COMP 6721

Mini Project 3 Report

Language Detection System

By

Basant Singh

(Student ID: 40047623)

1. Introduction :

In the project, I have built and experimented with a probabilistic language identification system to identify the language of a sentence. I have created a model using unigram, bigram, trigram, and quintagram characters to identify the language of a sentence.

To build the model, apart from English and French, I have selected the German language as the third language. The reason to select German as the third language is that all three languages are linguistic siblings and lots of root words are common. So having English, French, and German text to train the models, I thought it would be interesting to check how accurately these models can recognize the language of given test sentences correctly.

For experiment purpose, I have included one more language “Indonesian”. This language has only english characters but it is very dissimilar to other selected languages (German, French and English).

2. Data Analysis & Preprocessing

The text file has been prepared after remove diatrics. The first step is to analyze and preprocess the datasets as mentioned below.

- Using regular expression, I removed other than alphabetic characters. The processed sentences has characters as ‘a’ to’ z’.
- All characters have been converted to lowercase to build the model.
- Dictionary have been created with ngram as key and its count as value

For training set,

english corpora has **1227660** (1220066 + 7594) characters.

French corpora has **886428** (878834 + 7594) characters.

German corpora has **1209772** characters.

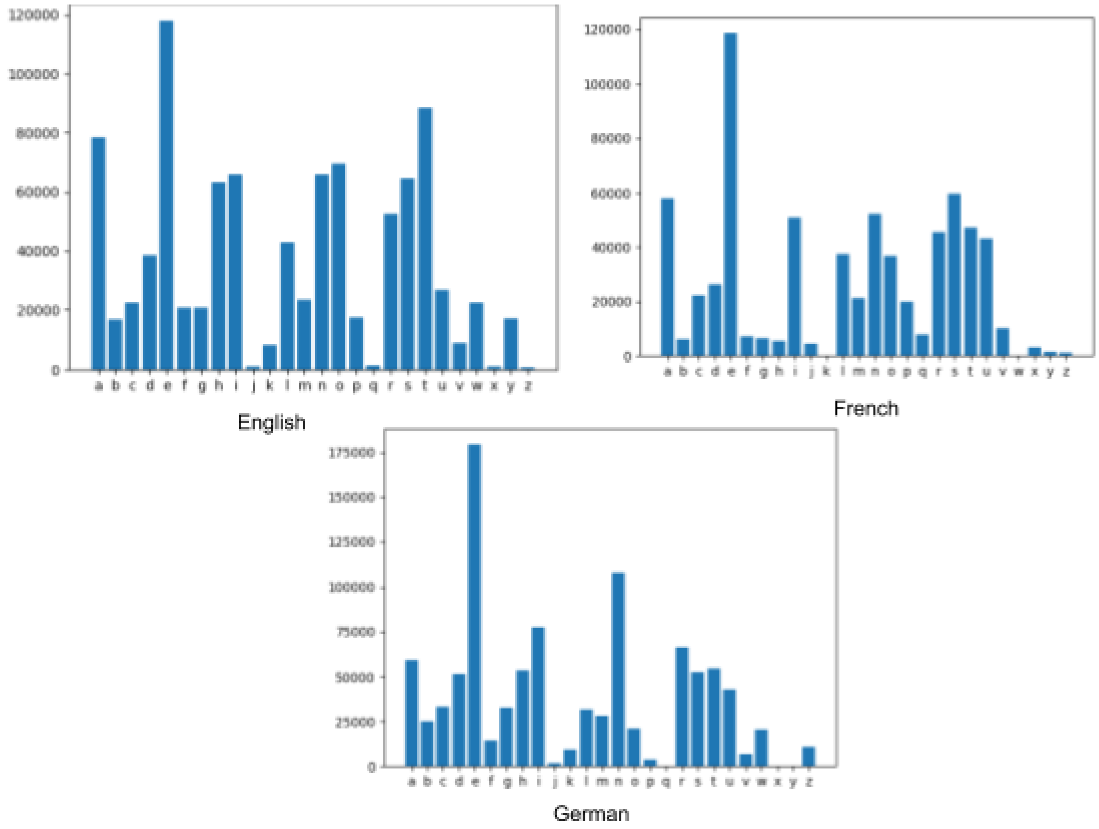
Indonesian corpora has **1122433** characters.

