**Part of speech processing**

The project originally wanted to use the stem of the word to convert the words of the same meaning into the same words such as ‘chips’ can be converted to ‘chip’. However, the project found that after using Python Stemming, some proper nouns were converted into the wrong words, such as ‘haggis’ was converted to ‘haggi’. As a result, the project decided to combine only the singular and plural nouns to minimize the risk of incorrect categorising words because of word stem. This uses the part-of-speech judgment and stem detection method in the Python nltk package.

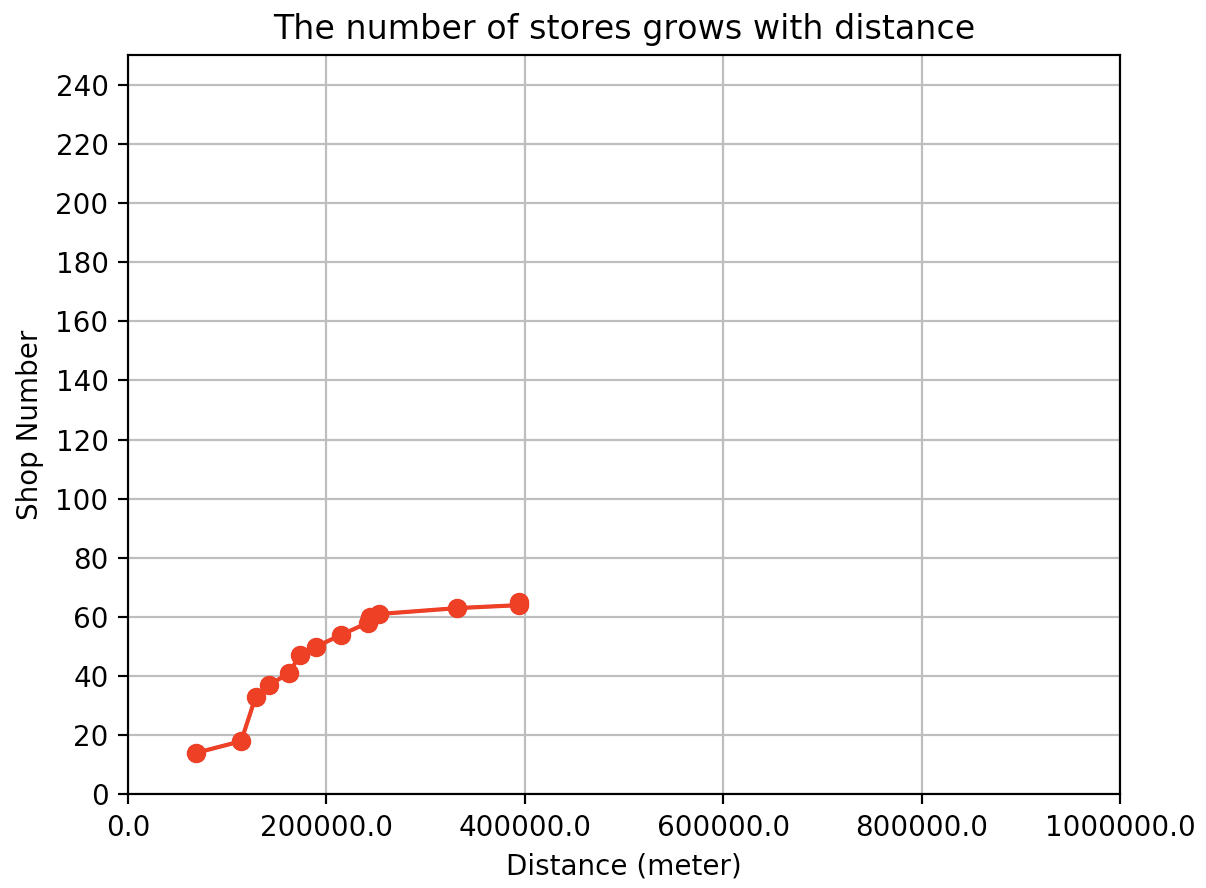
**Scoring method**

The project used another scoring method: Observing the trend of the number of shops increasing with distance (meter) from the centre point. The project used 6 regional words and 6 widely distributed words to observe the differences and patterns of regional words and non-regional words. Besides, in order to explore the method of judging words which distributed in a small number of shops (initially 15 or less), the project selected 6 examples to explore the pattern. The reason why the project wants to find this pattern is that if the number of shops of a word is too small, the frequency of this word will be very low, and all distributions may be below 200000 meters. If the proportion defined in regional words findings is used, all such words will be judged as regional word. In doing so, the possibility of misjudgement will increase. Thus, in order to reduce the false positive rate, the project will ignore some words with a small number of shops. If the project wants to know the categories of these words which were ignored, the project needs more data samples.

**Regional words findings:**

By observing these six regional words, the project found that the number of shops of regional words increases significantly within one-fifth of the largest distance (1,000,000 meters). The reason why the project selected 1,000,000 meters as the largest distance is to make the results easy to observe. There are few shops’ distance exceeds 1,000,000 meters and the project has calculated them in the total shop number. Thus, even if these shops are not shown in the figure, the results will not be affected.

Through the observation data below, the project found that in the range of 200,000 meters, if a word is a regional word, the coverage of the number of shops will reach 60% or more.

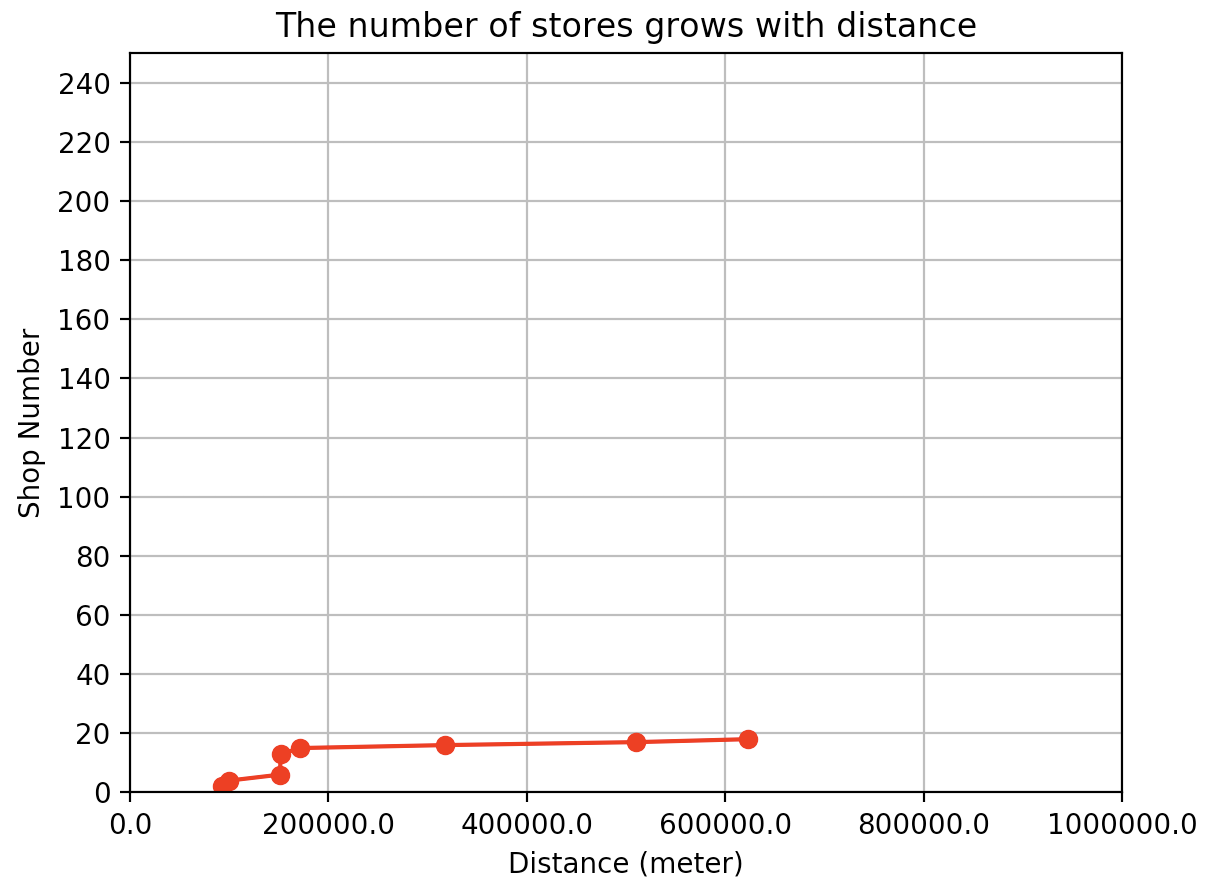


**Figure 1: The number of ‘haggis’ shops varies with distance**

less than 200000: 50.

total shop number: 65.0.

proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 76.9%.

