Assignment 3:

KRISHNA SANAGAVARAPU

Instructions for running the code and screenshots:

- Navigate to the folder language_detector/ in the submission zip:
- language_detetor_1.py is the code that implements bigrams language model.
- language_detector.py is the code that implements bigram language model with snowball stemming.
- language_detector_tri.py is the code that implements tri gram language model.
- language_detector_tri_stem.py is the code that implements tri gram language model with snowball stemming.

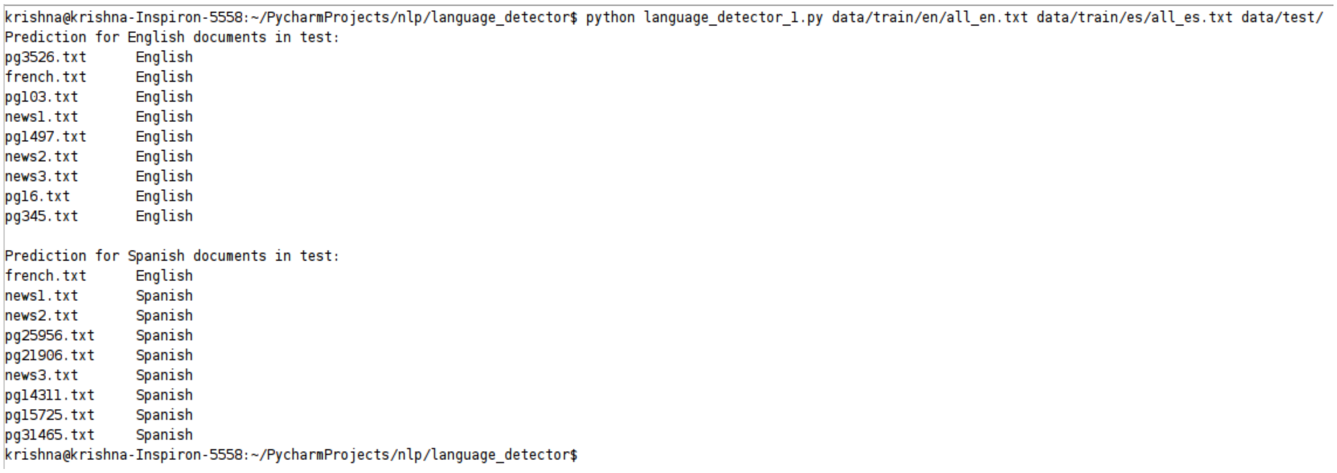
Requirements to run the code: python, matplotlib and nltk

Preprocessing: The entire data while training and testing is preprocessed by lowercasing, eliminating special characters and adding $ to start and end of each token.

Screen shots:

1) language model with bigrams:

- The bigram model implemented with Laplace smoothing predicts all English documents as English and Spanish documents as Spanish.
- This model identifies French document as English.
    - This could be due to bigrams in French being similar to the bigrams in English rather than Spanish



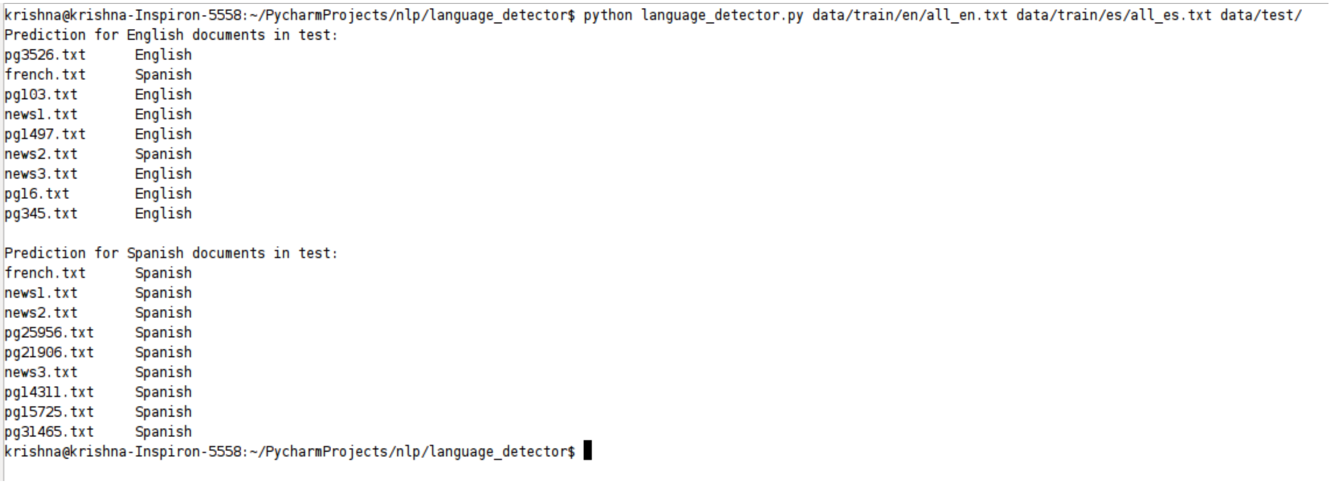
```
krishna@krishna-Inspiron-5558:~/PycharmProjects/nlp/language_detector$ python language_detector_1.py data/train/en/all_en.txt data/train/es/all_es.txt data/test/
Prediction for English documents in test:
pg3526.txt      English
french.txt      English
pg103.txt       English
news1.txt       English
pg1497.txt      English
news2.txt       English
news3.txt       English
pg16.txt        English
pg345.txt       English

Prediction for Spanish documents in test:
french.txt      English
news1.txt       Spanish
news2.txt       Spanish
pg25956.txt     Spanish
pg21906.txt     Spanish
news3.txt       Spanish
pg14311.txt     Spanish
pg15725.txt     Spanish
pg31465.txt     Spanish
krishna@krishna-Inspiron-5558:~/PycharmProjects/nlp/language_detector$
```

