* **Task1 :**

For this task I’ve downloaded reviews for three product type from <http://jmcauley.ucsd.edu/data/amazon/>

Automotive 5-core (20,473 reviews)

Office Products 5-core (53,258 reviews)

Digital Music 5-core (64,706 reviews)

In these reviews, I’ve considered only of reviews in 2013 to make corpus less huge.

For each product type below are the top-20 terms (after removing stop words and lemmatization), nouns and adjectives/ adverbs:

**Automotive:**

In total **9022** reviews processed, these are terms, nouns and adjectives/adverbs:

|  |  |  |
| --- | --- | --- |
| Terms | Nouns | Adjectives/Adverbs |
| 'nt': 4153,  'car': 3499,  'use': 3307,  'one': 3221,  'work': 3002,  'product': 2680,  'good': 2507,  'great': 2496,  'well': 2377,  'like': 2339,  'get': 2180,  'would': 2167,  'used': 2020,  'time': 1898,  'battery': 1734,  'easy': 1632,  'light': 1507,  'make': 1485,  'need': 1411,  'much': 1356 | 'i': 20714,  'nt': 4153,  's': 3666,  'use': 3307,  'car': 2929,  'product': 2103,  'time': 1545,  'works': 1510,  'work': 1492,  'battery': 1409,  'need': 1193,  'price': 1072,  've': 1027,  'bought': 996,  'light': 994,  'water': 970,  'oil': 964,  'quality': 946,  'fit': 908,  'put': 849 | 'great': 2496,  'good': 2494,  'easy': 1632,  'other': 1381,  'much': 1356,  'little': 1135,  'nice': 1057,  'new': 874,  'small': 708,  'last': 698,  'few': 651,  'best': 634,  'most': 596,  'old': 596,  'same': 580,  'many': 542,  'black': 516,  'high': 451,  'hard': 437,  'different': 418 |

**Office Products:**

In total **20714** reviews processed, these are terms, nouns and adjectives/adverbs:

|  |  |  |
| --- | --- | --- |
| Terms | Nouns | Adjectives/Adverbs |
| 'nt': 10323,  'printer': 9048,  'use': 7827,  'one': 7342,  'paper': 6266,  'like': 6026,  'work': 5061,  'would': 5003,  'ink': 4915,  'pen': 4721,  'great': 4611,  'print': 4546,  'good': 4490,  'well': 4312,  'get': 3995,  'color': 3964,  'need': 3866,  'time': 3834,  'also': 3683,  'easy': 3679 | 'i': 47074,  'nt': 10323,  's': 9508,  'use': 7827,  'printer': 7657,  'paper': 5514,  'ink': 4578,  'print': 3447,  'quality': 3267,  'time': 3217,  'tape': 3204,  'need': 3079,  'work': 2840,  've': 2752,  'price': 2603,  'color': 2458,  'm': 2418,  'pens': 2398,  'product': 2363,  'printing': 2306 | 'great': 4611,  'good': 4471,  'easy': 3679,  'other': 3335,  'much': 3059,  'nice': 2642,  'little': 2586,  'small': 2004,  'most': 1767,  'black': 1626,  'many': 1492,  'few': 1447,  'old': 1227,  'same': 1210,  'new': 1189,  'last': 1153,  'big': 1143,  'scan': 1120,  'best': 1107,  'different': 1107 |

**Digital Music:**

In total **4589** reviews processed, these are terms, nouns and adjectives/adverbs:

|  |  |  |
| --- | --- | --- |
| Terms | Nouns | Adjectives/Adverbs |
| 'song': 6391,  'album': 6129,  'nt': 2884,  'like': 2860,  'one': 2689,  'music': 2326,  'love': 1944,  'great': 1915,  '34': 1884,  'sound': 1791,  'track': 1763,  'good': 1729,  'band': 1594,  'time': 1512,  'cd': 1459,  'really': 1278,  'get': 1121,  'would': 1118,  'best': 1115,  'first': 1114 | 'i': 10537,  's': 6410,  'album': 5286,  'song': 3947,  'nt': 2884,  'songs': 2444,  'music': 2326,  'love': 1882,  'cd': 1367,  'band': 1350,  'sound': 1218,  'time': 1210,  'track': 967,  'albums': 843,  'rock': 841,  'listen': 820,  'tracks': 796,  'm': 736,  'lyrics': 691,  'way': 688 | 'great': 1913,  'good': 1721,  'best': 1114,  'new': 907,  'much': 894,  'most': 882,  'other': 847,  'many': 571,  'little': 506,  'classic': 467,  'few': 436,  'last': 411,  'same': 404,  'live': 391,  'such': 386,  'single': 368,  'original': 365,  'own': 364,  'hard': 361,  'nice': 358 |

**Observations:**

1. In all the categories we can see a pattern that most frequent nouns, adjectives will also be part of the terms but not vice versa.
2. As expected some words which are most related for that category occur frequently in terms and nouns such as ‘car’ for automotive category, ‘printer’ for office products and ‘song’ & ‘album’ for digital music category. However, we can see that adjectives/adverbs won’t tell much about the category, so most frequent of them are same in all categories because their place is to attribute extra information to sentence rather than concept of it.

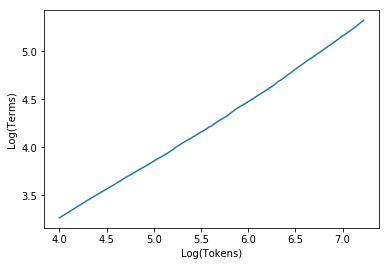
* **Task2:**

For this task I’ve considered below review category:

‘Tools and Home Improvement 5-core (134,476 reviews)’.

As part of this task I’ve captured total tokens and terms (distinct tokens) among all the reviews in this category. To count the number of terms I’ve used **‘hyperloglog’** counter (with 1% error) for memory efficiency. All the captured tokens and terms are saved in ‘task2\_tokens\_count.txt’ with review number at each 10K.

After all the reviews are processed, I’ve plotted the Log10(Tokens) and Log10(Terms)



**Observations:** Though the number of terms (210326) at the end is very less compared to tokens (16700000), both of them are linearly increasing with every review in log-scale. This shows that when corpus size increases even distinct tokens also starts to increase.