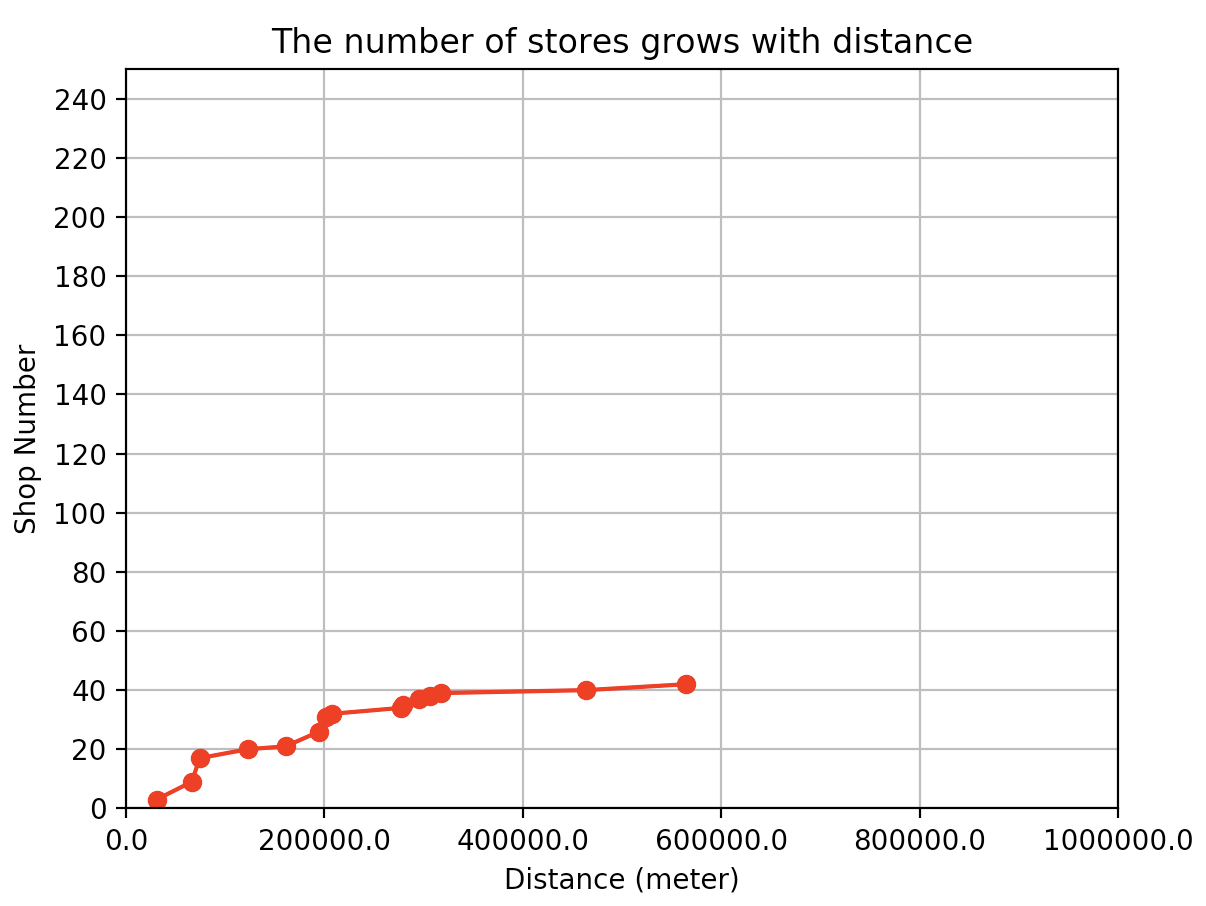
****

**Figure 2: The number of ‘balti’ shops varies with distance**

less than 200000: 15.

total shop number: 18.0.

proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 83.3%.

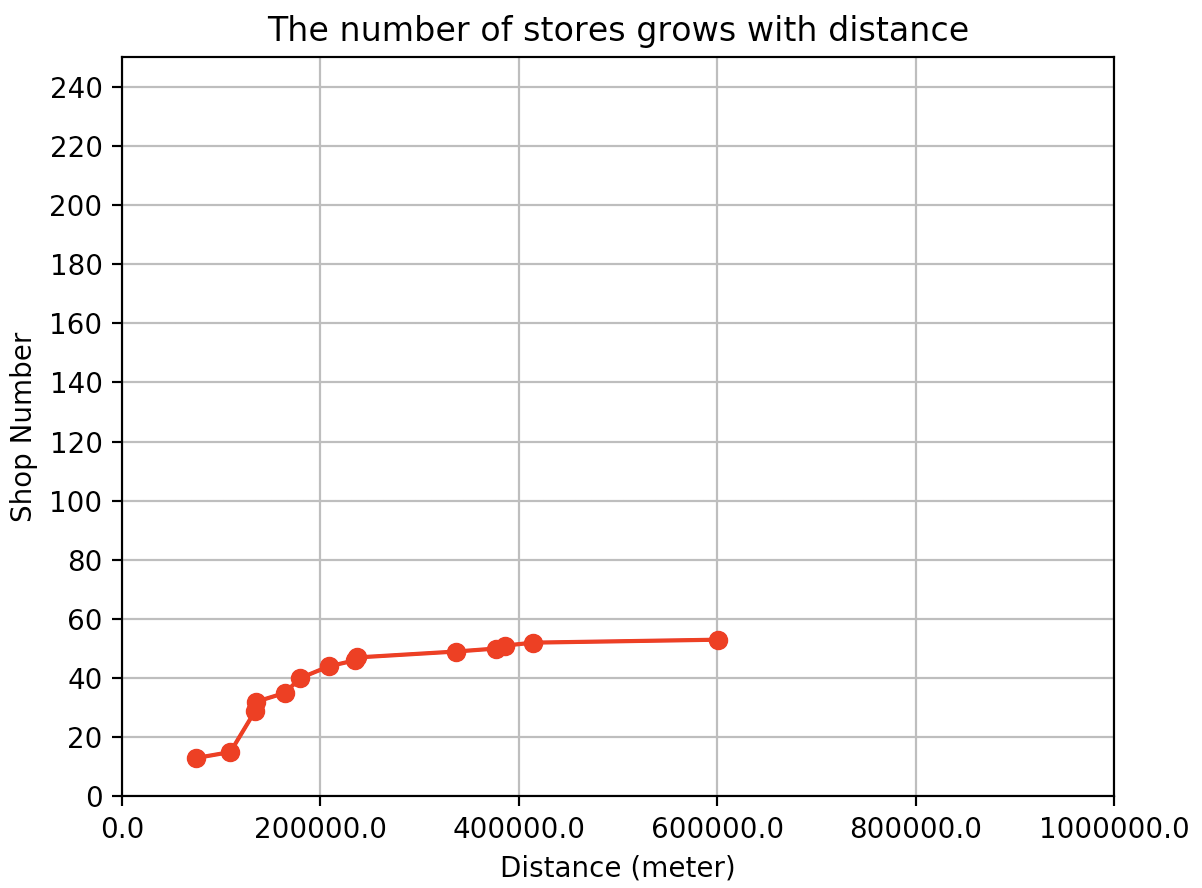
****

**Figure 3: The number of ‘roe’ shops varies with distance**

less than 200000: 26

total shop number: 42.0

proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 61.9%.

****

**Figure 4: The number of ‘irn’ shops varies with distance**

less than 200000: 40.

total shop number: 53.0.

proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 75.4%.

