Amazon ML challenge

**Data Preprocessing**

Dataset has 5 columns: Title | Description | Bullet\_Points | Brand | Browse\_node\_ID

Defects in dataset: Many columns has more than 5 entries, this is due to entries in

columns such as brand and bullet points that have commas in them

Important Points:

1. All rows are of text type
2. The last entry of every row will always be the Browse\_node\_id
3. The second to last will always be the Brand
4. The first entry will always be the beginning of Title
5. Bullet Points always starts and ends with square brackets []
6. All unnecessary commas are preceded by a backward slash, meaning whenever we encounter a backward slash, the term right next to it is part of that column, thus the title can be separated from the description.

