**Improving Sales of Samsung & Apple**

**Introduction:**

As a Data scientist our main goal is to improve the business and sales with different Data science techniques. In this project we choose Natural Language Processing(NLP) for solving the problem by using Topic modelling and with Word frequency techniques using Word Cloud.

**Problem Statement:**

Improving the sales of Samsung and Apple mobile sales.

**Methodology:**

We took the data from “Amazon public datasets” which has the data 400 thousand observations with 168 brands and among all the brands we choose Samsung & Apple with 50k observations each and performed all sorts of text data cleaning operation such as Tokenization, Removal of stop words, Removing the reviews that has less than 3 words, and Lemmatization for converting the words to their root words.

* We used VADER (Valence Aware Dictionary and Sentiment Reasoner) It is a lexicon and rule-based sentiment analysis tool that is specifically used for sentiments. This is one of the best sentiment analyzer which scores each and every word in the review and sum up the total score as compound score based on this compound score the reviews are filtered as Positive review (Compound score >0.5), Negative review (Compound score <-0.5) and Neutral reviews (-0.5<Compound score <0.5) and finally Compound score: is the sum of all of the lexicon ratings which have been standardized to range between -1 and 1.

**Topic Modelling techniques**:

* **LDA (Latent Dirichlet Allocation)** In this algorithm it maximizes the separation between the means of projected topics and minimizes variance within each projected topic.
* LDA, or Latent Derelicht Analysis is a probabilistic model, and to obtain cluster assignments, it uses two probability values: P(word | topics)and P( topics | documents). These values are calculated based on an initial random assignment, after which they are repeated for each word in each document, to decide their topic assignment. In an iterative procedure, these probabilities are calculated multiple times, until the convergence of the algorithm. Let’s look at an example and walk through a step of the algorithm
* It first initializes K- Clusters and assign each word in the document to one of the K- cluster
* Re-assigns words to new topic based on how the proportion of words for document to a topic is.
* How is the proportion of topic widespread across all the documents and this step will be repeated again and again till coherent topics results.

**NMF (Non-Negative matrix factorization)** is the topic modeling algorithm that the decomposition method for multivariant data such as text corpus.

Non-negative Matrix Factorization is a Linear-algeabreic model, that factors high-dimensional vectors into a low-dimensionality representation. Similar to Principal component analysis (PCA), NMF takes advantage of the fact that the vectors are non-negative. By factoring them into the lower-dimensional form, NMF forces the coefficients to also be non-negative.

Given the original matrix **A**, we can obtain two matrices **W**and **H**, such that A= WH. NMF has an inherent clustering property, such that W and H represent the following information about A:

A (Document-word matrix) — input that contains which words appear in which documents.

