**Word pair**

The way to generate a word pair is to combine HTMLPaser and spaCy which is industrial-strength natural language processing tool. SpaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. Independent research has confirmed that spaCy is the fastest in the world.

The project firstly used HTMLPaser to filter some useless content in the HTML page such as content in ‘script’ tag and content in ‘header’. Next, the project used HTMLPaser to extract useful content in the HTML ‘body’ tags. The spaCy will regard these useful contents as sentences and extract noun phrases through built-in methods.

**Word pair similarity**

If the project only based on the equality of string to judge whether different word pairs have the same meaning, the project will divide many word pairs with the same meaning into different categories. For example, ‘irn bru bottle l’ and ‘irnbru’ are the same drink, but they are different strings. Thus, the project planned to use a method which can judge the semantic similarity of two noun phrases and category word pairs which have the same meaning. In spaCy, there is a function which can calculate the similarity of two-word pairs. The project made a testing experiment which aims to test whether the similarity can correctly classify words. The project tested the similarity between ‘mini southern fried chicken wrap’ and ‘chicken wrap’, ‘sweet chili chicken wrap’ and ‘chicken wrap’, and ‘a inch chicken tikka’ and ‘chicken wrap’. The similarities between these words and ‘chicken wrap’ are 0.72, 0.84 and 0.78. Obviously, ‘mini southern fried chicken wrap’, ‘sweet chili chicken wrap’ are ‘chicken wrap’, but ‘a inch chicken tikka’ and ‘chicken wrap’ are totally different dishes. Thus, if the project uses similarity to continue classify noun phases, many words that are not in the same category will be classified into same category. As a consequence, the project planned to further filter the noun phrases, such as filtering adjectives and articles in noun phrases. The project firstly used the method provided by spaCy to identify part of speech of each word in the noun phase and then filtered non-nominal words.

However, the project has generated 15551-word pairs, and this makes the comparison inefficient. Thus, the project plan to improve the calculating similarity algorithm.