 의미있는 n-gram은 품사 정보를 이용하는 것이 좋습니다
  - "있는 – 영화"는 유의미한 n-gram이 아닙니다
  - 첫 단어가 명사이거나, 조사/어미가 아닌 n-gram을 선택할 수 있습니다
    - "재미 – 있는"

Korean n-gram

```
tokenize('라라랜드는 재미있는 영화입니다')
```

```
라라랜드/명사, 는/조사, 재미/명사, 있는/형용사, 영화/명사, 입니다/형용사
```

- 의미있는 n-gram을 추출하기 위하여 templates을 이용해도 좋습니다
  - (명사, 조사) / (명사, 명사) 처럼 미리 정의한 templates에 품사가 매칭되는 n-grams 을 추출할 수 있습니다

Korean n-gram

```
tokenize('라라랜드가 개봉 했습니다')
```

```
라라랜드/명사, 가/조사, 개봉/명사, 했습니다/동사
```

- 조사/어미를 skip 하는 templates도 유용합니다
  - (명사, [조사], 명사)를 이용하면

Korean n-gram

- 조사/어미를 skip 하는 templates도 유용합니다
  - (명사, [조사], 명사)를 이용하면 아래 두 문장에서 모두 "라라랜드 – 개봉" 을 추출할 수 있습니다

라라랜드/명사, 가/조사, 개봉/명사, 했습니다/동사

진짜/부사, ?/기호, 라라랜드/명사, 개봉/명사, 했어/동사, ?/기호