1. Find special words such as ‘Haggis’ which is a regional word and ‘chips’ which is a widely distributed word to calculate central and radius.

Calculate central algorithm:

Use all coordination of the word.

First convert the longitude and latitude to radian.

Calculate the sum of coordinates x, y, z in the sphere.

Radian of longitude is atan2(y, x).

Radian of latitude is atan2(z, (x \* x + y \* y))).

convert radian to degree.

Radius algorithm:

Calculate the distance between all coordinates and the centre point and find the largest distance as the radius.

Use Euclidean distance to calculate distance between two points, because the Basemap has converted the Earth's sphere into a plane. The project firstly choose the Haversine distance to calculate the distance and found that the result is not correct.

1. Set the filter percentage: Used to filter some points which are far from the centre. After filter, recalculate centre point and radius. The Fig. 1, Fig. 2, Fig. 3 and Fig. 4 shows the distribution of ‘haggis’ and ‘chips’ under different filter ratios.

Evaluation:

The percentage is set to 90% because each word is distributed in up to 39 cities and there are up to 4 cities are filtered.

Some cities that are close to the north or south may be filtered every time.

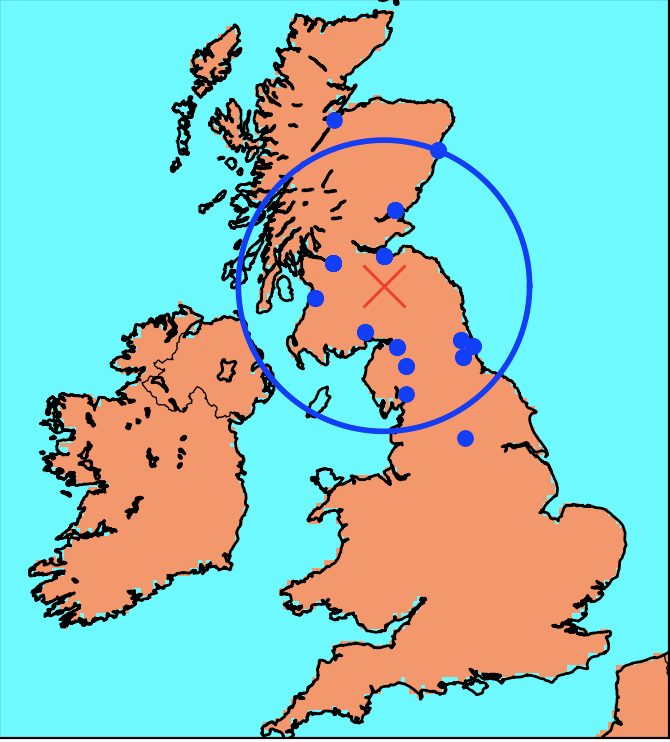


Figure.1 Haggies distribution (Retains the nearest 90% of points from the center)

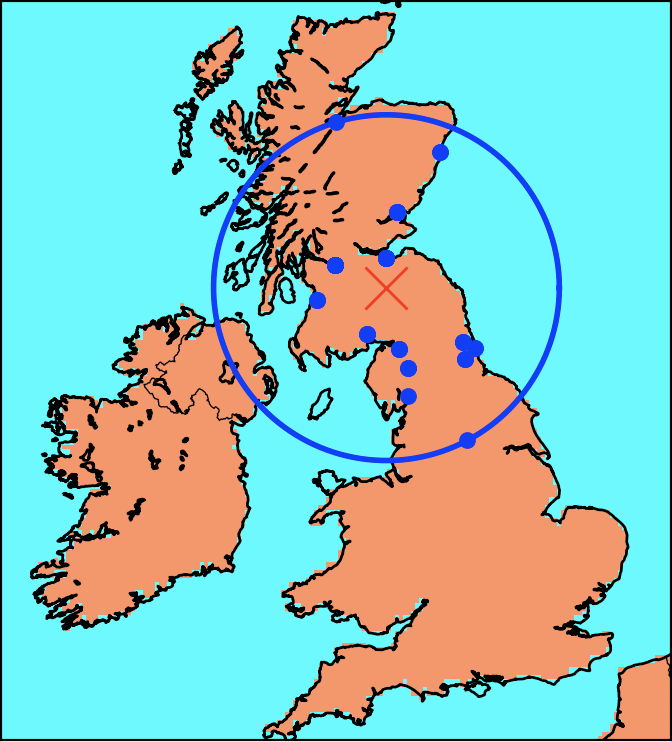


Figure.2 Haggies distribution (Retains all cities)

