

AI in Facebook

- FastText

김동길 김혁민 이우성 최재건



Contents



1

초/중/종성 분리

2

코드 실행 및 구현

FastText

Facebook AI팀의 자연어처리 프로젝트

The logo for FastText, featuring the word 'fast' in a red, italicized sans-serif font, followed by 'Text' in a blue, bold sans-serif font.

Library for efficient text classification and representation learning

모든 단어의 각 n -gram에 대해서 워드 임베딩을 하는 인공 신경망

FastText

Fasttext in Google colab

Word Representation.ipynb ☆

파일 수정 보기 삽입 런타임 도구 도움말 모든 변경사항이 저장됨

+ 코드 + 텍스트

RAM 디스크 수정 가능

```
!pip install fasttext
```

```
Collecting fasttext
  Downloading https://files.pythonhosted.org/packages/10/61/2e01f1397ec533756c1d893c22d9d5ed3fce3a6e4af1976e0d86bb13ea97/fasttext-0.9.1.tar.gz (57kB)
    |#####| 61kB 2.0MB/s
Requirement already satisfied: pybind11>=2.2 in /usr/local/lib/python3.6/dist-packages (from fasttext) (2.4.3)
Requirement already satisfied: setuptools>=0.7.0 in /usr/local/lib/python3.6/dist-packages (from fasttext) (41.6.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.6/dist-packages (from fasttext) (1.17.4)
Building wheels for collected packages: fasttext
  Building wheel for fasttext (setup.py) ... done
  Created wheel for fasttext: filename=fasttext-0.9.1-cp36-cp36m-linux_x86_64.whl size=2387870 sha256=1eae2b123b9c616a9ac40e4fae5805713c2e4c2c400a534c97d8d9
  Stored in directory: /root/.cache/pip/wheels/9f/f0/04/caa82c912aee89ce76358ff954f3f0729b7577c8ff23a292e3
Successfully built fasttext
Installing collected packages: fasttext
Successfully installed fasttext-0.9.1
```

```
[2] !git clone https://github.com/facebookresearch/fastText.git
```

