Srilm N-gram

Srilm 编译

export SRILM=/home/nghon/srilm
export PATH=$PATH:$SRILM/bin:$SRILM/bin/i686-m64

sudo apt install make

sudo apt update # 更新包列表
sudo apt install gcc # 安装 GCC

sudo apt update # 更新包列表
sudo apt install g++

make

数据准备

数据合并

```
cat data/news.train data/news.2007.en.shuffled > data/combined.train
```

文本清洗 - 全部小写，去除标点符号

stopwords - All English Stopwords (700+) | Kaggle

```
awk 'BEGIN{IGNORECASE=1} NR==FNR{stopwords[tolower($0)]; next} {gsub(/[[:punct:]]/, ""); for (i=1; i<=NF; i++) {word=tolower($i); if (!(wor
```

文本分割

```
perl -C -lne 'print join(" ", split(""))' data/combined.train > data/combined.train.split
```

Word-Based

无数据预处理

```
bin/i686-m64/ngram-count -order 3 -text data/news.train -lm data/models/news_word.lm

bin/i686-m64/ngram -lm data/models/news_word.lm -order 1 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/news_word.lm -order 1 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/news_word.lm -order 2 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/news_word.lm -order 2 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/news_word.lm -order 3 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/news_word.lm -order 3 -ppl data/news.test
```

```
bin/i686-m64/ngram -lm data/models/news_word.lm -order 1 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/news_word.lm -order 1 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/news_word.lm -order 2 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/news_word.lm -order 2 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/news_word.lm -order 3 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/news_word.lm -order 3 -ppl data/news.test
warning: discount coeff 1 is out of range: 0
file data/news.train: 300000 sentences, 6571469 words, 0 OOVs
0 zeroprobs, logprob= -2.319247e+07 ppl= 2372.376 ppl1= 3382.728
file data/news.test: 91233 sentences, 1970648 words, 34583 OOVs
0 zeroprobs, logprob= -6674795 ppl= 1960.916 ppl1= 2802.91
file data/news.train: 300000 sentences, 6571469 words, 0 OOVs
0 zeroprobs, logprob= -1.496623e+07 ppl= 150.6693 ppl1= 189.433
file data/news.test: 91233 sentences, 1970648 words, 34583 OOVs
0 zeroprobs, logprob= -5228048 ppl= 379.1628 ppl1= 501.5884
file data/news.train: 300000 sentences, 6571469 words, 0 OOVs
0 zeroprobs, logprob= -1.393731e+07 ppl= 106.7302 ppl1= 132.0939
file data/news.test: 91233 sentences, 1970648 words, 34583 OOVs
0 zeroprobs, logprob= -5080458 ppl= 320.645 ppl1= 420.8386
```

数据预处理加平滑

```
awk 'BEGIN{IGNORECASE=1} NR==FNR{stopwords[tolower($0)]; next} {gsub(/[[:punct:]]/, ""); for (i=1; i<=NF; i++) {word=tolower($i); if (!(wor
```

```
bin/i686-m64/ngram-count -text data/processed_news.train -order 3 -lm data/models/tuned_news_word.lm -kndiscount -interpolate -gt3min 1 -gt

bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 1 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 1 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 2 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 2 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 3 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 3 -ppl data/news.test
```

```
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 1 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 1 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 2 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 2 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 3 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_news_word.lm -order 3 -ppl data/news.test
data/models/tuned_news_word.lm: line 3638: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train: 300000 sentences, 6571469 words, 0 OOVs
0 zeroprobs, logprob= -1.837309e+07 ppl= 471.8699 ppl1= 625.0105
data/models/tuned_news_word.lm: line 3638: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test: 91233 sentences, 1970648 words, 0 OOVs
0 zeroprobs, logprob= -5471638 ppl= 450.5177 ppl1= 597.8165
data/models/tuned_news_word.lm: line 3638: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train: 300000 sentences, 6571469 words, 0 OOVs
0 zeroprobs, logprob= -1.541459e+07 ppl= 175.0952 ppl1= 221.6583
data/models/tuned_news_word.lm: line 3638: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test: 91233 sentences, 1970648 words, 0 OOVs
0 zeroprobs, logprob= -4861567 ppl= 227.9456 ppl1= 293.082
data/models/tuned_news_word.lm: line 3638: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train: 300000 sentences, 6571469 words, 0 OOVs
0 zeroprobs, logprob= -1.315496e+07 ppl= 82.11689 ppl1= 100.4223
data/models/tuned_news_word.lm: line 3638: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test: 91233 sentences, 1970648 words, 0 OOVs
0 zeroprobs, logprob= -4722563 ppl= 195.1707 ppl1= 249.1447
```

