Comp 472/6721 **Mini-Project 2 Report** **Fall 2018**

Prof. Leila Kossiem **Suyash Malthankar**

Student ID: 40058018

Project Specification:

Purpose: For this mini project we will be using a machine learning framework to experiment with different machine learning algorithm and different datasets.

Languages chosen:

As per given in the handout I have used:

a) **English**

b) **French**

c) **Italian**

As we are already given English and French as the primary languages to test on, I have chosen Italian as 3rd language. The basic reason is the character alphabets are similar and on the other hand we can test our classifier because there are lot words and even phrases in Italian, French and English those who are same.

[1]So, there it will be interesting to check whether classifier works well or how it will behave upon them. The lexical similarity between French and Italian is around 85-90%. That means that almost 9/10s of the two languages' words are similar but does not mean that they are necessarily mutually intelligible to native speakers due to big differences in pronunciation and syntax.
The Italian alphabet consists of 21 letters (with 5 vowels and 16 consonants) whereas French alphabet consists from 26 letters, as modern English alphabet.

Statistics of Language Characters: -

```
English :
Total characters: 2855913
Counter({'e': 351276, 't': 263988, 'a': 233748, 'o': 207978, 'n': 196851, 'i': 196302, 's': 192693, 'h': 188688, 'r': 156402,
'l': 128379, 'd': 114657, 'u': 80091, 'm': 69831, 'c': 67521, 'w': 66666, 'f': 62499, 'g': 62460, 'p': 51765, 'b': 50631, 'y':
50616, 'v': 25794, 'k': 24177, 'q': 4668, 'j': 3246, 'x': 3090, 'z': 1896})


French :
Total characters: 698472
Counter({'e': 119441, 's': 60189, 'a': 58580, 'n': 52674, 'i': 51204, 't': 47648, 'r': 45700, 'u': 43445, 'l': 37606, 'o': 371
65, 'd': 26544, 'c': 22466, 'm': 21463, 'p': 20215, 'v': 10380, 'q': 8011, 'f': 7127, 'g': 6552, 'b': 6203, 'h': 5480, 'j': 43
58, 'x': 3205, 'y': 1482, 'z': 1048, 'k': 201, 'w': 85})


Italian :
Total characters: 6961766
Counter({'i': 816163, 'e': 794262, 'a': 748072, 'o': 632824, 'n': 503893, 't': 486521, 'r': 459642, 'l': 440846, 's': 358912,
'c': 299675, 'd': 258847, 'p': 207166, 'u': 203717, 'm': 186786, 'g': 128187, 'v': 108608, 'z': 81869, 'f': 75577, 'b': 63798,
'h': 60455, 'q': 26067, 'k': 6583, 'w': 5107, 'y': 3947, 'x': 2580, 'j': 1662})
```

Analysis of the basic setup:

Sentences	Unigram	Bigram
What will the Japanese economy be like next year?	English	English
She asked him if he was a student at this school.	English	English
I'm OK.	English	English
Birds build nests.	**French**	English
I hate AI.	English	English
L'oiseau vole.	**Italian**	French
Woody Allen parle.	**English**	**English**
Est-ce que l'arbitre est la?	French	French
Cette phrase est en anglais.	**English**	French
J'aime l'IA.	French	**Italian**

(fig. a)

The wrong classifications are highlighted in bold letters. There are several factors affecting the classification, sometimes it depends upon the characters used because in some languages some characters are used often so their probabilities will be more than the occurrence in actual language, in the case of unigram model. If we talk about bigram then it can be more accurate because it is taking two characters into account so that it makes sure to get better results every time. Conditional probability may lead to better performance.

Taking one sentence from the above table into account, “Birds build nests.”, it is indeed English

For unigram model:

French: - log(probability) = -19.06

English: - log(probability) = -19.36

This is how it classifies it as French, but it is very close to English as well and here there are so many characters involved in the sentence formation those who have more probability to occur in French label. Characters like “s”, “b”, “u”, “l”, “n”, “r” have the greater probability to occur in French than English.

For bigram model:

According to bigram model, it is been classified as English. Since bigram model is taking two characters’ conditional probabilities at a time and that’s why it is performing well as compared to unigram model.

Another example of, “Woody Allen parle.” This should go to French but according to both the models classified as English. Here Woody Allen would be a name and it will be having probabilities according to the occurrence in the language models but the word which can make a difference would be “parle” because it occurs in English and French as well.

For unigram model:

French: - log(probability) = -21.09

English: - log(probability) = -19.19

This is how it classifies it as English, but it is French actually. The characters like “w” and “y” are having very less probabilities to occur in French label. These probabilities made a difference, since probabilities of w and y in French are -3.91 and -2.67 approx respectively as compared to probabilities in English as -1.63 and -1.75 approx respectively, which is a very low score.

Analysis of the other sentences:

As per requirements, other 20 sentences those who are not classified correctly (10) and classified correctly (10).

First, analysis of 10 false classified sentences.

Sentences	Unigram	Bigram
Bonjour Hi.	**English**	**French**
è una persona borghese.	**English**	Italian
I'm stable.	**French**	English
That guy is a chauffeur	English	**French**
Buona notte.	**English**	Italian
Leila Kossiem is our pr ofessor for Artificial Intelligence class.	**Italian**	**Italian**
Not all treasure is sil ver and gold, mate.	**Italian**	English
D'accord.	**Italian**	**Italian**
Parlo français.	**Italian**	**French**
Pardon me.	**Italian**	**French**

(fig. b)

“Bonjour Hi”, taking one French and one English word common in a phrase, results are as expected since unigram works traditionally to classify on the basis of individual probabilities but in case of bigram it is interesting to see that it classified as French, reason being “bonjour” is a bigger word in length than “hi” and there will be more probability of it going into French.

