Word Statistics Challenge

Intro

You have 2 fields,

- Tariff Number field – Can be 9 or 10 digits, if it is 9 digits assume there should be a zero in front of the first number.
- Description field – A string field that has various words either separated by space, comma, or other characters.

The Tariff Number field value example – 0901.21.0010

Chapter	Heading	Subheading	Duty Rate	Tariff / Tariff Number
09	0901	090121	09012100	0901210010

The Challenge

The goal is to parse every word out of the Description field, and then do:

1- A word count overall – Count how many times that word is being used.
2- **The unique count** and **the max count** of word based on the Tariff Number:
   a. how many words share the first 2 digits of the Tariff Number (called Chapters),
   b. how many words share the first 4 digits of the Tariff Number (called Heading),
   c. how many words share the first 6 digits of the Tariff Number (called Subheading),
   d. how many words share the first 8 digits of the Tariff Number (called the Duty Rate) and
   e. how many words share all 10 digits (called the Tariff Number).

Example of Result

Let us say you have 4 entries like this:

TariffNumber	Description
0100405060	The Red Elephant Market
0100405040	Elephant the market
0410500030	The Red violin
3900339302	The Yellow Potato the Elephant

The result table with the statistics would have these columns:

1. Id – internal identifier
2. Word – Extracted word from description
3. TotalCount – total number of descriptions in which word occurs
4. SingleChapterMaxCount – the highest number of word occurrences for a single Chapter
5. UniqueChaptersCount – number of unique Chapters in which word occurs
6. SingleHeadingMaxCount – the highest number of word occurrences for a single Heading
7. UniqueHeadingCount – number of unique Headings in which word occurs
8. SingleSubheadingMaxCount – the highest number of word occurrences for a single Subheading
9. UniqueSubheadingCount – number of unique Subheadings in which word occurs
10. SingleDutyRateMaxCount – the highest number of word occurrences for a single Duty Rate
11. UniqueDutyRateCount – number of unique Duty Rates in which word occurs
12. SingleTariffMaxCount – the highest number of word occurrences for a single Tariff
13. UniqueTariffCount – number of unique Tariffs in which word occurs

And the table would look something like this:

Id	Word	Total Count	Single Chapter Max-Count	Unique Chapter Count	Single Heading Max-Count	Unique Heading Count	Single Subheading Max-Count	…	Single Tariff Max-Count	Unique Tariff Count
1001	red	2	1	2					1	2
1005	elephant	3	2	2	2	2	2		1	3
1015	the	5	2	3					1	4
1026	market	2	2	1					1	2
2008	violin	1	1	1					1	1
4009	yellow	1	1	1					1	1
6010	potato	1	1	1					1	1

Additional Info

Obviously, there may be several other ways of solving this with no database or with database. You may also use several temporary tables which you may want to create first before you do a “group by”, “order by” or “select distinct” to count all the various numbers up together to create this final summary table as described above.

The result is a CSV file with all the field as described above, or even more if you find them important.