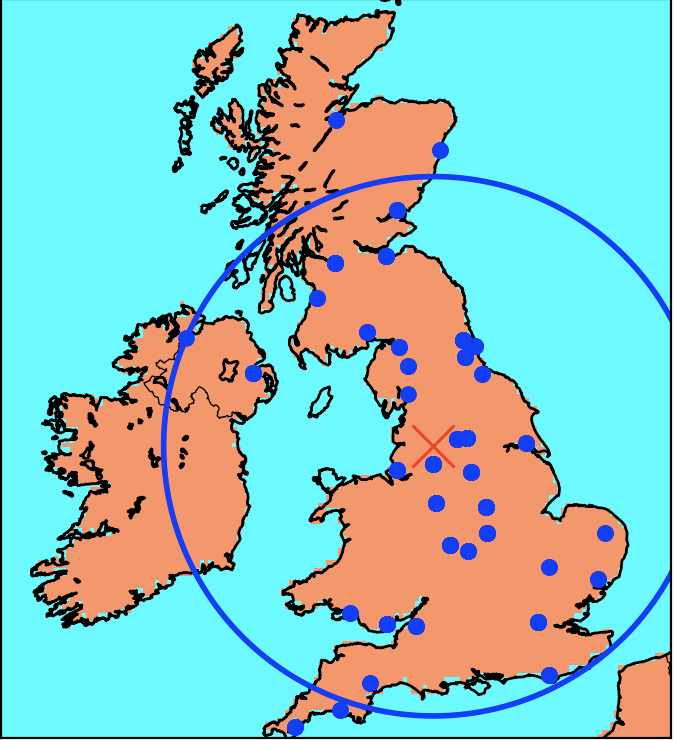


Figure.2 Chips distribution (Retains the nearest 90% of points from the center)

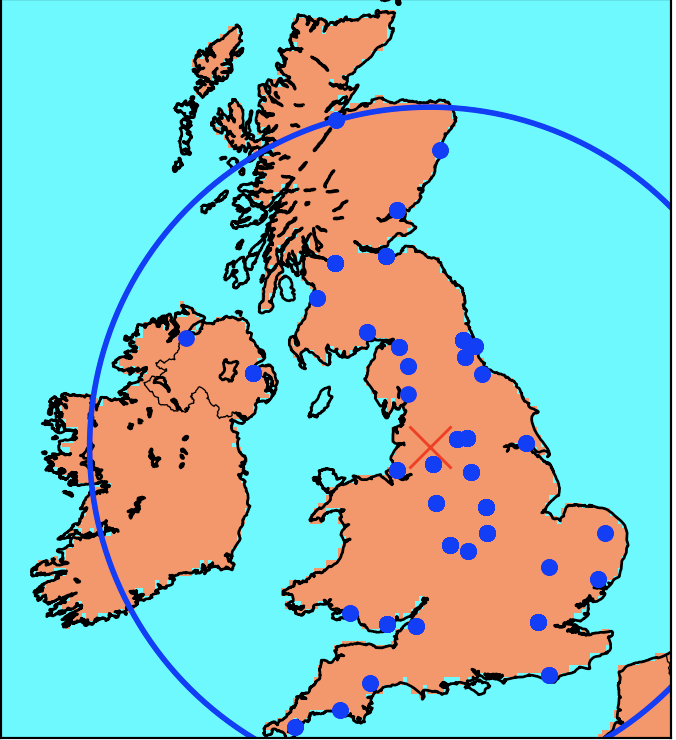


Figure.4 Chips distribution (Retains all cities)

1. Determine if all words are regional or not: Use radius to judge.

Evaluation: This method can just tell the word is a regional word or not. It cannot say the precise area which the word distributed.

1. Use Anaconda to manage Python libraries to solve the conflicts of packages.
2. Use NLP to filter useless words in the HTML: Replace regular expressions with methods provided by NLP.

Evaluation:

Use regular expressions will miss a lot of useless words.

Use nltk package can filter words based on part of speech.

Still need to filter, because some word like ‘vary’ and ‘try’ are identified as nouns.