05 - BoW

April 24, 2020

1 CountVectorizer and BoW

Analysis of Social Media Contents

Alessandro Ortis - University of Catania

Explore the CountVectorizer class to implement BoW text representation.

```
In [1]: from sklearn.feature_extraction.text import CountVectorizer

        corpus = [
        'All my dogs in a row',
        'When my dog sits down, she looks like a Furby toy!',
        'The dog from outer space',
        'Sunshine loves to sit like this for some reason.'
        ]

        vectorizer = CountVectorizer()
        vectorizer.fit(corpus)
```

```
Out[1]: CountVectorizer(analyzer='word', binary=False, decode_error='strict',
                        dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
                        lowercase=True, max_df=1.0, max_features=None, min_df=1,
                        ngram_range=(1, 1), preprocessor=None, stop_words=None,
                        strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
                        tokenizer=None, vocabulary=None)
```

```
In [2]: print(vectorizer.get_feature_names())
```

```
['all', 'dog', 'dogs', 'down', 'for', 'from', 'furby', 'in', 'like', 'looks', 'loves', 'my', '
```

```
In [3]: print(vectorizer.get_stop_words())
```

```
None
```

```
In [4]: sentence = "My dog loves this toy"
        data = [sentence]
        print(vectorizer.vocabulary_)
```

```
        bow_dict = vectorizer.transform(data)
        print(bow_dict)
        bow_feat = bow_dict.toarray()
        print(bow_feat[0])
```

```
{'all': 0, 'my': 11, 'dogs': 2, 'in': 7, 'row': 14, 'when': 25, 'dog': 1, 'sits': 17, 'down': 3
  (0, 1)        1
  (0, 10)        1
  (0, 11)        1
  (0, 22)        1
  (0, 24)        1
[0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0]
```