**FINAL PROPOSAL PYTHON PROJECT**  
  
**Introduction**  
Our project proposal is for a programme in which people can enter a Dutch text to check whether they use a varied language. If the word frequency of a certain word is higher than 1%, then the programme will suggest a list of synonyms for that word. Since this is more useful for adjectives and nouns, we will only focus on these word classes.  
   
**Problem analysis**     
***Input***

* **Data requirements**: We used OpenDutchWordnet for synonyms. We thought about combining it with e.g. synoniemen.net, but for copyright reasons that isn’t possible. Therefore, we just used OpenDutchWordnet, because it contains a lot of data.
* **User input**: People need to enter an URL from the Femma website in the programme for it to work. Since SublimeText does not work with User Input, you have to write the URL into the code.

***Output***  
The programme calculates the word frequency and when the percentage is too high, the programme gives a list of synonyms for each high-frequent word. (To know what percentage we use for this, we have first manually calculated the word frequency of a badly-written text and looked at the percentage of high-frequent words in that text. We then used this percentage (1%) in our programme.)

   
***Processing:  steps involved in going from input to output.***

After receiving the URL, we scrape this website for the text we need to test.

Then we need to process the text using TreeTagger to lemmatise and tag the text.

From the output of the TreeTagger, we then need to exclude common words - such as articles, particles, adverbs - from this list. The programme will only keep the word classes adj, nounsg, nounpl and nounprop.

After that, the remaining list of words is used in the programme to calculate the word frequency of each of these words.

The programme then sees which words have a high percentage and looks for synonyms for these words in the inputted data (OpenDutchWordnet).

**Example**  
People can enter a URL of the Blog page on the Femma website with a text they wrote themselves and the programme will read the text, calculate the word frequency for every word in the text and suggest a list of synonyms for words with a high word frequency. The user is free to make use of these synonyms or not, these are just suggestions.