Векторное представление слов – вектор в пространстве с фиксированной размерностью. На вход подаются коллекция документов, на выходе получаем векторное представление из коллекции документов.

One-hot encoding – метод представления в векторном виде. Создается словарь фиксированного размера *N*, где для каждого слова из словаря соответствует вектор размера *N*, каждый вектор состоит из 1 единицы и (*N – 1*) нулей на -м месте, – номер слова в словаре. Таким образом, запись данного метода представлена на рисунке 1.1.

ТУТ ДОЛЖЕН БЫТЬ РИСУНОК

Рисунок 1.1 – One-hot encoding

Из приведённом описание и примере можно сделать вывод, что данный метод обладает множеством недостатков:

* 1. Размерность векторов могут быть довольно большой;
  2. Слова, не попавшие в словарь, не будут обработаны;
  3. Векторы слов не отражают смысл слов;
  4. Семантическая значимость слов между собой отсутствует.

Bag to Words (BoG) – Мешок слов, представление предложений, основанный на методе one-hot encoding. В данном случае вектором предложений является сумма векторов слов. Или же можно сказать, что мешок слов для предложения – это вектор, у которого длина равна длине словаря и на -м месте стоит такое число сколько раз -е слово из словаря встречается в данном предложении. Запись данного метода можно увидеть на рисунке 1.2.

ТУТ ДОЛЖЕН БЫТЬ РИСУНОК

Рисунок 1.2 – Bag to Words

Следовательно, данный метод наследуется от предыдущего, будет иметь такие же недостатки, но только уже для предложений:

1. Приложения, не попавшие в словарь, не будут обработаны;
2. Векторы предложений не отражают смысл слов;
3. Семантическая значимость предложений между собой отсутствует.

И для того, чтобы перейти к следующему методу, стоит обратить внимание на еще один недостаток – различные слова в любом тексте имеют разную важность/вес для передачи контекста и смысла этого текста. Например, любой текст на английском языке имеет множество артиклей, предлогов, который в свою очередь не несут никакого смыслового контекста, т.к. мы все же сможем понять, о чем идет речь. На рисунке 1.3 красным выделено, какие слова в каждом предложение имеют высокую важность для понимания, зеленным те, которые не имеют высокой важности.

ТУТ ДОЛЖЕН БЫТЬ РИСУНОК

Рисунок 1.3 – Важность слов в предложениях

http://www.machinelearning.ru/wiki/images/b/b3/word2vec.pdf