Character-Based

数据预处理加平滑

```
perl -C -lne 'print join(" ", split(""))' data/processed_news.train > data/processed_news.train.split
perl -C -lne 'print join(" ", split(""))' data/news.train > data/news.train.split
perl -C -lne 'print join(" ", split(""))' data/news.test > data/news.test.split
```

```
bin/i686-m64/ngram-count -text data/processed_news.train.split -order 6 -lm data/models/tuned_news_char.lm -kndiscount
```

```
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 1 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 1 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 2 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 2 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 3 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 3 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 4 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 4 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 5 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 5 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 6 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 6 -ppl data/news.test.split
```

```
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 1 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 2 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 2 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 3 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 3 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 4 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 4 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 5 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 5 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 6 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_news_char.lm -order 6 -ppl data/news.test.split
file data/news.train.split: 300000 sentences, 34824611 words, 2049043 OOVs
0 zeroprobs, logprob= -6.523275e+07 ppl= 93.80668 ppl1= 97.78808
file data/news.test.split: 91233 sentences, 10414486 words, 607492 OOVs
0 zeroprobs, logprob= -1.951596e+07 ppl= 93.68335 ppl1= 97.72473
file data/news.train.split: 300000 sentences, 34824611 words, 2049043 OOVs
0 zeroprobs, logprob= -4.749321e+07 ppl= 27.28351 ppl1= 28.12181
file data/news.test.split: 91233 sentences, 10414486 words, 607492 OOVs
0 zeroprobs, logprob= -1.420395e+07 ppl= 27.22696 ppl1= 28.07688
file data/news.train.split: 300000 sentences, 34824611 words, 2049043 OOVs
0 zeroprobs, logprob= -4.414029e+07 ppl= 21.60372 ppl1= 22.21999
file data/news.test.split: 91233 sentences, 10414486 words, 607492 OOVs
0 zeroprobs, logprob= -1.321038e+07 ppl= 21.60832 ppl1= 22.23498
file data/news.train.split: 300000 sentences, 34824611 words, 2049043 OOVs
0 zeroprobs, logprob= -4.000245e+07 ppl= 16.19668 ppl1= 16.61484
file data/news.test.split: 91233 sentences, 10414486 words, 607492 OOVs
0 zeroprobs, logprob= -1.197597e+07 ppl= 16.21477 ppl1= 16.6405
file data/news.train.split: 300000 sentences, 34824611 words, 2049043 OOVs
0 zeroprobs, logprob= -3.62367e+07 ppl= 12.46158 ppl1= 12.75267
file data/news.test.split: 91233 sentences, 10414486 words, 607492 OOVs
0 zeroprobs, logprob= -1.086119e+07 ppl= 12.51083 ppl1= 12.80837
file data/news.train.split: 300000 sentences, 34824611 words, 2049043 OOVs
0 zeroprobs, logprob= -2.353029e+07 ppl= 5.145293 ppl1= 5.22302
file data/news.test.split: 91233 sentences, 10414486 words, 607492 OOVs
0 zeroprobs, logprob= -7137629 ppl= 5.261406 ppl1= 5.343307
```

