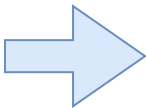


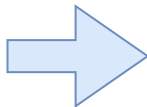
'The résultat:  
123 people,  
123456 CATS  
and 123.45  
pizzas!\n'

**PDFs**



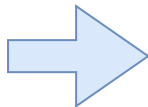
'The résultat:  
123 people,  
123456 CATS  
and 123.45  
pizzas!\n'

**Strip accents**



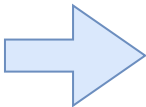
'The résultat:  
123 people,  
123456 CATS  
and 123.45  
pizzas! '

**Strip newlines**



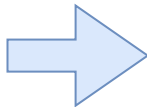
'the resultat:  
123 people,  
123456 cats  
and 123.45  
pizzas! '

**Lowercase**



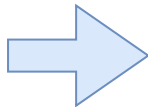
'the resultat:  
SMALLNUMBER  
people,  
BIGNUMBER  
cats and FLOAT  
pizzas! '

**Discretize numbers**



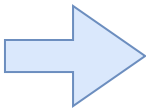
'the resultat  
SMALLNUMBER  
people  
BIGNUMBER  
cats and FLOAT  
pizzas '

**Remove punctuation**



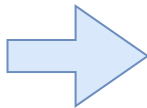
'the resultat  
SMALLNUMBER  
people  
BIGNUMBER  
cats and FLOAT  
pizzas '

**Remove subsequent  
whitespaces**



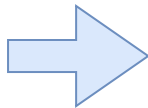
'the resultat  
<SMALLNUMBER>  
people  
<BIGNUMBER>  
cats and  
<FLOAT>  
pizzas '

**Change number  
encodings**



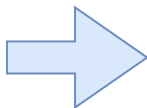
'resultat',  
<SMALLNUMBER>,  
'people',  
<BIGNUMBER>,  
'cats', <FLOAT> ,  
'pizzas'

**Remove stop words**



'resultat',  
<SMALLNUMBER>,  
'people',  
<BIGNUMBER>,  
'cat', <FLOAT> ,  
'pizza'

**Lemmatization**



'resultat  
<SMALLNUMBER>  
people  
<BIGNUMBER>  
cat <FLOAT>  
pizza'

**preprocessed  
PDFs**