Project Semester 1:

Estimating a Language Model to Generate Wine Reviews

Joy-Anne FOSTER
Student Number: 22203455
Shami THIRION SEN (née SEN)
Student Number: 22200036

Corpus Linguistics in English
Armand STRICKER

The aim of this project is to build an algorithm that generates reviews from a language model using the probabilities of a word given the two words before it. These probabilities can be calculated by the Bayes theorem which can be explained with the following formula:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Thus, we can calculate the probability a word from the two previous words with the following logic: P(A|B) would represent the probability of knowing a word given the words before it. We then need to multiply the probabilities of A and that of B knowing A , and divide this result by P(B) in order to calculate the probability of knowing a *wi* given the two words prior (*w1*, *w2*). Here we have used wine2.txt to study probability of a what precedes and follows a word.

1.a) P(wi | wi-2, wi-1) = count (wi , wi-2, wi-1) / count (wi-2, wi-1)

In order to calculate the probability of a word considering the two words before it, we need two words "Begin" and "Now" at the beginning of each sentence ; this will allow the language model to predict the probability of which word or words should be at the beginning of the sentence. We can test this hypothesis with the test.txt data. Some of the sentences obtained from the algorithm were the following: "I do not like chocolate pudding ." ; "I like chocolate pudding . " ; "I like chocolate ice-cream . " As can be observed, the sentences obtained are quite satisfactory for this tiny dataset. As all the sentences in the test file begin with "I", the probability that the first word in a sentence would begin with "I" is 1.0, meaning that every sentence produced in the test started with "I".

The difference between the sentences generated with the test data and wine reviews data was noticeable. Sentences generated with the wine reviews data were much longer and were grammatical, however we can see that some sentences lack meaning. An example of a sentence generated by the wine data was: "*Filled with fiji apple flavors giving way to toast , crème brûlée , green fig flavors and cream follow through on the floral-tinged finish has lots to offer gorgeous apple , sliced pear , floral scents to the kirsch and plum aromas give way to orange sherbet , yellow fruits and finds a good rosé for your money ‘s worth of releases to look like a turnaround in recent vintages have been standouts .* " Here, we can observe that the generator accurately makes use of punctuations, capitalisation and grammar, but we can see that as the words are generated randomly and the sentences are quite long, the semantics are lacking for the longer sentences.

To conclude, it can be observed that it is possible to generate grammatical sentences from such algorithms. However, as the sentences are randomly generated by the computer, some of the longer sentences lack meaning. This is representative of a common problem of semantics in Machine Learning, or even Computational Semantics. Computers do not know the inherent meaning of words or their co-occurrence pattern for natural languages, so language generated by the machine can sometimes lack perfection.