Combined-Word-Based

```
cat data/news.train data/news.2007.en.shuffled > data/combined.train
```

```
awk 'BEGIN{IGNORECASE=1} NR==FNR{stopwords[tolower($0)]; next} {gsub(/[[:punct:]]/, ""); for (i=1; i<=NF; i++) {word=tolower($i); if (!(wor
```

```
bin/i686-m64/ngram-count -text data/processed_combined.train -order 3 -lm data/models/tuned_combined_word.lm -kndiscount -interpolate -gt3m

bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 1 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 1 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 2 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 2 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 3 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 3 -ppl data/news.test
```

```
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 1 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 1 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 2 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 2 -ppl data/news.test
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 3 -ppl data/news.train
bin/i686-m64/ngram -lm data/models/tuned_combined_word.lm -order 3 -ppl data/news.test
data/models/tuned_combined_word.lm: line 110027: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train: 300000 sentences, 6571469 words, 0 OOVs
0 zeroprobs, logprob= -1.943954e+07 ppl= 674.5632 ppl1= 908.1821
data/models/tuned_combined_word.lm: line 110027: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test: 91233 sentences, 1970648 words, 0 OOVs
0 zeroprobs, logprob= -5778060 ppl= 634.3441 ppl1= 855.1872
data/models/tuned_combined_word.lm: line 110027: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train: 300000 sentences, 6571469 words, 0 OOVs
0 zeroprobs, logprob= -1.664649e+07 ppl= 264.5767 ppl1= 341.3074
data/models/tuned_combined_word.lm: line 110027: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test: 91233 sentences, 1970648 words, 0 OOVs
0 zeroprobs, logprob= -5118313 ppl= 303.6348 ppl1= 395.6164
data/models/tuned_combined_word.lm: line 110027: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train: 300000 sentences, 6571469 words, 0 OOVs
0 zeroprobs, logprob= -1.458372e+07 ppl= 132.5434 ppl1= 165.6715
data/models/tuned_combined_word.lm: line 110027: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test: 91233 sentences, 1970648 words, 0 OOVs
0 zeroprobs, logprob= -4953115 ppl= 252.4826 ppl1= 326.1707
```

Combined-Character-Based

```
perl -C -lne 'print join(" ", split(""))' data/combined.train > data/combined.train.split
```

```
bin/i686-m64/ngram-count -text data/combined.train.split -order 6 -lm data/models/tuned_combined_char.lm -kndiscount -interpolate -gt3min 1
```

```
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 1 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 1 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 2 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 2 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 3 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 3 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 4 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 4 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 5 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 5 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 6 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 6 -ppl data/news.test.split
```

```
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 1 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 2 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 2 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 3 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 3 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 4 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 4 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 5 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 5 -ppl data/news.test.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 6 -ppl data/news.train.split
bin/i686-m64/ngram -lm data/models/tuned_combined_char.lm -order 6 -ppl data/news.test.split
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train.split: 300000 sentences, 34824611 words, 0 OOVs
0 zeroprobs, logprob= -7.329204e+07 ppl= 122.0757 ppl1= 127.2345
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test.split: 91233 sentences, 10414486 words, 0 OOVs
0 zeroprobs, logprob= -2.186876e+07 ppl= 120.6716 ppl1= 125.8462
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train.split: 300000 sentences, 34824611 words, 0 OOVs
0 zeroprobs, logprob= -5.665778e+07 ppl= 41.0252 ppl1= 42.35907
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test.split: 91233 sentences, 10414486 words, 0 OOVs
0 zeroprobs, logprob= -1.695904e+07 ppl= 41.14033 ppl1= 42.50198
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train.split: 300000 sentences, 34824611 words, 0 OOVs
0 zeroprobs, logprob= -5.026544e+07 ppl= 26.98115 ppl1= 27.75802
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test.split: 91233 sentences, 10414486 words, 0 OOVs
0 zeroprobs, logprob= -1.506857e+07 ppl= 27.18444 ppl1= 27.98242
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train.split: 300000 sentences, 34824611 words, 0 OOVs
0 zeroprobs, logprob= -4.317301e+07 ppl= 16.94878 ppl1= 17.36708
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test.split: 91233 sentences, 10414486 words, 0 OOVs
0 zeroprobs, logprob= -1.2975e+07 ppl= 17.18073 ppl1= 17.61411
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train.split: 300000 sentences, 34824611 words, 0 OOVs
0 zeroprobs, logprob= -3.639814e+07 ppl= 10.8707 ppl1= 11.09646
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test.split: 91233 sentences, 10414486 words, 0 OOVs
0 zeroprobs, logprob= -1.100228e+07 ppl= 11.14977 ppl1= 11.38781
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.train.split: 300000 sentences, 34824611 words, 0 OOVs
0 zeroprobs, logprob= -2.085146e+07 ppl= 3.923217 ppl1= 3.969687
data/models/tuned_combined_char.lm: line 41: warning: non-zero probability for <unk> in closed-vocabulary LM
file data/news.test.split: 91233 sentences, 10414486 words, 0 OOVs
0 zeroprobs, logprob= -6469244 ppl= 4.128412 ppl1= 4.180011
```