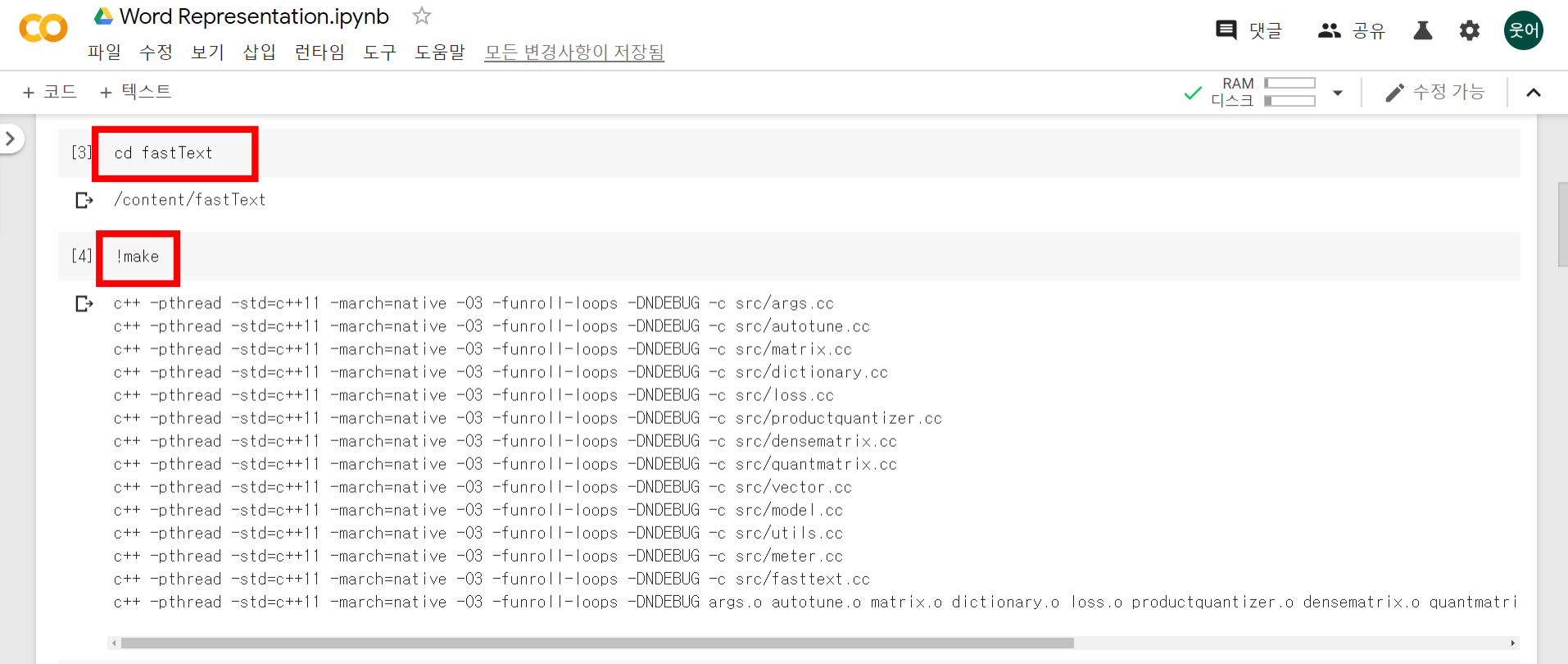
```
Cloning into 'fastText'...
remote: Enumerating objects: 3531, done.
remote: Total 3531 (delta 0), reused 0 (delta 0), pack-reused 3531
Receiving objects: 100% (3531/3531), 8.02 MiB | 5.41 MiB/s, done.
Resolving deltas: 100% (2225/2225), done.
```

1. pip install을 통해 fasttext 패키지를 다운

2. git clone을 통해 fastText 저장소를 colab환경에 가져오기

FastText

Fasttext in Google colab



Word Representation.ipynb ☆

파일 수정 보기 삽입 런타임 도구 도움말 모든 변경사항이 저장됨

+ 코드 + 텍스트

RAM 디스크

수정 가능

```
[3] cd fastText
```

```
[4] make
```

```
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/args.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/autotune.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/matrix.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/dictionary.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/loss.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/productquantizer.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/densematrix.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/quantmatrix.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/vector.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/model.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/utils.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/meter.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG -c src/fasttext.cc
c++ -pthread -std=c++11 -march=native -O3 -funroll-loops -DNDEBUG args.o autotune.o matrix.o dictionary.o loss.o productquantizer.o densematrix.o quantmatr
```