An observation from the plot in below Table 2.1, is that all three language from the text file has a relatively high number of character ‘e’.

For the sake of argument in below sections, I have listed last 4 characters with the lowest probability of occurrence in a corpus.

For English : ‘j’, ‘q’, ‘x’, ‘z’
For French: ‘k’, ‘w’, ‘y’, ‘z’
For German: ‘j’, ’q’, ‘x’, ‘y’

Below shows the frequency of the characters for each language



English

French

German

Table 2.1

3. **Reports and analysis of the results of the basic setup:**

Unigram language models:

It is an extremely simple class of models but for simple tasks such as language identification it can be considered good enough. A unigram language model assumes that each character is generated independently of the other character. The model assumes that the probability of a character does not depend on its position i in the document.

Bigram language models

A bigram language model generates a sequence one character at a time, starting with the first character and then generating each succeeding characters overlapping the previous one.

ngram- language models

As defined in above bigram language model, n gram model is generalization to generates a sequence of a character at a time, starting with the a character and then generating each n-1 succeeding characters overlapping the previous one.

	Unigram			**Bigram**		
Sentences	**English**	**French**	**German**	**English**	**French**	**German**
1.What will the Japanese economy be like next year?	**-53.12650411**	-60.2537536	-61.74366322	**-43.38583302**	-56.08532226	-65.2902261
2.She asked him if he was a student at this school.	**-46.27193363**	-53.79646726	-48.40499656	**-39.9906939**	-53.03757715	-47.89788423
3. I'm OK.	**-5.984116877**	-7.459092532	-6.3494087	**-4.071081291**	-5.772445985	-5.611017702
4. Birds build nests.	-19.40215956	**-19.12200196**	-18.74564818	**-17.61637943**	-19.30249612	-19.66284521
5. I hate AI.	**-7.623197341**	-8.461681343	-7.897635224	**-7.482075455**	-7.946260046	-8.049264242
6. L'oiseau vole.	-13.80781041	**-12.92902785**	-14.9111097	-14.12852771	**-10.59706572**	-13.34392648
7. Woody Allen parle.	**-19.25541976**	-21.89693258	-23.68384937	**-17.00483022**	-19.90554836	-21.65014341
8.Est-ce que l'arbitre est la?	-26.26852205	**-24.34370465**	-27.4223194	-23.28352503	**-20.54067773**	-29.96088946
9.Cette phrase est en anglais.	**-26.73033268**	-26.75280058	-27.28893791	-24.85693028	**-23.81425976**	-29.12891259
10. J'aime l'IA.	-11.31241223	**-10.18082364**	-11.23063701	-8.132993322	**-7.704256948**	-10.23012802

Table 3.1

Above table 3.1, I have 10 sentences, where sentence 1-5 are in English, 6-10 are in French. The tables shows the probability of the sentence for ngram in english, french and german and the probability value has been converted in $\log_{10}$ space. The probabilities highlighted in green/ red are highest. Green highlight represents, the correctly classified sentences and red means incorrectly classified.

For possible 30 case(10 sentence, in 3 languages), only 4 case has been wrongly classified, for rest 26 case it has been correctly classified.

We can state that overall accuracy of bigram is 90% and for unigram it is 70%.

Another observation is that for english sentences, both unigram and bigram model has 100% accuracy. There were incorrect classification of french sentence as to english. One possible reason can be due to the size of training corpus, English corpus is relatively bigger than the french corpus (as stated in section 2)

4. Justification and analysis of the results of my experiments.

Experiment 1 : With different n gram.