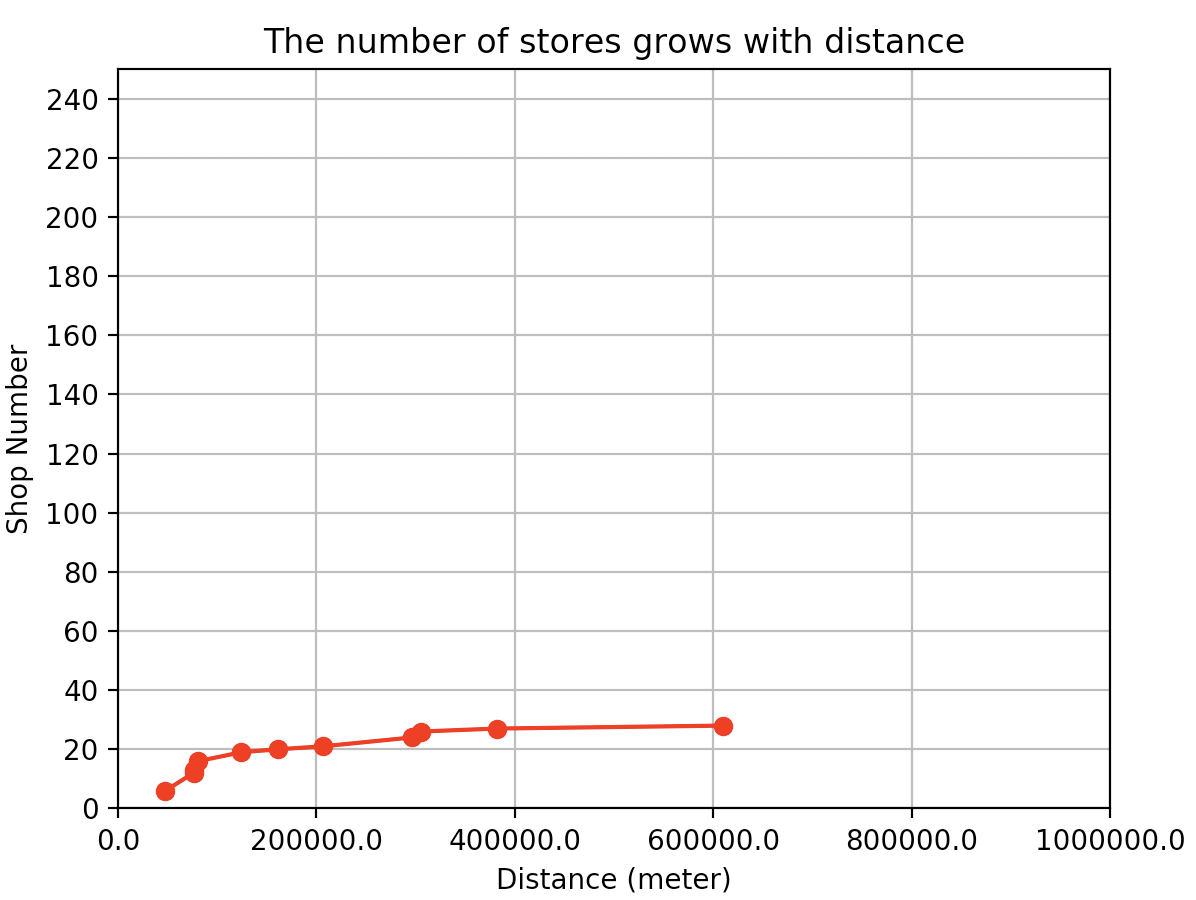
****

**Figure 5: The number of ‘naan’ shops varies with distance**

less than 200000: 28.

total shop number: 31.0.

proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 90.3%.

****

**Figure 6: The number of ‘kiev’ shops varies with distance**

If the project uses median to judge ‘kiev’, the word will be judged as wide distributed word (less/total=67%<70%).

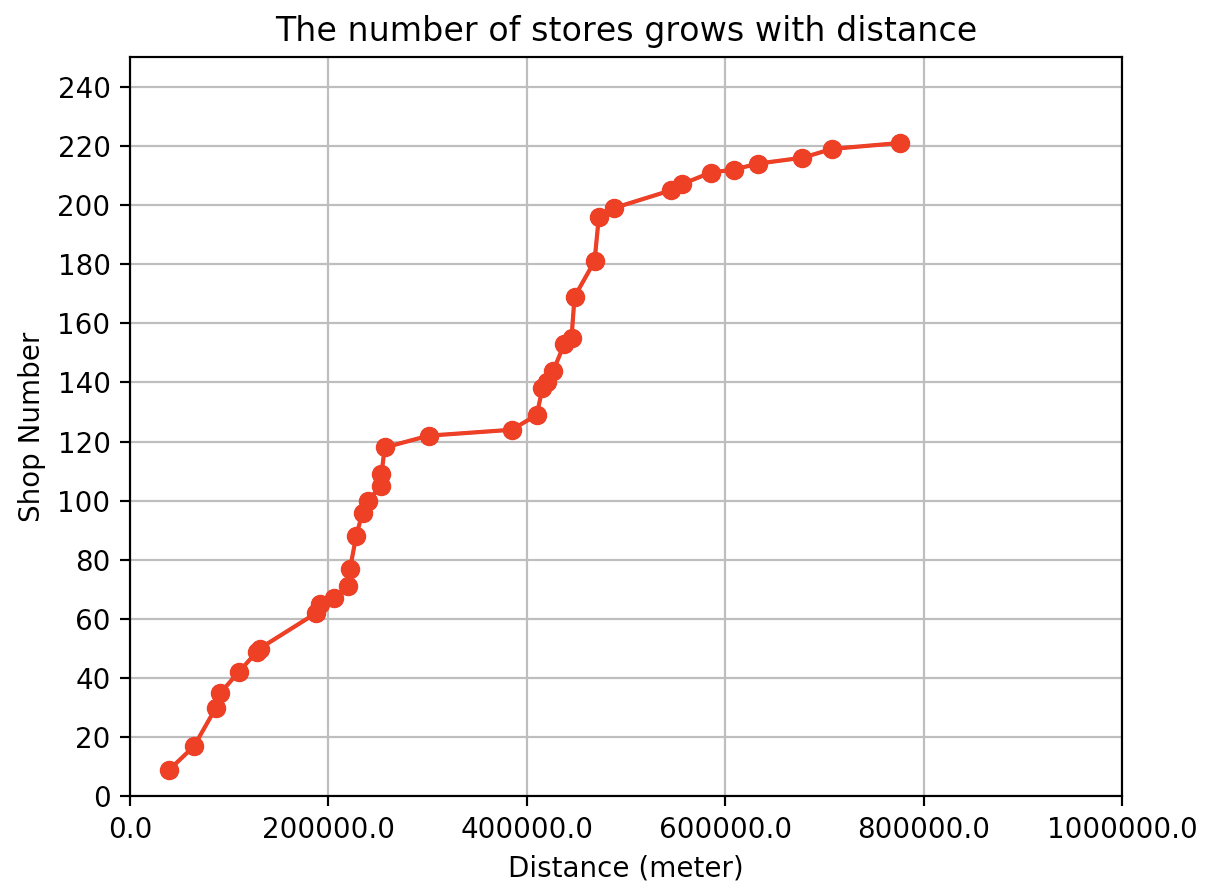
less than 200000: 20.

total shop number: 28.0.

proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 71.4%.

**Widely distributd words findings:**

In terms of widely distributed words, in the range of 200,000 meters, if a word is a widely distributed word, the coverage of the number of shops is not more than 40%.

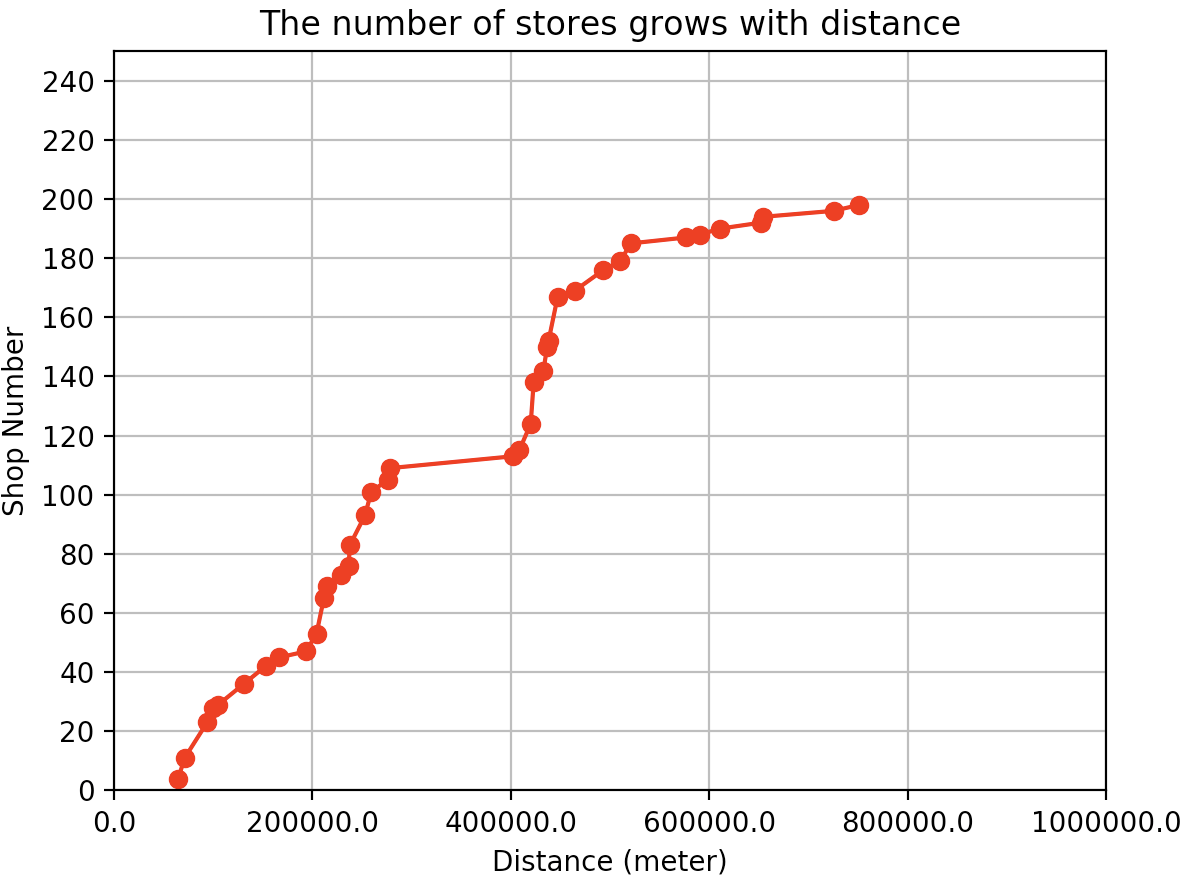
****

**Figure 7: The number of ‘chip’ shops varies with distance**

less than 200000: 65.

total shop number: 221.0.