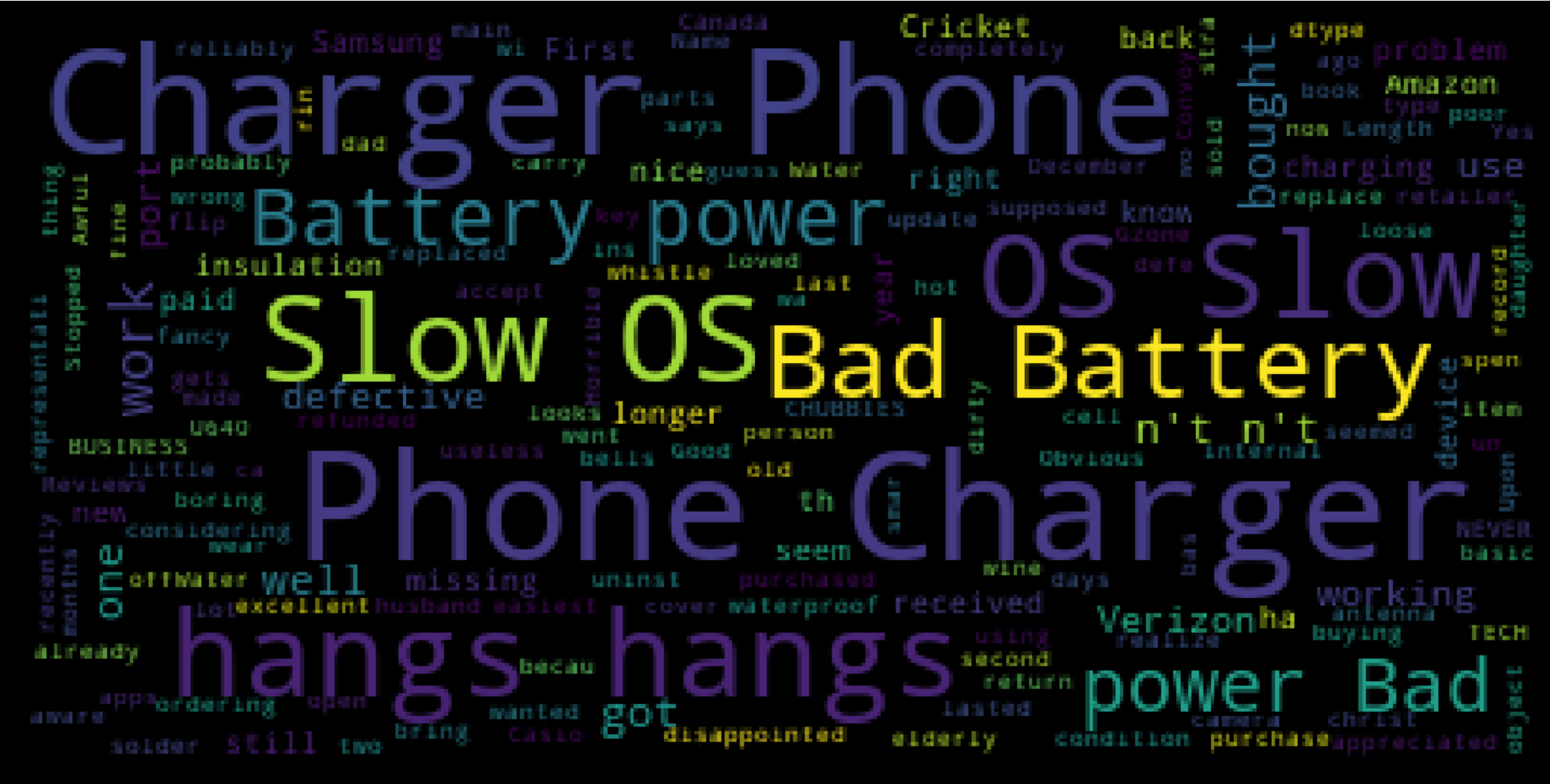
W (Basis vectors) — the topics (clusters) discovered from the documents.

H (Coefficient matrix) — the membership weights for the topics in each document.

We calculate W and H by optimizing over an objective function (like the EM algorithm), updating both W and H iteratively until convergence.

**Word Cloud:**

In this part we took only the negative reviews from Samsung and Apple and converted all the data into tokens and then to dictionaries and performed word frequency for finding the how many times that the particular word repeated in the processed document. Based on this frequency words we plotted them as a word cloud which give the clear picture of high frequent words with big letters and if the frequency reduces the owrd size in the cloud also reduces. Finally, we come with the perfect topics with this method

