ARTIFICIAL INTELLIGENCE

Fall 2018 Mini-Project3 Report

Implementing Probabilistic Language Identification System

First Name	Last Name	Student Id	Email Id
Pulkit	Wadhwa	40082832	Pulkitwadhwa95@gmail.com

TABLE OF CONTENTS

Objective 3

Selection of third language and Basic Setup 3

Result Analysis 4

Experimentation 5

References 7

.

Objective

The main objective of the project is to implement a probabilistic language identification system to identify language of a sentence using character level n-gram models for 3 languages: English, French and a language of our own choice to build two models:

1.Unigram Character based-language model
2.Bigram Character based-language model

Selection of third language and Basic Setup

First step was to choose the third language, the language I chose was German as these 3 languages are closely related to each other in terms, English language core vocabulary and Grammar is all Germanic, while talking about modern English vocabulary it comes from French (a third or half of English Vocabulary can be somewhat intelligent to a French reader and a lot of remainder would be Germanic words).

Therefore after choosing the language the first step involved pre-processing of training corpus it involved various steps:
1.Converting all the characters to lower case
2.Removing digits from sentences
3.Removing punctuation and special characters
4.Removing all the white spaces

Then the processed corpus was used to count the unigram and bigram pairs, for the missing characters in unigram and missing pairs in bigram, add delta smoothing was used with delta=0.5 for computation of results and various other values of delta were used check if there are any effects of smoothing on both the models.

The smoothed count was used to generate the models i.e. calculate the probabilities of occurrence of each character in training corpus and bigram pairs in training corpus. Bigram probabilities for occurrence of a sequence of two characters C1,C2 is count of C1,C2 together divided by count of C1

Bigram_prob = count(C1,C2) /count(C1)

These models are used to generate the probabilities for a sentence in testing corpus, In total 6 different models are generated i.e. each language has its own unigram and bigram model, The language model which gives the maximum probability for a given sentence, sentence is identified to be of that language

Result Analysis

The results were better were better for testing corpus with bigram models as compared to unigram models, as unigram model only takes into account the current character so it gave incorrect identification results for few sentences. For various sentences the probability differences for various languages in unigram model was very less. For e.g. For the sentence "I'm ok" the unigram model predicted this sentence as English and had very close probability to that of German, for the same sentence bigram also predicted the sentence is in English.(Here I have displayed probability upto 3 decimals)

Sentence	Unigram Probability	Bigram Probability
I'm Ok	English= -13.770	English= -9.437
	French= -17.177	French = -13.373
	German= -13.949	German = -11.358

Smoothening didn't had much effect on the identification of language, practically for a large training corpus smoothing for character level unigram is of no use because when we talk of a large training corpus (I used around 30k sentences) almost all the characters are already present and have a large count so there were no differences of smoothening unigram pairs, while for bigram pairs there was a minute differences in probability of unsmoothed pairs and smoothened pairs while the output was correct for both of them this was probably due to no occurrence of missing pairs in test corpus.

For the 10 given sentences almost all the sentences were correctly identified by bigram models except one and for unigrams 3 sentences were incorrectly identified. The models identified language incorrectly for small sentences and for words, which were common in any of the languages, small sentences like "woody Allen Parle" was predicted to be in English by both the models which was in French.

Sentence	Unigram Probability	Bigram Probability
woody Allen Parle	English= -44.382	English= -39.651
	French= -50.386	French = -47.247
	German= -49.863	German = -43.408

However, when we look at the probabilities of both the models, for unigram there is a difference of 6 in probabilities of English and French but for bigram models probability given sentence in English is -39.651 and for French it is -47.247 but probability being larger of English so sentence is classified as English.

For the sentence "cette phrase est en anglais" this French sentence was classified as German by unigram model and correctly classified by bigram model, It was classified correctly by bigram due to ordering of characters i.e. we followed occurrence of sequence of 2 characters as compared to individual character in unigram these characters had more number of occurrences in German as compared to French corpus, but there was not much difference in probability of all the 3 languages.