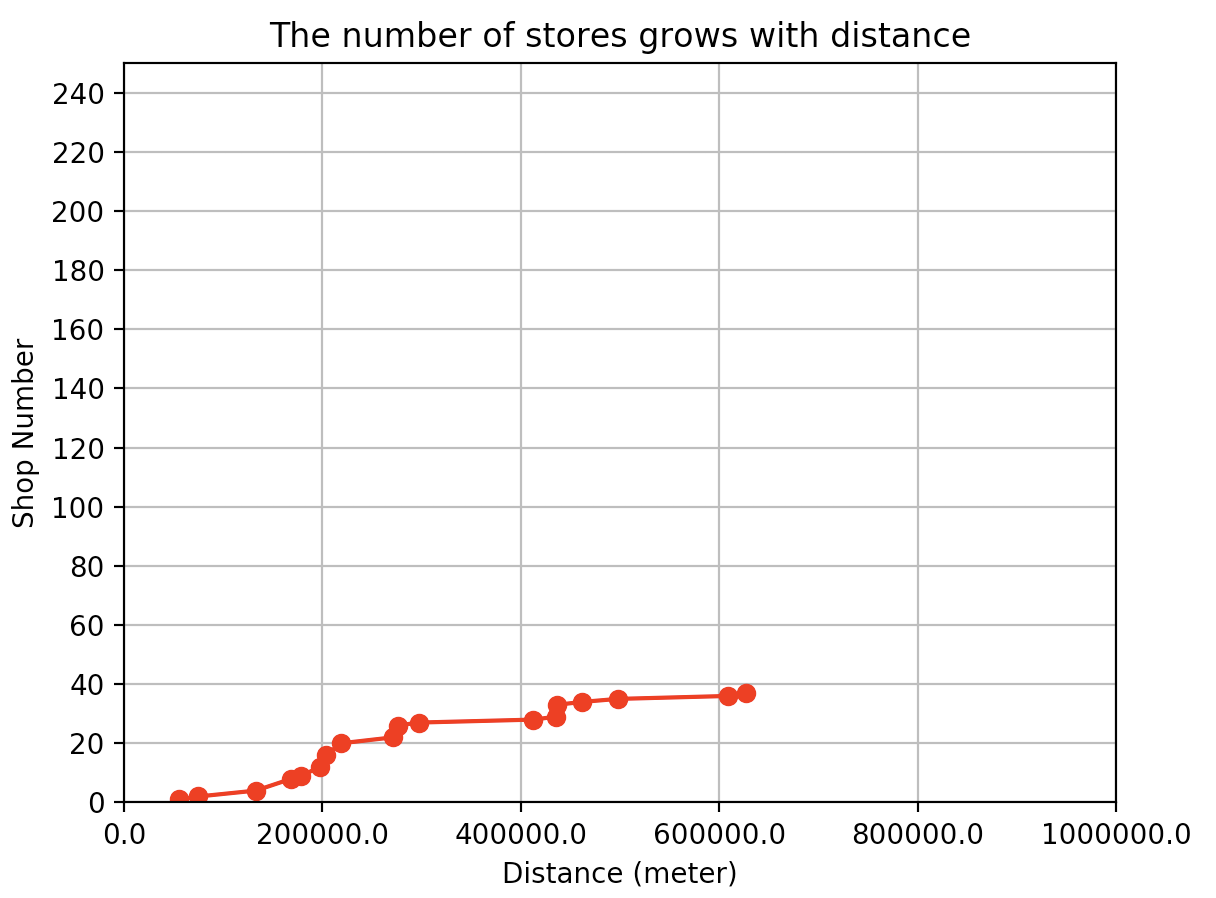
proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 29.4%.

****

**Figure 8: The number of ‘sausage’ shops varies with distance**

less than 200000: 47.

total shop number: 198.0.

proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 23.7%.****

**Figure 9: The number of ‘supreme’ shops varies with distance**

less than 200000: 12

total shop number: 37.0

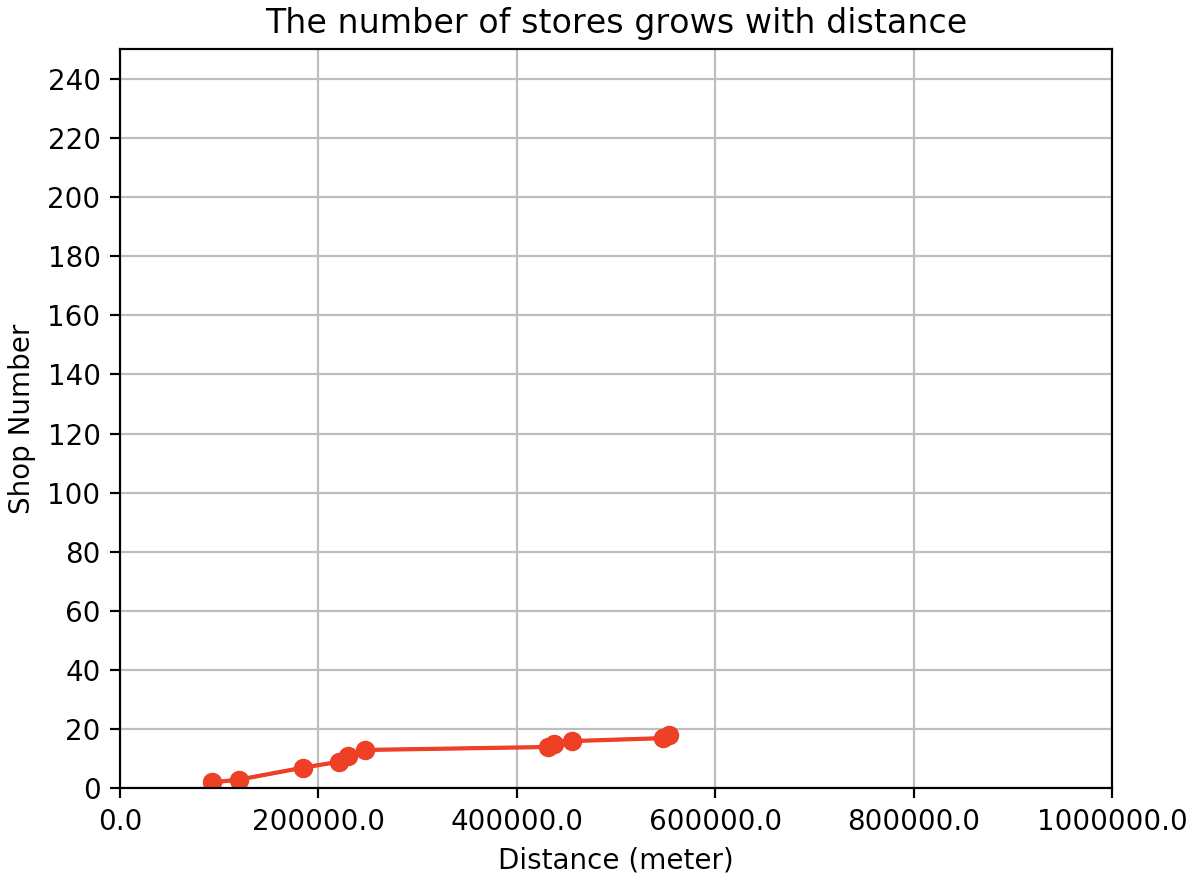
proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 32.4%.

****

**Figure 10: The number of ‘gherkin’ shops varies with distance**

less than 200000: 12.

total shop number: 49.0.