3. colab으로 옮겨온 fastText 저장소로 디렉토리 이동

4. Make를 통해 colab환경에서 파일 관리 유틸리티 설정

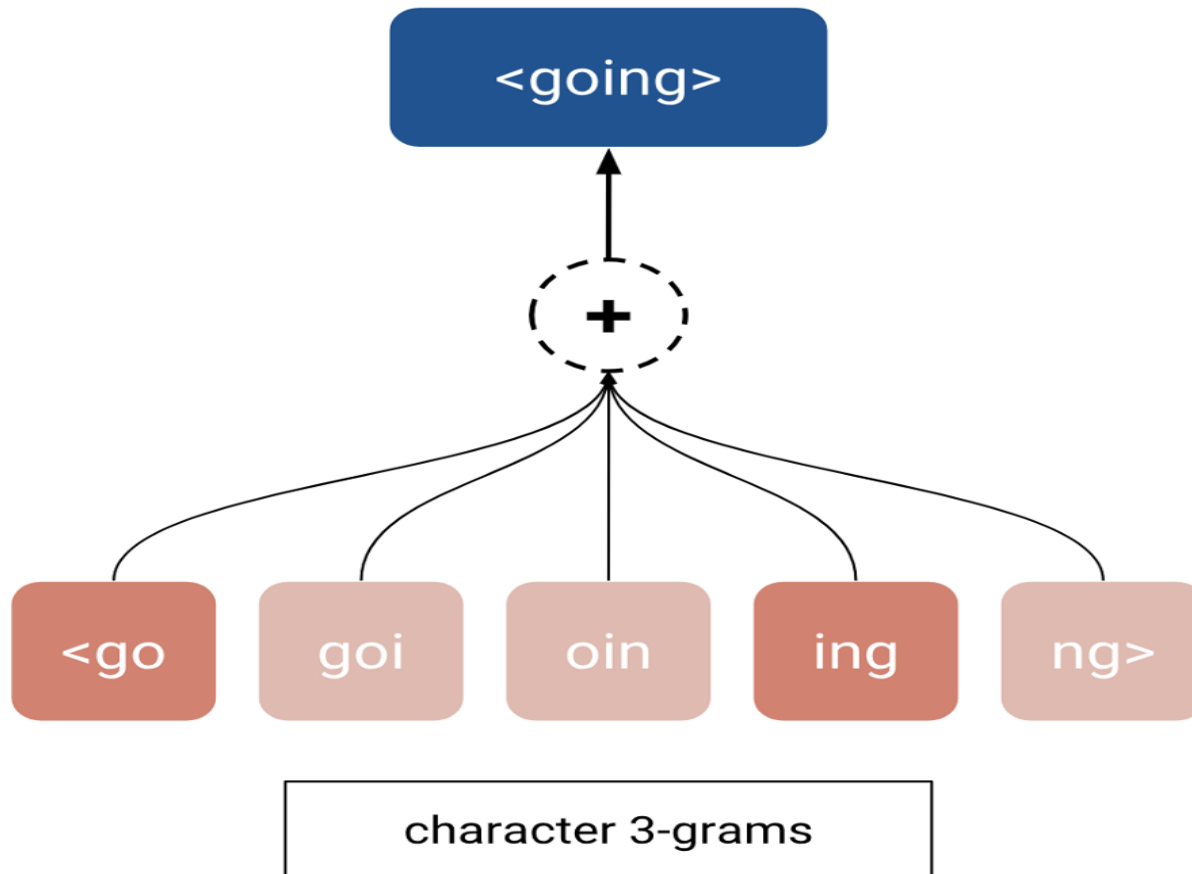
1

초 / 중 / 종성 분리



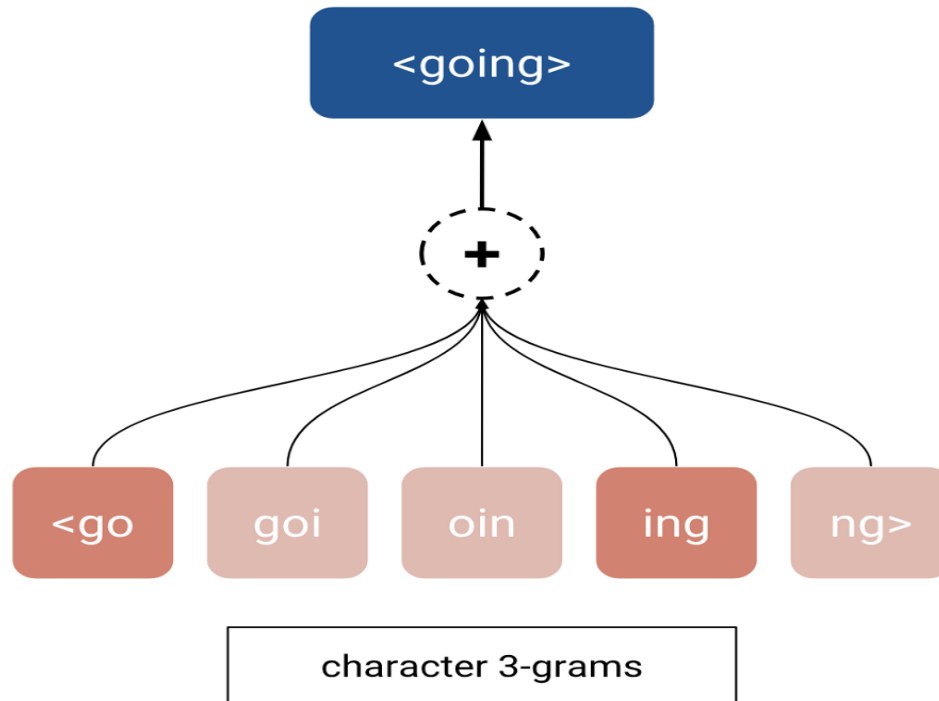
Word Embedding

단어를 밀집 표현으로 변환하는 방법



Word Embedding

오타자에 민감한 한국어



어디야 학습 -> 어디 / 디야

어딘야 학습 -> 언디 / 딘야

→ 어디야와 어딘야를 비슷한 단어로 묶어주지 못한다!

2

코드 구현 및 성능 비교



Word Embedding for Korean

한국어를 위한 FastText

```

from sys import stdin
import re

double_space_pattern = re.compile('##s+')

def jamo_sentence(sent):

    def transform(char):
        if char == ' ':
            return char
        cjj = decompose(char)
        if len(cjj) == 1:
            return cjj
        cjj_ = ''.join(c if c != ' ' else '-' for c in cjj)
        return cjj_

    sent_ = ''.join(transform(char) for char in sent)
    sent_ = double_space_pattern.sub(' ', sent_)
    return sent_

print(jamo_sentence('어이고ㅋㅋ큅큅아이고오'))

print(jamo_sentence('한글분리'))

print(jamo_sentence('이재우교수님'))

```

한글의 자모를 분리함

Word Embedding for Korean

한국어를 위한 FastText

```
[69] !./fasttext skipgram -input ratings_train.txt -output ratings_train.model
```

```
!./fasttext nn ratings_train.model.bin
```

```
... Query word? 영화
영화♥ 0.814993
