Name: Anvesh

Id: 11116468

**1st Answer**)

NO all documents are not written in one language. In Spanish documents some amount of data was in English.

**2nd Answer)**

In my program 225 tokens is the minimum number of tokens need to process to always make the right prediction for all the input test files. After 75 tokens I am getting right prediction for major test files but some are giving wrong predictions but at 225 tokens all are giving right predictions.

**3rd Answer)**

If we have different models with different training data for English and Spanish. Then take one input file from English. Find probability that documents belongs to English by using different models. The model which gives maximum probability that it belongs to English and if a spanish document was given as input it should give least probability that it belongs to English among remaining once , that one is the best model.

Example

En\_News.txt is an English document.

Sp\_News.txt is a Spanish document.

Model 1 (English trained data):

Probability that En\_News.txt belongs to English is 0.4

Probability that Sp\_News.txt belongs to Spanish is 0.3

Model 2 (English trained data):

Probability that En\_News.txt belongs to English is 0.8

Probability that Sp\_News.txt belongs to Spanish is 0.2

Model 3 (English trained data):

Probability that En\_News.txt belongs to English is 0.6

Probability that Sp\_News.txt belongs to Spanish is 0.3

In this case Model2 is best Model among all. It has given highest probability that En\_News belongs to English and least probability that SP\_News it belongs to English.

**4th Answer)**

In graph I have taken a “news1.txt” as input test file and X axis show training size (Number of tokens taken during training) In Y axis it shows difference between log values of probabilities that with how much confidence we predict that the given text file is belongs to particular model.

**5th Answer)**

I am getting French documents as Spanish if I removed some preprocessing techniques like punctuation marks. Otherwise I am getting it as English why because the document contains some symbols and preprocessor thinking it as punctuation and removing all those symbols.

**Challenges faced**:

* Understanding the problem and how to solve it.
* Choosing data structure for model
* Accessing the elements of bigrams, trigrams
* Knowing details about python predefined libraries

**Bug Fixes:**

* By observing output probability values of different files
* By tracking values stored in bigrams, trigrams

**Results:**

Got list of files which belong to English and files which belong to Spanish. By changing length of training data the results were getting changed. After some threshold range the training data length does not effect the output. The more the input, more the training data we get better results.