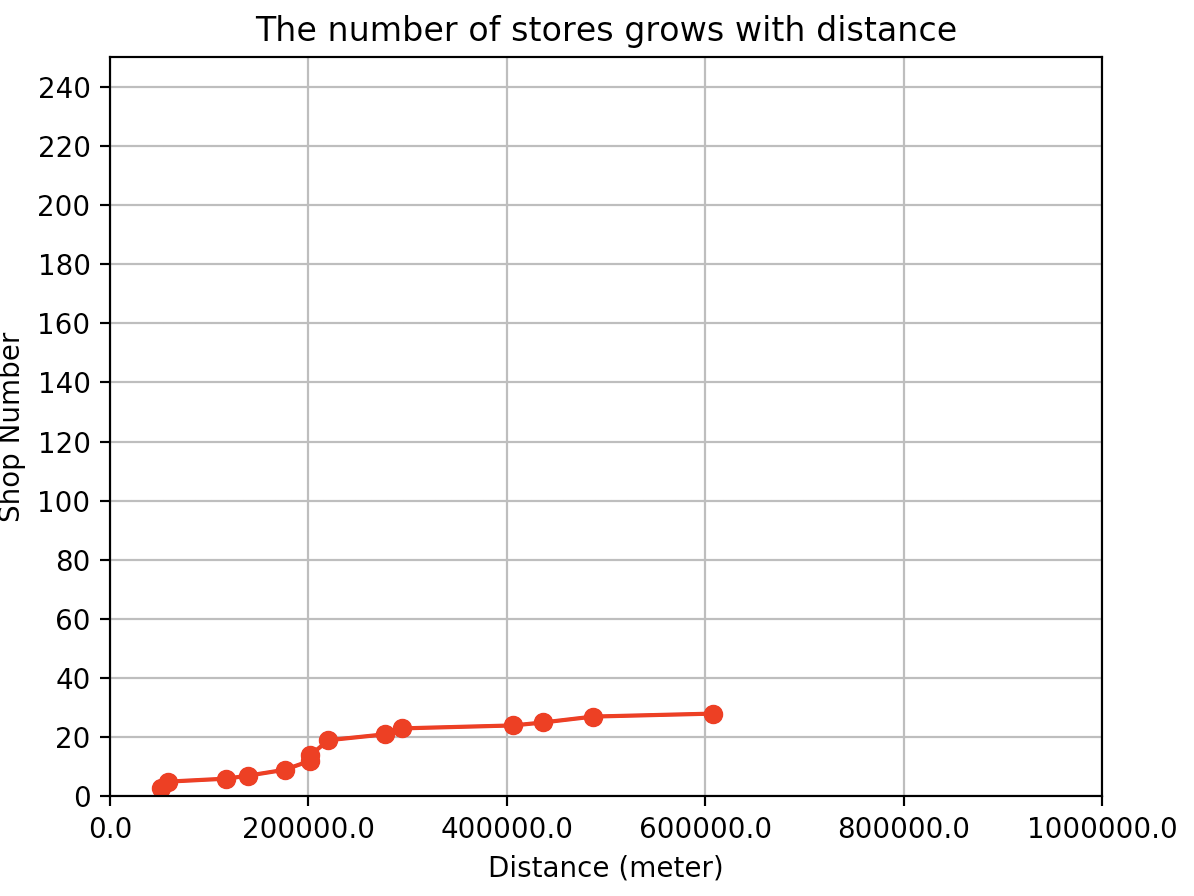
proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 24.4%.****

**Figure 11: The number of ‘yoghurt’ shops varies with distance**

less than 200000: 7.

total shop number: 18.0.

proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 38.8%.

****

**Figure 12: The number of ‘gift’ shops varies with distance**

less than 200000: 9.

total shop number: 28.0.

proportion (the number of shops whose distance is less than 20000 meters from the center point/ total shop number): 32.1%.

**Words which distributed in a small number of cities:**

‘securely’ (Fig. 13, Fig. 14) should be a common word, but because there are too few samples, it will be divided into widely distributed words by the proportion.

‘quattro Stagioni’ (Fig. 15, Fig. 16) is a traditionally Italian recipe. It widely distributed in Scotland, the project will judge ‘Stagioni’ and ‘Quattro’ as regional words. However, this may not be true, because these two words are Italian.

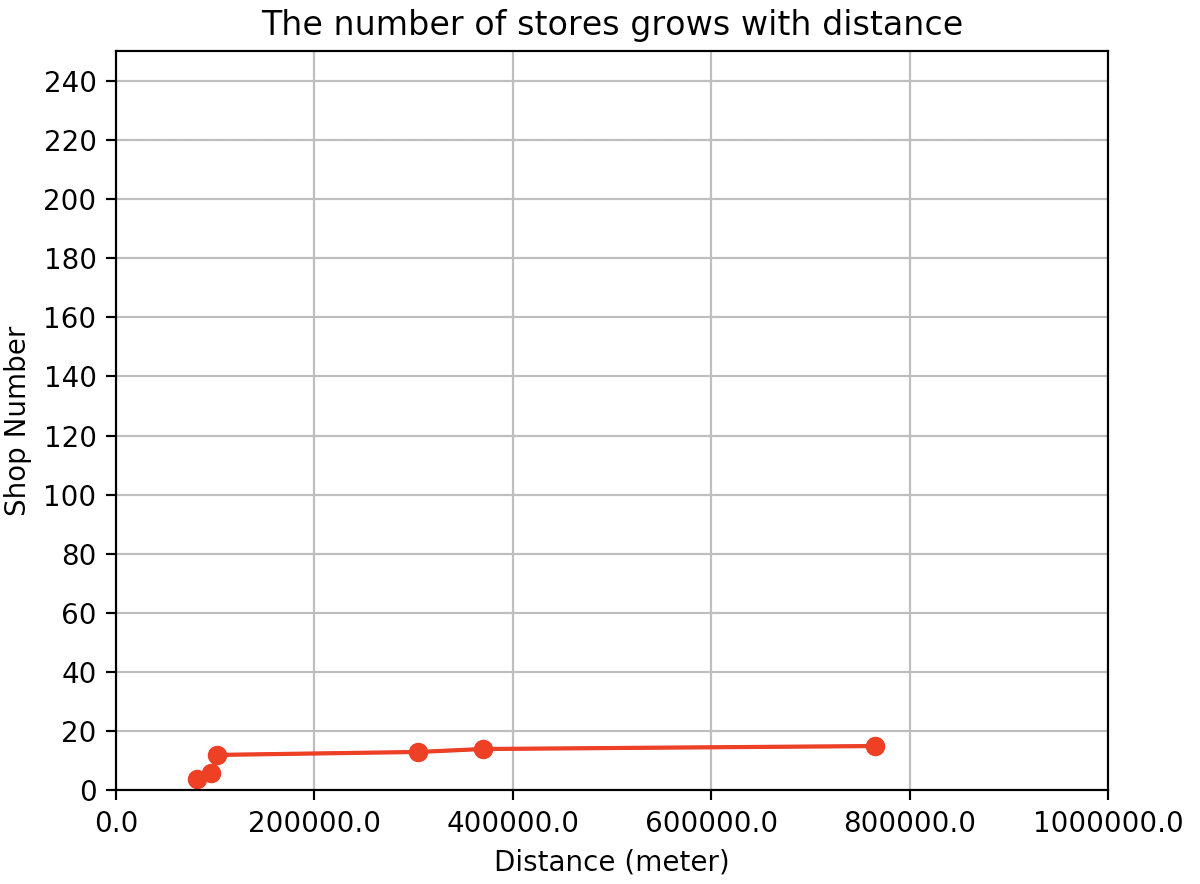
Because of lack of data, ‘farm’ (Fig. 17, Fig. 18) is looked as a regional word, but the proportion will judge it as a wide distributed word.

‘keema’ (Fig. 19, Fig. 20) has 11 shops, but it can be judged as a regional word.