영화ㅠ 0.808703
영화- 0.805596
영화ㅋ 0.805174
영화ㅎ 0.804743
영화군 0.804033
영화긴 0.799356
영화과 0.798061
영화' 0.79769
영화췌 0.797546
Query word? 봉준호
마이클베이의 0.935089
다케시 0.933121
임상수 0.927648
더스틴 0.925874
조선의 0.923667
스콜세지의 0.923354
놀란의 0.923269
저수지의 0.922861
올바른 0.922245
천재다 0.92224
```

```
Query word? 
```

원본 데이터 학습 및 유사도 분석

Word Embedding for Korean

한국어를 위한 FastText

```
[56] !./fasttext skipgram -input ratings_soynlp.txt -output ratings_soynlp.model
```

```
Read 2M words  
Number of words: 20274  
Number of labels: 0  
Progress: 100.0% words/sec/thread: 22116 lr: 0.000000 loss: 2.240380 ETA: 0h 0m
```

```
!./fasttext nn ratings_soynlp.model.bin
```

```
... Query word? 영화  
그영화 0.811201  
.영화 0.78649  
할영화 0.769467  
컨보영화 0.762247  
쿼어영화 0.755589  
태국영화 0.754246  
한영화 0.74841  
일세 0.747367  
C급영화 0.746042  
멜로영화 0.738266  
Query word? 봉준호  
이산리투 0.835542  
현들 0.834433  
이창동 0.832838  
박철수 0.827747  
임권택 0.822618  
멜 0.820413  
장진의 0.819861  
커리어에 0.819625  
상발은 0.818349  
자질 0.818149  
Query word? 문준호  
정준호 0.917452  
테니스 0.871294  
피디가 0.864308  
갑. 0.861759  
자존심도 0.859886  
어쓰케 0.859068  
봉준호 0.856369  
김준호 0.854757  
동호 0.848227  
으이구 0.845584
```

```
Query word? 
```