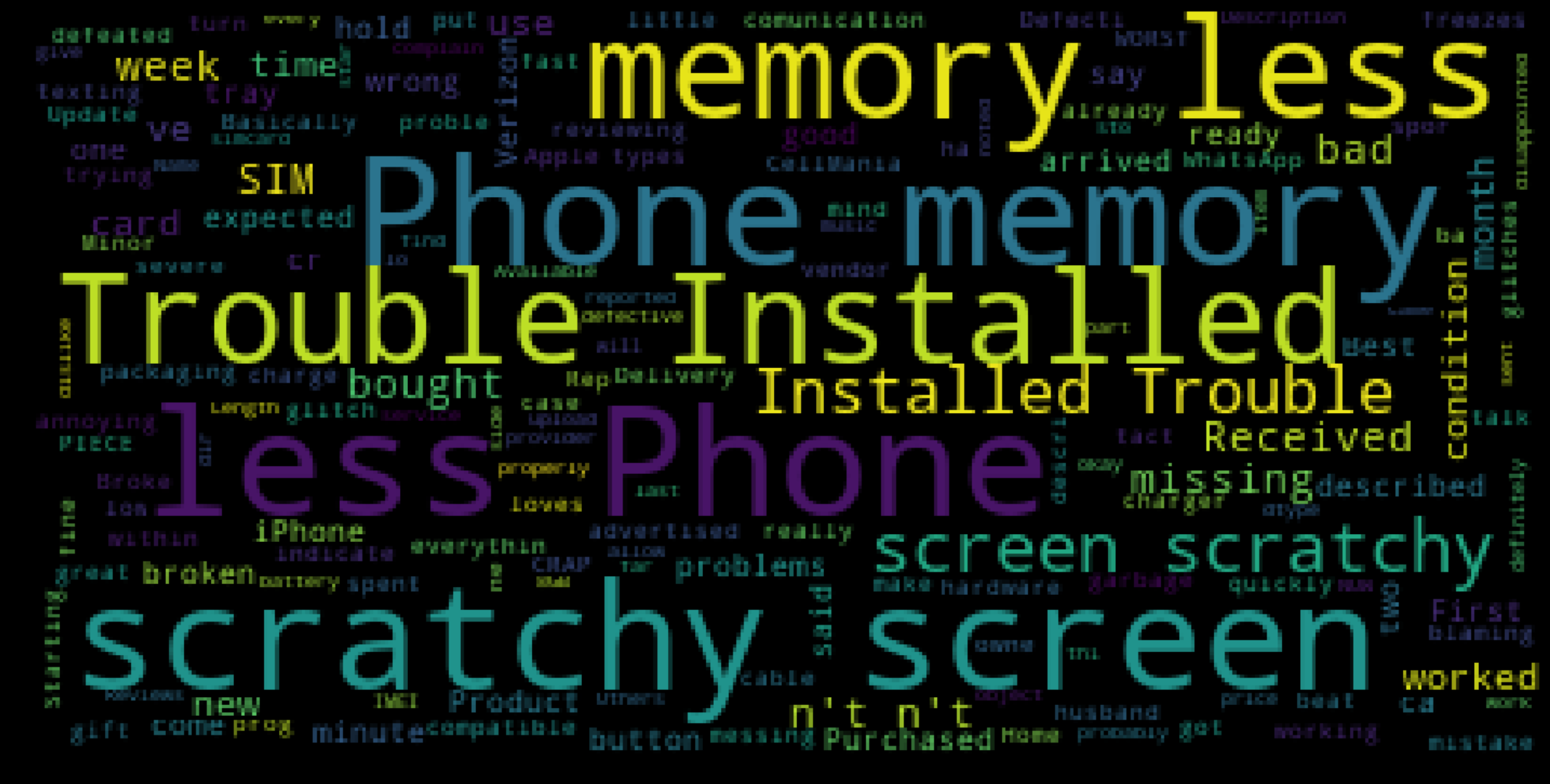


**Result:**

Word Cloud for Samsung Mobile

* (Slow OS, OS bad, hangs)
* (Bad Battery,Phone Charger, disappointing Battery power, power Bad)

Word Cloud for Apple Mobile:



* Trouble Installation
* Conditional Installation Trouble
* Scratchy Screen, Sim slot not working,
* Less memory

Conclusion:

We found for Samsung the most discussed negative topics are

**OS problems, Battery Power, and Missing Items**

Apple**: Installation problem, Damaged Screens, Damaged Battery, Less memory and missing Items**

**Common Topic: MISSING ITEMS**