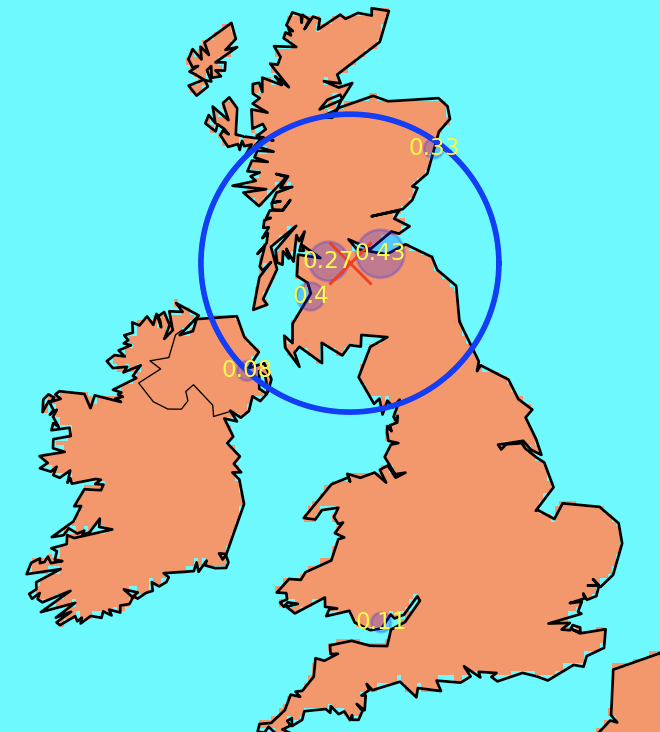
‘haagen’ (Fig. 21, Fig. 22) has 13 shops, this word should have been linked to Dazs (Haagen-Dazs). It should be a widely distributed word, but it is judged as a regional word.

‘carbonara’ (Fig. 23, Fig. 24) has 15 shops, but it can be judged as a regional word.

Based on the above results, using 15 as the smallest number of shops that can be judged, the project will miss some regional words, but will also classify some words incorrectly. Thus, the project plan to use 10 as the smallest number of shops that can be judged a word and evaluate the classification results.

****

**Figure 13: The number of ‘securely’ shops varies with distance**

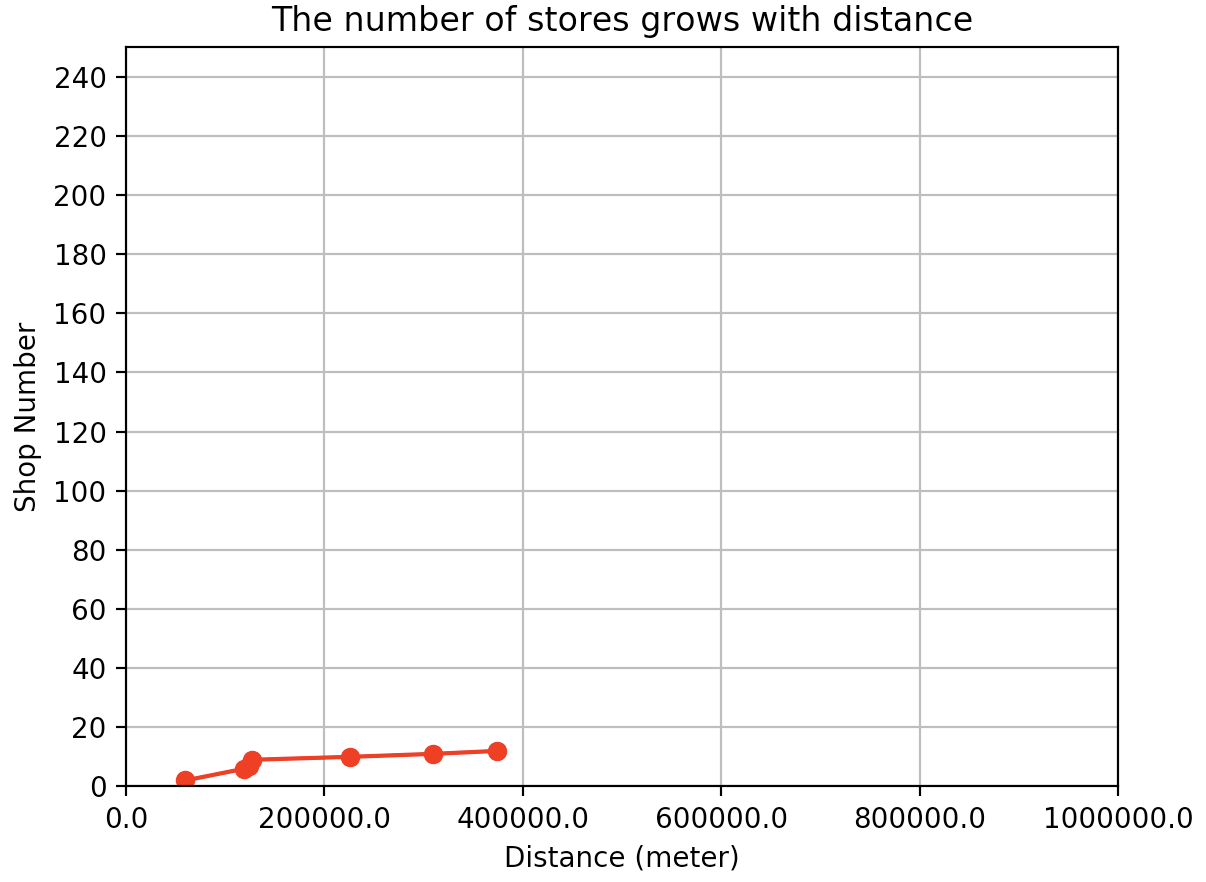
****

**Figure 14: The distribution of ‘securely’**

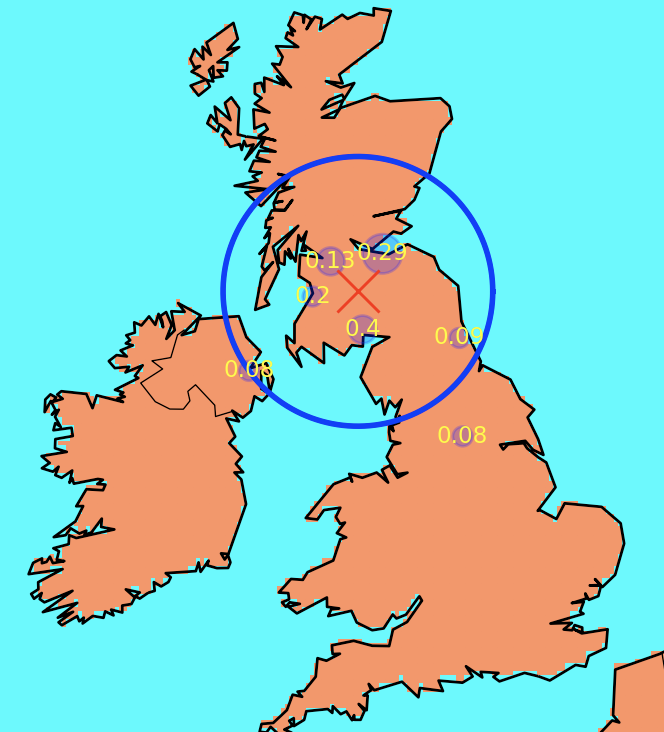
The number of cities: 6.

less than 200000: 12.

total shop number: 15.