In basic set up, bigram performed better than the unigram, so it encouraged me to check the results for the case of trigram and quantagram. As shown in table 4.1 results for trigram and quantgram are even better and there is only one wrong prediction.

So this lead to check me with even higher n gram (i.e. quintagram , heptagram) but results were not good as some of english sentences were getting classified as german.The possible reason for this is because with increase in size of ngram, the probability to find the ngram in the class is lesser so the model may not be able to correct prediction.

	Trigram			**Quatagram**		
Sentences	**English**	**French**	**German**	**English**	**French**	**German**
1. What will the Japanese economy be like next year?	**-59.15950379**	-99.36942186	-100.9299748	**-139.444269**	-187.2823823	-186.4276996
2. She asked him if he was a student at this school.	**-54.0150399**	-88.83655887	-77.18758386	**-124.582517**	-170.8499088	-163.14436
3. I'm OK.	**-4.420082971**	-5.752962147	-6.807194021	**-5.659893392**	-5.659893392	-5.659893392
4 .Birds build nests.	**-23.65816332**	-30.60132467	-28.62647899	**-49.5905641**	-58.92961297	-58.08146868
5. I hate AI.	**-7.474109896**	-10.63209155	-10.89701174	**-17.14004417**	-18.31538861	-18.53230316
6. L'oiseau vole.	-21.83123502	**-14.54912507**	-20.93734797	-38.31715105	**-31.45191494**	-39.3735961
7. Woody Allen parle.	**-22.0531037**	-28.74489997	-33.9977456	**-47.7099931**	-51.89878104	-56.47205551
8. Est-ce que l'arbitre est la?	-34.63450339	**-27.45816542**	-40.78040221	-76.2470903	**-63.45899431**	-85.89256395
9.Cette phrase est en anglais.	-35.67723274	**-33.1641236**	-40.96268213	-82.69739593	**-71.4158513**	-83.27563145
10. J'aime l'IA.	-12.28946089	**-11.15544288**	-14.06127063	-22.01824802	**-21.64591113**	-24.47689739

Table 4.1

So based on observation in table 4.1, it can be suggested that trigram is best for the character based language model predictor.

Another observation is that sentence 7 was incorrectly classified in all models. In order to understand the cause, I checked the probability of characters (unigram) for the words in the sentence. Below are my observations:

1. The probability for the occurrence of 'w' in French is too low when compared to the other two languages, this fact can be easily observed from the frequency table in section 2.
2. For sentence 7, the probability of occurrence of 'o' is high in English when compared to French.
3. However, when I removed 'wood' from the sentence, all 4 ngram models were able to correctly classify the sentence.

In section 2, we have listed four characters with the lowest probability, so for unigram, if the sentence of a language has high characters from this class, its least likely that the model will be able to correctly classify the sentence. Similar, a conclusion can be derived from the above-stated scenario for ngrams.

To extend my experimentation on the model, I randomly selected 25 sentences(for each language). The result was similar and I did not find any interesting fact. One minor observation was that the model is very accurate if the test sentence is large. It can be obviously understood as for small sentences if most words are from rare class, it reduces overall probability of the sentence for being the class (as in the case of sentence 7 from table 3.1 and 4.1). To get more precise accuracy of the model, we can run these models on the large size of test corpus.

Experiment 2: With smoothing value delta

For 30 sentences, below table shows number of missed classification of sentences for their respective language

Delta = 1	**Unigram**	**Bigram**	**Trigram**	**Quatagram**
English	**1**	**0**	**0**	**0**
French	**4**	**2**	**3**	**2**
German	**5**	**1**	**1**	**1**

Table 4.2

Delta = .5	**Unigram**	**Bigram**	**Trigram**	**Quatagram**
English	**1**	**0**	**0**	**0**
French	**4**	**2**	**3**	**3**
German	**5**	**1**	**1**	**1**

Table 4.3

Above table 4.2 and 4.3 shows the number of missed classification of models.
I tried to understand the impact of delta on the model from the data derived from the experiment. However, i could not extract the dependency of the model on delta.

