**IMPLEMENTING BAG OF WORDS (BOW) FROM SCRATCH:**

Question: What the heck is bag of words and why should I read this concept as a Data scientist?

Answer: What if I say data science is all about converting the data into mathematical equation and solving it. And now the question is how do we convert the data into math. It will be very confusing if you not aware about the domain of data science in depth. So in this blog I will convince you guys why BOW is important and how we connecting mathematics logic with the data.

**Let’s start reading-------🡪**

**What is BOW?**

Ans: According to Wikipedia : The **bag-of-words model** is a simplifying representation used in [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) and [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval) (IR). In this model, a text (such as a sentence or a document) is represented as the [bag (multiset)](https://en.wikipedia.org/wiki/Multiset) of its words, disregarding grammar and even word order but keeping [multiplicity](https://en.wikipedia.org/wiki/Multiplicity_(mathematics)). The bag-of-words model has also been [used for computer vision](https://en.wikipedia.org/wiki/Bag-of-words_model_in_computer_vision).

Important note: In order to solve real world problems, we have the data and to use mathematical knowledge on that data we need to convert that data into vectors, now vectors enable us to use mathematical concepts and bag of words is one of the techniques to convert data into vector format. I hope now most of us understood why mathematics is crucial for building machine learning models. If you don’t know math then it’s obvious that you don’t know what’s happening behind the bush and we will learn how to convert data into vector form using Bag of words from scratch.

Before moving ahead we should be clear that this technique is mostly used to convert text data into vectors. Now if we all think in which all domain text data is present then we can find tons and tons of examples.

1. Google news: When you read a news on google news, what you see is texts and as a data scientist your question should that how are we finding out which news is important and coming at top in our search results and yes behind the hood complex mathematical concepts are being used to build these systems.
2. Amazon product reviews: How are we able to which are the top positive comments or top negatice comments on amazon and yes here also maths is being used.
3. Email spam classification: How gmail or any mail is able to distinguish between spam or not spam mails, here we use maths.

Note: We all aware of various machine learning algotithms like KNN,linear regression.logistic regression, SVN and all. Can these algorithms be able to work if give text data or any data directly into these?

No, we need to convert the data text data into vector format so that these algorithms can work.

**Understanding the intuition of Bag of words:**

Example: string = [“I love data science and we are building BOW from scratch”, “I love to do math in my free time”]

**Step 1: Creating the dictionary of all the unique words in our list :**

The following code will help us building the dictionary of all the unique words in our list.

We created function for all the unique words and we will consider this as vocab.

**Step 2: Creating a Custom BOW:**

We will understand it clearly with the following example.

**Comparing our customer implementation with Sklearn countvectorizer:**

**The following code is of Bag of words in Sklearn and we have same results.**

**Question: How this is useful to us?**

**Answer: This technique is used to build spam or not spam classifier and modifications of it is being used.**

**DRAWBACKS OF BAG OF WORDS:**

1. **It don’t consider the sementic meaning of words.**
2. **We need to create vocab very carefully**

**But there many other state of the art techniques which are used in NLP like word2vec and all.**