**Figure 15: The number of ‘Stagioni’ shops varies with distance**

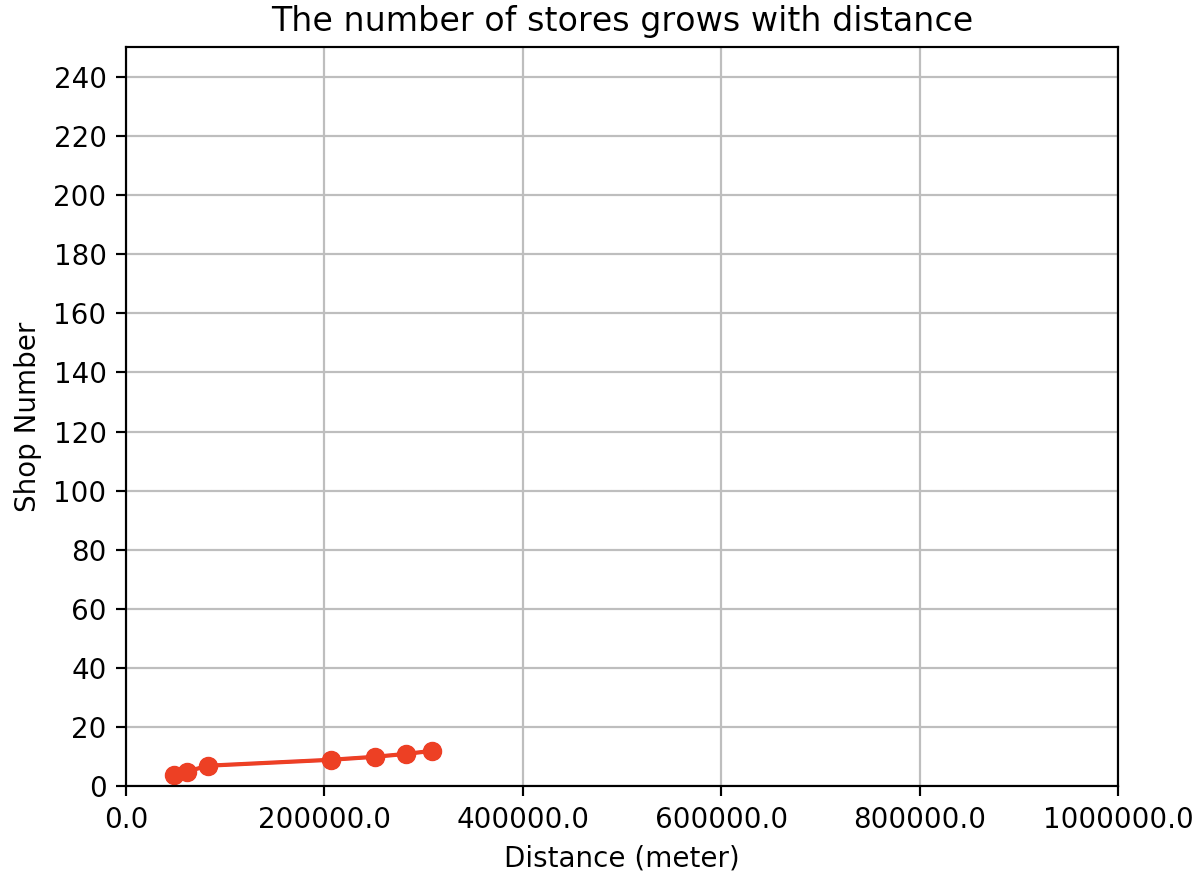


**Figure 16: The distribution of ‘Stagioni’**

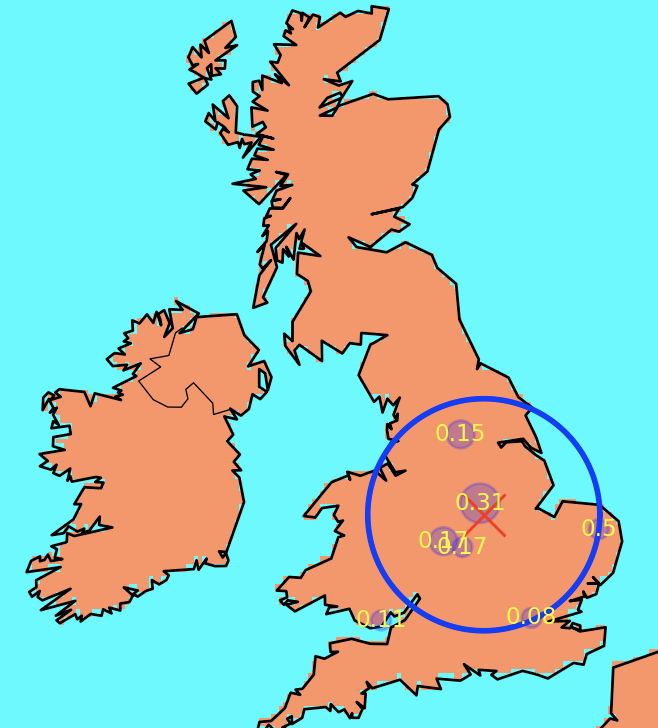
The number of cities: 7.

less than 200000: 9.

total shop number: 12.



**Figure 17: The number of ‘farm’ shops varies with distance**



**Figure 18: The distribution of ‘farm’**

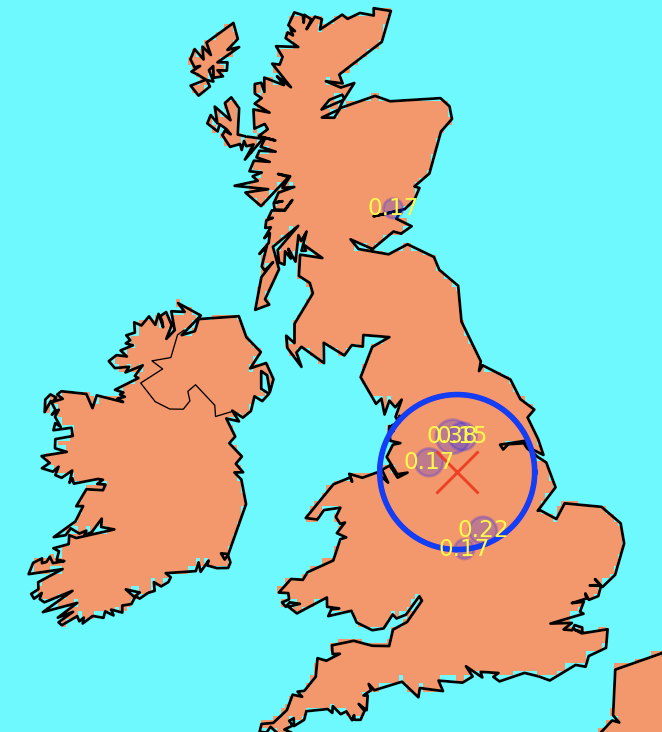
The number of cities: 7.

less than 200000: 7.

total shop number: 12.