Experiment 3: Using Indonesian language(*bahasa Indonesia*)):

The Indonesian language has English characters but it has it is quite dissimilar to English with respect to the similarity between English and other European languages (French and German). I thought it would be interesting to see how the model is able to recognize the language. For this experiment, I have taken English, French and 'Bahasa Indonesia".

In below table 4.4, I have 15 sentences, where sentence 1-5 are in English, 6-10 are in French and 11-15 are in the Indonesian language. Similar to the table in section 3, I have shown the probability (in log base 10). The probabilities highlighted in green/ red are highest. Green highlight represents, the correctly classified sentences and red means incorrectly classified.

Here we are presenting results for trigram and quantgram based models as earlier experiment shows that trigram and quantum give a better result with respect to uni and bigram models.

	Trigram			**Quantgram**		
Sentences	**English**	**French**	**German**	**English**	**French**	**German**

1. What will the Japanese economy be like next year?	**-59.15950379**	-99.36942186	-99.36942186	**-139.444269**	-187.2823823	-179.8456039
2. She asked him if he was a student at this school.	**-54.0150399**	-88.83655887	-78.81054363	**-124.582517**	-170.8499088	-159.8775144
3. I'm OK.	**-4.420082971**	-5.752962147	-4.627281362	**-5.659893392**	-5.659893392	-5.182935569
4 .Birds build nests.	**-23.65816332**	-30.60132467	-34.49151964	**-49.5905641**	-58.92961297	-63.9689065
5. I hate AI.	**-7.474109896**	-10.63209155	-10.10986926	**-17.14004417**	-18.31538861	-17.66683394
6. L'oiseau vole.	-21.83123502	**-14.54912507**	-23.91743744	-38.31715105	**-31.45191494**	-42.67440568
7. Woody Allen parle.	**-22.0531037**	-28.74489997	-28.57933650	**-47.7099931**	-51.89878104	-56.4336438
8. Est-ce que l'arbitre est la?	-34.63450339	**-27.45816542**	-46.29044905	-76.2470903	**-63.45899431**	-88.56579488
9.Cette phrase est en anglais.	-35.67723274	**-33.1641236**	-39.43579194	-82.69739593	**-71.4158513**	-84.45897287
10. J'aime l'IA.	-12.28946089	**-11.15544288**	-9.36884219	-22.01824802	**-21.64591113**	-18.48186998
11. Artikel ini telah tayang di	-38.08557948	-51.89493757	**-24.99180138**	-87.96134701	-94.29359106	**-62.04361087**
12. Bareskrim Polri Periksa Saksi Terkait Bahar bin Smith di Polda Sumsel	-115.8714713	-143.6289707	**-94.99652545**	-258.8894274	-281.7054683	**-222.5904790**
13. Namun, menurutnya, Bahar bin Smith tidak memiliki kediaman di Palembang.	-361.0696570	-371.2658973	**-223.8071087**	-116.2415532	-138.6405070	**-80.98638097**
14. mencampuri urusan politik suatu negara	-103.7430077	-116.4338086	**-73.91984009**	-231.4132794	-243.6928964	**-173.6960651**
15. Artikel ini telah tayang di Tribunjateng	-132.217631	-144.9200392	**-97.23805610**	-283.553496	-292.1322227	**-224.5109332**

Table 4.4

The result in above shows that there was 100% accuracy in identification of Indonesian language were as there was one instance where a french sentence was classified as English. The reason behind this is obvious that languages are very different from common root-derived languages so there is very less chance of having similar trigram or bigram. For this reason, the model was easily able to distinguish been the Indonesian language with the French or English language.

5. **Conclusion**:

We conclude that smoothed character n-gram language models can work well for especially for bigram and trigram and relatively longer words. In future, I can expand my baseline testing to execute on the large testing corpus.

References:

1. http://aclweb.org/anthology/Q14-1003
2. https://en.wikipedia.org/wiki/N-gram

3. https://sookocheff.com/post/nlp/n-gram-modeling/