형태소 분석 데이터 학습 및 유사도 분석

Word Embedding for Korean

한국어를 위한 FastText

```
!./fasttext nn alignment/word-embeddings/fasttext-jamo/ fasttext-jamo.bin
```

```
... Query word? ㅇㅋㅇㅎㅏ  
ㅇㅋㅇㅎㅏㄹ 0.847467  
ㅇㅋㅇㅎㅏ-ㄱㅑ 0.807159  
ㅇㅋㅇㅎㅏㄴ 0.793214  
ㅇㅋㅇㅎㅏ-ㅇㅋㅇ 0.787161  
ㅇㅋㅇㅎㅏ- 0.785971  
ㅇㅋㅇㅎㅏ-ㄱ-ㄱ 0.774965  
ㅂㅏㅑㅑㅑㅑㅑㅑㅑ 0.774352  
ㄷㅑㅑㄹ | ㅂㅇㅋㅇㅎㅏ- 0.772896  
ㅇㅋㅇㅎㅏ-ㄱㅏㅇ 0.769094  
ㅇㅋㅇㅎㅏㄴ ㄷㅑ- 0.751629  
Query word? ㅂㅏㅇㅑㅑㅑㅑㅑ  
ㅂㅏㅇㅑㅑㅑㅑㅑㅑ- 0.958443  
ㄱㅏㅇㅑㅑㅑㅑㅑㅑ- 0.788131  
ㄱㅏㅑㅑㅑㅑㅑㅑㅑㅑ 0.787735  
ㅂㅏㅑㅑㅑㅑㅑㅑㅑㅑ 0.777712  
ㅑㅑㅑㅑㅑㅑㅑㅑㅑㅑ 0.773031  
ㅑㅑㅑ-ㅑ | ㅑㅑㅑ | ㅑ 0.771743  
ㄱ | ㅑㅑ | -ㅑㅑㅑ 0.766479  
ㅑㅑ | ㅇㅑㅑㅑㅑㅑㅑㅑ 0.766007  
ㅑㅑㅇㅑㅑㅑㅑㅑㅑㅑ 0.765264  
ㅎㅑㅇㅑㅑㅑㅑㅑㅑㅑ 0.764595  
Query word? ㅂㅏㅇㅇㅑㅑㅑㅑㅑㅑ  
ㅇㅑㅇㅑㅑㅑㅑㅑㅑㅑㅑ 0.837831  
ㅑㅑㅇㅑㅑㅑㅑㅑㅑㅑㅑ 0.837329  
ㄱㅑㅇㅑㅑㅑㅑㅑㅑㅑㅑ 0.832619  
ㅇ | -ㅇ-ㅑㅑㅑㅑㅑㅑㅑ 0.823627  
ㄱ | ㅑㅑ | -ㅑㅑㅑ 0.821812  
ㄱㅑㅇㅑㅑㅑㅑㅑㅑㅑㅑ 0.816032  
ㅂㅑㅇㅑㅑㅑㅑㅑㅑㅑㅑ 0.815988  
ㄱ | ㅑㅑㅑㅑ-ㅎㅑㅑㅑ 0.815709  
ㅎㅑㅑ-ㅑㅑㅑㅑㅑㅑㅑㅑ 0.814419  
ㅎㅑㅇㅑㅑㅑㅑㅑㅑㅑㅑ 0.812532  
Query word? 
```

자모 분리 데이터 학습 및 유사도 분석

Reference

1. <https://fasttext.cc/docs/en/supervised-tutorial.html>
2. FastText를 적용한 한국어 단어 임베딩(서울대학교)
3. <https://ratsgo.github.io/embedding>

감사합니다!