**Figure 19: The number of ‘keema’ shops varies with distance**

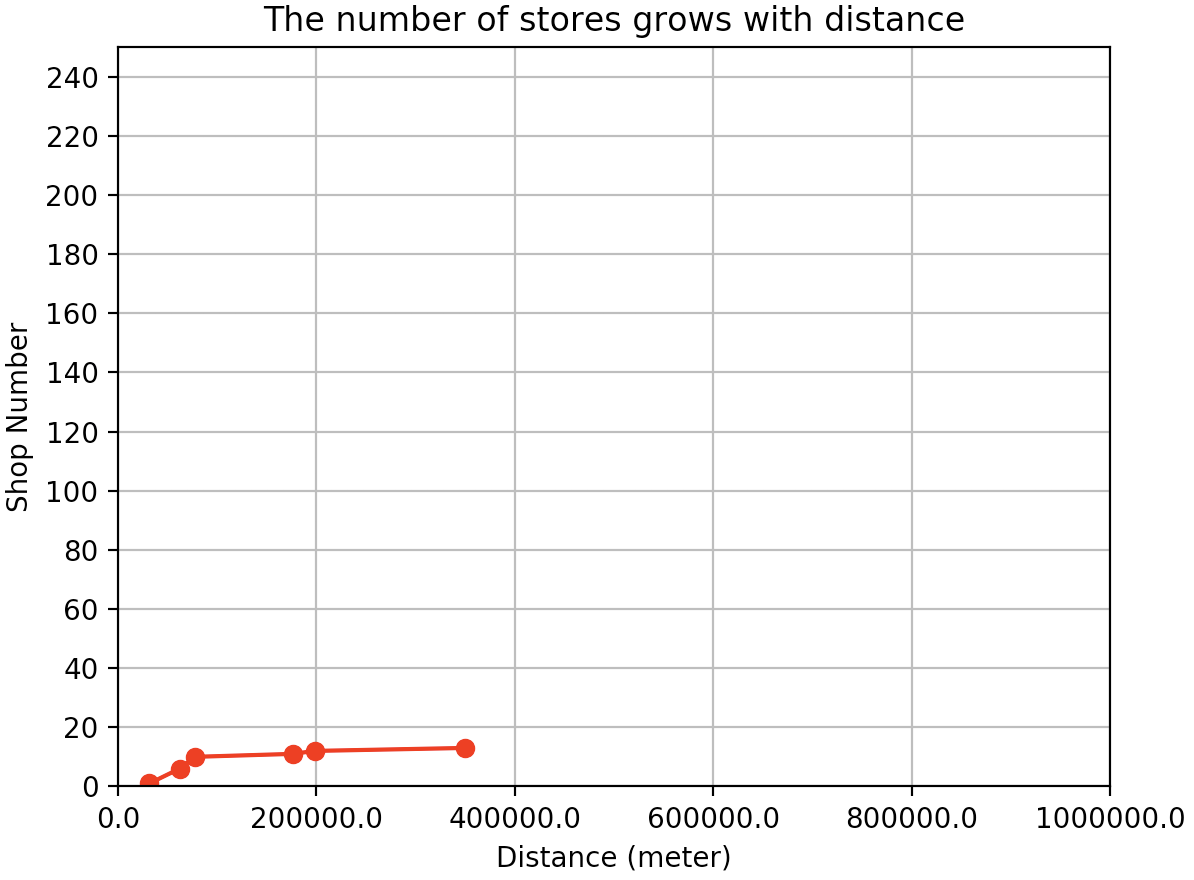


**Figure 20: The distribution of ‘keema’**

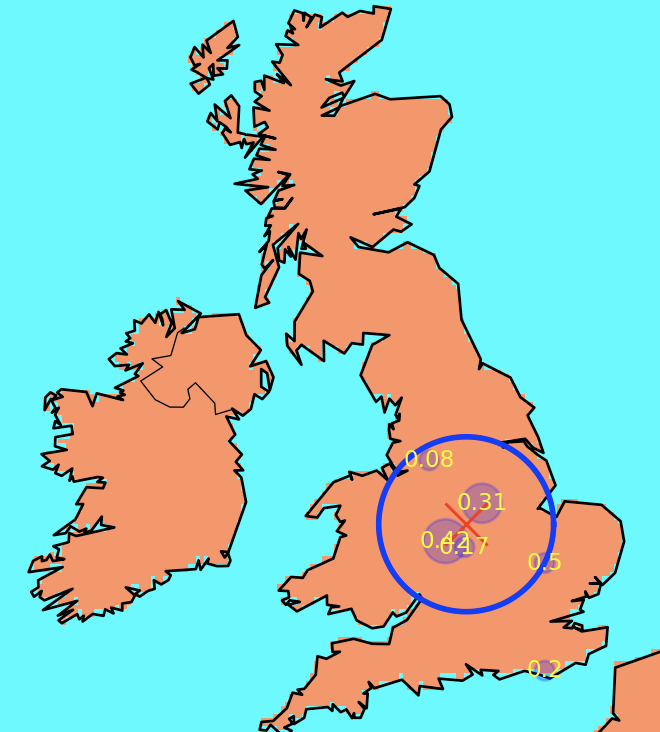
The number of cities: 6.

less than 200000: 9.

total shop number: 11.



**Figure 21: The number of ‘haagen’ shops varies with distance**

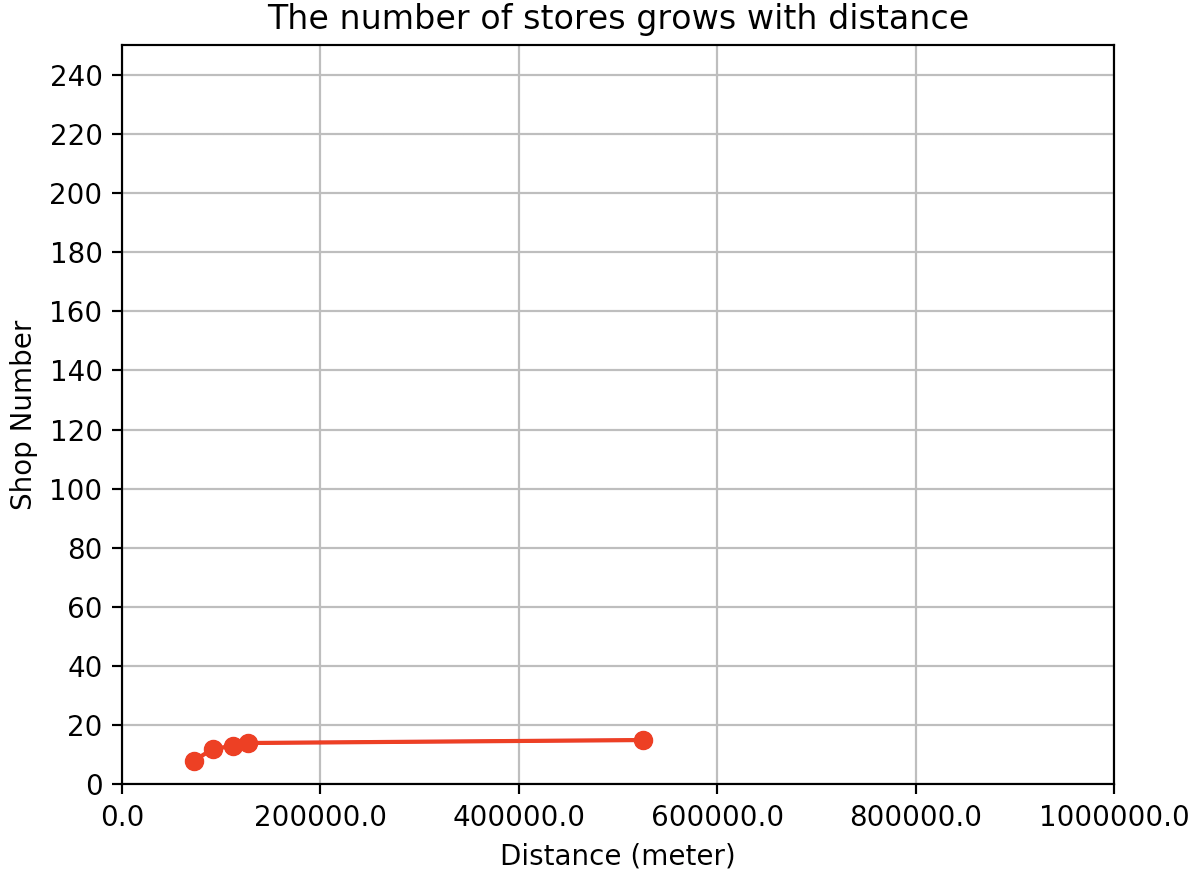


**Figure 22: The distribution of ‘haagen’**

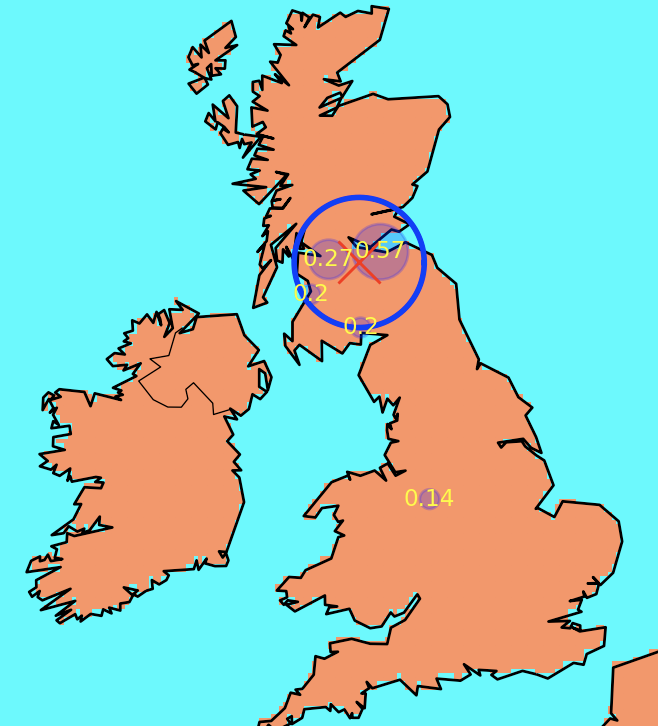
The number of cities: 6.

less than 200000: 12.

total shop number: 13.



**Figure 23: The number of ‘carbonara’ shops varies with distance**



**Figure 24: The distribution of ‘carbonara’**

The number of cities: 5.

less than 200000: 14.

total shop number: 15.

**Result**

Under the following standards:

1. Selection of proportion: The number of shops within 200,000 meters from the centre point/ The total shops number.
2. Words which have more than 10 shops can be marked as regional word or widely distributed word. Otherwise, they will be marked as unknown words.
3. If the proportion greater than 60, the word will be marked as regional word.
4. If the proportion less than 60, the word will be marked as widely distributed word.

There are 72 words judged as the regional words and 664 words judged as widely distributed words. However, because of lack of samples, 4552 words are judged as unknown.

**Evaluation**

The evaluation of the results is based on experience.

There are 36 words that should appear in the widely distributed words category appearing in regional words category. (Hamburger, hamburger, dazs, instantly, haagen, rump, suey, scallop, shot, splash, securely, rock, skate, give, smokey, passion, value, chosen, bull, cucumber, facebook, bit, under, smoked, love, serving, telephone, including, piece, enjoy, spam, per, well, then, supper, site).

In the result of regional words, only 50% of the regional words are judged correctly. Thus, the project wants to improve the accuracy of the judgment. The project attempts to use the average distance from the center point (250000) as metric, and then filter again based on the proportion. The result is that there are 49 words judged as the regional words and 687 words judged as widely distributed words.

There are 23 words that should appear in the widely distributed words category appearing in regional words category (Hamburger, hamburger, dazs, instantly, haagen, scallop, shot, splash, rock, skate, give, smokey, passion, value, cucumber, facebook, bit, under, love, serving, telephone, including, spam). In the result of regional words, 53% of the regional words are judged correctly and the improvement is very small.

The project also attempts to change the determining distance (200000 to 150000) and the proportion of shops within the distance metric to improve accuracy. However, the final result is basically the same as the above result.

The main reason for this result is that some words appear at a low frequency, and the project mainly judges according to the way of the distribution of the word. Thus, this method will cause misjudgment. If each word has as many samples as ‘chip’ or ‘haggis’, the judgement which according to the distribution will be more accurate.