Sentence	Unigram Probability
cette phrase est en anglais	English= -61.600
	French= -61.605
	German= -61.517

Sentences were classified wrongly when there were common used vocabulary words i.e. words which were common in any of the 3 languages for e.g. words like attention, information, communication, addition, depot etc are common in French and English, for French and German words like appetite, passage, pommes, balance etc are common. In German we use addition, aluminium, atom, parameter etc are common with English these common words affect the character level n-gram models for small sentences even if a sentence is one language but it may be incorrectly identified as of other language due to these common vocabulary words. For e.g. Consider the sentence “Depot is done” this sentence is classified as French by both unigram as well bigram models just because “Depot” is used in French language also in English language for unigram models probability difference was very less, but when we look at bigram models there was difference of 5 in probability between English and French where probability for French was -17.783 and probability for English was -22.621.

Sentence	Unigram Probability	Bigram Probability
Depot Done	English= -24.972	English= -22.621
	French= -24.744	French = -17.783

In German there is a sentence “Gutten appetit” also appetit is used in French, but both the unigrams and bigrams identified it as sentence in French, also the probability was more of English as compared to German and was highest for French

Sentence	Unigram Probability	Bigram Probability
Gutten appetit	English= -34.543	English= -29.094
	French= -33.783	French = -25.791
	German= -35.147	German = -30.507

EXPERIMENTS

For the experiments purpose I also Implemented trigram models, trigrams gave the best result as compared to unigrams and bigrams, trigram had the highest probability difference between the correctly identified language probability and probability for other two languages for almost all of the instances.For e.g. consider the sentence above “cette phrase est en anglais” which was identified as German by the unigram models, had the probability difference of roughly 20 in trigrams as compared to 0.830 in unigram models

Sentence	Unigram Probability	Bigram Probability	Trigram Probability
cette phrase est en anglais	English= -61.600	English= -57.177	English= -70.919
	French= -61.605	French= -55.812	French= -65.724
	German= -61.517	German= -61.413	German= -85.293

Smoothening didn’t produce much effect on language identification except small differences in probability between unsmoothed and smoothed models. I changed the value of delta between 0.1 to 1.0 this didn’t produce much effect on output except small change in probability for each model.

While talking about common vocabulary words trigram models gave better output for common vocabulary words as compared to bigrams and trigrams. For e.g for the sentence “Addition is done” unigram predicted the sentence as English,bigram predicted the sentence as French, while trigrams gave it as English sentence with greater differences in probability from other two languages

Sentence	Unigram Probability	Bigram Probability	Trigram Probability
Addition is done	English= -37.998	English= -57.818	English= -48.043
	French= -38.0675	French= -55.126	French= -55.645
	German= -38.523	German= -61.542	German= -52.675

But there was a sentence which wasn’t correctly identified by unigrams and bigrams i.e. woody allen parle which was supposed to be in French this sentence wasn’t correctly identified even by trigrams

Sentence	trigram Probability
woody Allen Parle	English= -59.055
	French= -64.316
	German= -70.743

Conclusion:

The main purpose of this mini project was to implement the language identification system and find the language identified for a sentence with test corpus. For character level n-grams there is a increase in the computation time when the n increase but greater the value of n better the results, Unigrams are not efficient in finding the correct language as they don’t take into account sequence of characters, bigrams gave better results as compared to unigrams and trigrams was above unigrams and bigrams in finding the correct language.

REFERENCES

1. http://www.practicalcryptography.com/miscellaneous/machine-learning/tutorial-automatic-language-identification-ngram-b/

2. https://appliedmachinelearning.blog/2017/04/30/language-identification-from-texts-using-bi-gram-model-pythonnltk/

3. https://cs.nyu.edu/courses/spring17/CSCI-UA.0480-009/lecture3-and-half-n-grams.pdf

4. http://www.decontextualize.com/teaching/rwet/n-grams-and-markov-chains/