1. **Phone data as the .csv file:**Contains the consumer activity information

* **overall:**  Overall rating for the particular product given by the user
* **verified:** Whether the user is a verified reviewer or not
* **reviewerID:** ID of the reviewer, e.g. A2SUAM1J3GNN3B
* **asin:**  ID of the product, e.g. 0000013714. ASIN stands for Amazon Standard Identification Number. It is a 10-character alphanumeric unique identifier that is assigned by Amazon.com and its partners. It is primarily used for product-identification within their product catalogue
* **style:** Information about physical features like colour
* **reviewerName:** Name of the reviewer
* **reviewText:**Review provided by the reviewer
* **summary:**Summary of the review
* **unixReviewTime:** Time of the review (Unix time)
* **vote:** Upvotes or Downvotes of the review. A review with more upvotes can hold a higher significance
* **image:** Image of type .png or .jpg, etc provided by the reviewer
* **review\_sentiment**: Pre-labelled sentiment of the review which can be either positive or negative. The labels will not be available for the real-time data, and a model needs to be built for such classification

2. **Phone metadata as zipped .json file:**This data contains the product information and is independent of the consumer/reviewer activity and includes description, price, sales-rank, brand info, and co-purchasing links etc.

* **Category:** List of categories the product belongs to, e.g., ['Cell Phones & Accessories', 'Cell Phones']
* **tech1:**  The specifications and technology of the product, e.g., Phablet ( a hybrid of phone and a tablet)
* **tech2:**  Contains data similar to that of tech1
* **description:** A short description of the product
* **fit:** A column which has no values for a mobile phone dataset
* **title:** Name of the product
* **also\_buy:** Product IDs of the products that are generally bought along with the particular product
* **image:** Image of the product by the supplier
* **brand:** Brand of the product
* **feature:** Features of the product including specifications such as the 4G/3G, language, apps supported etc.
* **rank:**Ranking of the product as given by Amazon based on various parameters in different segments. It is a metric that explains the relationship among products within a category based on their sales performance. Sales rank is updated hourly, it can range from 1 to over 1 million, and is influenced by seasonality, and its algorithm remains unrevealed.
* **also\_view:** Products which are also viewed while searching for this product.
* **details:**Contains details from a delivery point of view such as product dimension, shipping weight etc.
* **main\_cat:** The main category to which the product belongs, e.g., “All Electronics”
* **similar\_item:**Related products.
* **price:**Price of the product.
* **asin:**ID of the product, e.g., 0000031852

3. **pos\_words.txt :**This is a text file that contains a corpus of positive words. This file can help you in extracting some features out of the review text. There are multiple sources available online that gives you a corpus of positive negative words. You can also customise it based on your application.

4.  **neg\_words.txt :**Similar to the corpus of positive words, this text file contains a list of words that can be a part of the negative reviews of the users.

5. **Stop\_words\_long.txt:**In the previous modules, you had learnt to remove stopwords using the NLTK library. One drawback of it is, the negation words like nor, not, never are considered as stopwords. Remove those words may not affect a spam/ham classification, but it will definitely affect the sentiment classification. Hence, we have provided you with a customised list of stop words which you can use it for the analysis.