Preprocessing

11.09.2020

Фитц С.Ю.

Векторизация

Процесс превращения текста на естественном языке в числовой вектор.

Rule-based
Обучение

Pipline

- Нормализация (tokenization, stemming, lemmatization)
- Чистка (stop-words, punctuation)
- N-grams
- Векторизация (BOW, tf-idf)
- Уменьшение размерности (hash)

Предобработка текстов

1. Токенизация(tokenization)
2. Стеммизация(stemming)
3. Лемматизация(lemmatization)
4. Чистка

Tokenization

What is text?

Токенизация - разделение текста на токены.

- Буквы
- Слова
- Фразы
- Параграфы
- etc...

Tokenization

sentence:

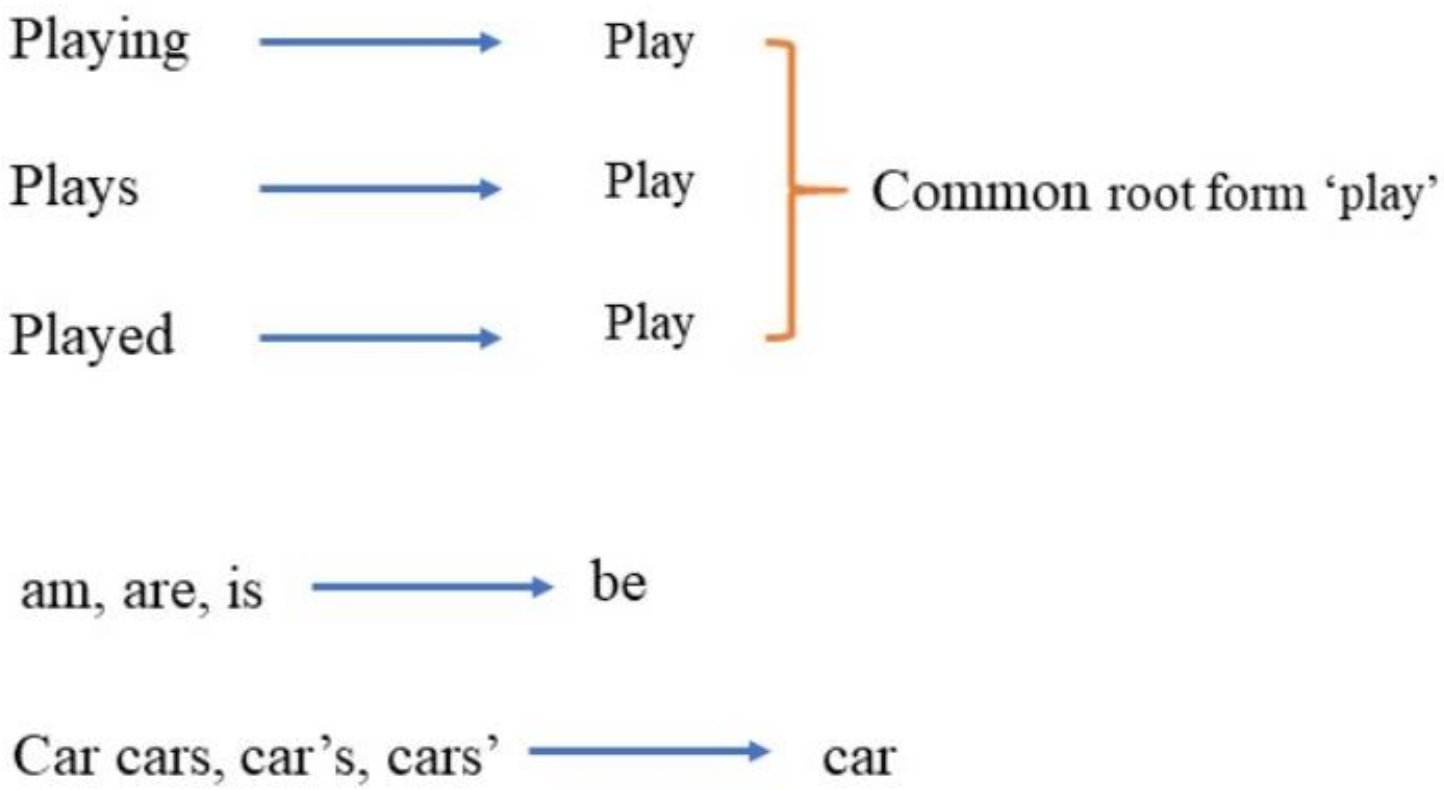
```
After sleeping for four hours, he decided to sleep for another four
```

In this case, the tokens are as follows:

```
{'After', 'sleeping', 'for', 'four', 'hours', 'he', 'decided', 'to', 'sleep', 'for', 'another', 'four'}
```

Stemming

приведение слова к его корню
путем устранения придатков
(суффикса, приставки, окончания).



Stemming

```
Step 1a

    SSES -> SS                      caresses   ->  caress
    IES  -> I                       ponies     ->  poni
                                    ties       ->  ti
    SS   -> SS                      caress     ->  caress
    S    ->                         cats       ->  cat

Step 1b

    (m>0) EED -> EE                 feed       ->  feed
                                    agreed     ->  agree
    (*v*) ED  ->                    plastered  ->  plaster
                                    bled       ->  bled
    (*v*) ING ->                    motoring   ->  motor
                                    sing       ->  sing
```

Stemming

```
AT -> ATE
BL -> BLE
IZ -> IZE
(*d and not (*L or *S or *Z))
   -> single letter

(m=1 and *o) -> E
```

```
conflat(ed)   ->  conflate
troubl(ed)    ->  trouble
siz(ed)       ->  size

hopp(ing)     ->  hop
tann(ed)      ->  tan
fall(ing)     ->  fall
hiss(ing)     ->  hiss
fizz(ed)      ->  fizz
fail(ing)     ->  fail
fil(ing)      ->  file
```

An algorithm for suffix stripping

Lemmatization

приведение слова к смысловой канонической форме слова (инфинитив для глагола, именительный падеж единственного числа — для существительных и прилагательных).

зарезервированный — резервировать

грибами — гриб

лучший — хороший

Lemmatization vs Stemming

Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study
niñas	Feminine gender, plural number of the noun niño	niño
niñez	Singular number of the noun niñez	niñez

Cleaning

- Артикли
- Междометья
- Союзы
- Предлоги
- И т.д.

Bag-of-words

Steps:

- Make a dictionary.
- Create zero vectors.
- Fill with values corresponding the words.
- Scoring words

Bag-of-words

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

TF-IDF

Term Frequency: is a scoring of the frequency of the word in the current document.

Inverse Document Frequency: is a scoring of how rare the word is across documents.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \qquad \text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

N-grams

Example: «It was the best of times»

it was

was the

the best

best of

of times

Хэшинг

Хэшинг - алгоритм представления строк в виде числа.

- есть разные имплементации
- можно варьировать длину и размер массива
- хорошо подходит для уменьшения размера больших корпусов
- замена словарю