Bag of words (BOW)

Let's count occurrences of a particular token in our text

- Motivation: we're looking for marker words like "excellent" or "disappointed"
- For each token we will have a feature column, this is called **text vectorization**.

good movie
not a good movie
did not like

good	movie	not	a	did	like
1	**1**	0	0	0	0
1	**1**	**1**	**1**	0	0
0	0	**1**	0	**1**	**1**

- Problems:
  - we loose word order, hence the name "bag of words"
  - counters are not normalized