For “D’accord”, this should go into French without any doubt, but this goes to Italian in both models. Since “d”, “a”, “c”, “o”, “r” all the characters have higher probabilities for Italian than French in unigram model.

Bigram model also classifies it as Italian reason being,

French: - sum(log(probability)) = -7.01 (0.97+1.33+1.73+0.65+1.02+1.31) approx.

Italian: - sum(log(probability)) = -6.39 (0.84+1.26+1.23+0.51+0.94+1.61) approx.

That’s very close call actually but as per the numbers bigram model has to go with Italian.

Correctly classified 10 sentences by both models.

Sentences	Unigram	Bigram
come ti chiami?	Italian	Italian
Amo l'intelligenza artificiale.	Italian	Italian
Io non pretendo di essere un capitano strano. Io faccio solo quello che faccio.	Italian	Italian
I appreciate this genre	English	English
En route.	French	French
Menu in the restaurant is huge.	English	English
Come sta?	Italian	Italian
Hessam Amini è assistente alla didattica della nostra conferenza sull'intelligenza artificiale.	Italian	Italian
On the table.	English	English
Je ne prétends pas être capitaine bizarre. Je fais juste ce que je fais.	French	French

(fig. c)

Experimentations and Analyses:

Delta Smoothing:

Taking given first 10 sentences, putting delta = 0.1,

All the results were same as above first table.

When delta smoothing changed to 0.9 then all the results seem same as the results of the first table (fig. a) above. It proves that delta smoothing was not much of impact on the previous results.

When delta = 0, then mathematically, log (0) = -infinity, which clearly disqualifies the language from the race, which doesn't contain those characters.

n-gram Variations: -

Taking value of n =3 as it makes a tri-gram model.

Sentences	Unigram	Bigram	Trigram
What will the Japanese economy be like next year?	English	English	English
She asked him if he was a student at this school.	English	English	English
I'm OK.	English	English	**Italian**
Birds build nests	**French**	English	English

I hate AI.	English	English	English
L'oiseau vole.	**Italian**	French	French
Woody Allen parle	**English**	**English**	French
Est-ce que l'arbi tre est la?	French	French	French
Cette phrase est en anglais.	**English**	French	French
J'aime l'IA.	French	**Italian**	French

(fig. d)

Evaluating the results from the above table we elicit that trigram model worked best for the 10 sentences given, but we must also consider that trigram model classified one of them in the wrong language class. So, we can conclude that it completely depends on the **type of data** we have, **total number of characters** we have for training and **length of the sentence** as well.

Trigram model column shows that it worked well for the long sentences as well as the ones containing names and short words from a language, but it didn't work for the short sentence like "I'm OK." and let us explore why it happened so?

TRIGRAM MODEL:

TRIGRAM : imo

FRENCH: P(o|im) = 0.09530386740331492 ==> log prob of sentence so far: -1.0208894754598916

English: P(o|im) = 0.045454545454545456 ==> log prob of sentence so far: -1.3424226808222062

Other: P(o|im) = 0.17527333894028596 ==> log prob of sentence so far: -0.7562841399912047

TRIGRAM : mok

FRENCH: P(k|mo) = 0.0003117206982543641 ==> log prob of sentence so far: -4.527123835072017

English: P(k|mo) = 0.027777777777777776 ==> log prob of sentence so far: -2.8987251815894934

Other: P(k|mo) = 0.014285714285714285 ==> log prob of sentence so far: -2.6013821800054617

According to the trigram model, the sentence is in Other (Italian)

When we analyse the situation above, we know that model is more leaning towards Italian language because of the huge corpora as compared to English or French. So, it gave more probability into Italian than English, since the results are very close that -2.8987251815894934 for English and -2.6013821800054617 for Italian, hence Italian wins here. Moreover, we can conclude that the trigram model had accuracy of 9/10 for the given sentences which is more than any other model.

Using Different Language:

Here I have used German as the experimentation language. Two electronic text data sets are included as German corpora.

Sentences	Unigram	Bigram
`come ti chiami? (IT)`	**German**	**German**
`Ich liebe künstliche Intelligenz.(GR)`	German	German
`Ich gebe nicht vor, ein merkwürdiger Kapitän zu sein. Ich mache nur was ich tue.`	German	German
`I appreciate this genre`	English	English
`En route.`	French	French
`Menu in the restaurant is huge.`	**German**	English
`Come sta?`	**French**	**French**
`Hessam Amini ist Assistent in unserer Konferenz über künstliche Intelligenz.`	German	German
`On the table.`	English	English
`Je ne prétends pas être capitaine bizarre. Je fais juste ce que je fais.`	French	French

(fig. e)

This fig. (e) is same as fig. (c), where all the results were classified correctly, let's analyze fig. (e).

When sentence was in Italian i.e. "`come ti chiami?`", it classified as German, since only three labels were given English, French and German, so, it bent towards German as per the sum of logarithmic probabilities.

So, as it is for "`Come sta?`", it bent towards French as per the sum of logarithmic probabilities. These outcomes are still okay since we are asking a language which not been labeled then anyhow it has to go with some language.

It went wrong only once evaluating "`Menu in the restaurant is huge.`", since it contains characters like "n", "u" those who are having higher probability to occur under label "German" than "English" or "French" that's why it is been classified as German.

References: -

[1] http://tsarexperience.com/how-different-or-similar-are-french-and-italian/
[2] http://computational-linguistics-class.org/assignment5.html