2)language model with bigrams and stemming:

- The bigram model implemented with Laplace smoothing and stemming predicts 7 out of 8 English documents as English and all Spanish documents as Spanish.

- This model identifies French document as English.
  - This could be due to bigrams in French after stemming are similar to the bigrams in English rather than Spanish
- Snowball stemmer with both English and Spanish is used, but gives poor results

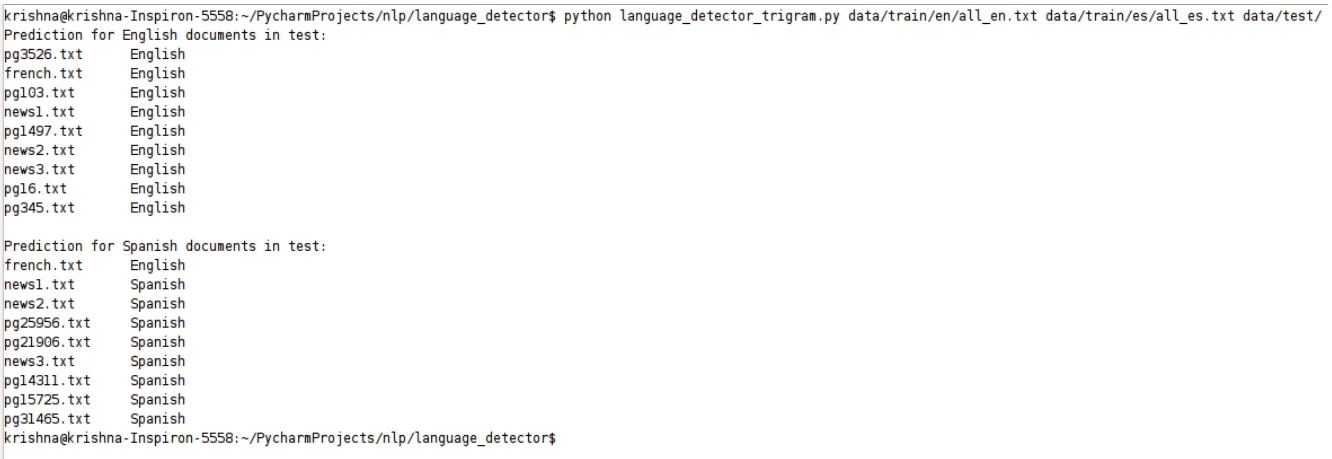


```
krishna@krishna-Inspiron-5558:~/PycharmProjects/nlp/language_detector$ python language_detector.py data/train/en/all_en.txt data/train/es/all_es.txt data/test/
Prediction for English documents in test:
pg3526.txt      English
french.txt      Spanish
pg103.txt       English
news1.txt       English
pg1497.txt      English
news2.txt       Spanish
news3.txt       English
pg16.txt        English
pg345.txt       English

Prediction for Spanish documents in test:
french.txt      Spanish
news1.txt       Spanish
news2.txt       Spanish
pg25956.txt     Spanish
pg21906.txt     Spanish
news3.txt       Spanish
pg14311.txt     Spanish
pg15725.txt     Spanish
pg31465.txt     Spanish
krishna@krishna-Inspiron-5558:~/PycharmProjects/nlp/language_detector$ █
```

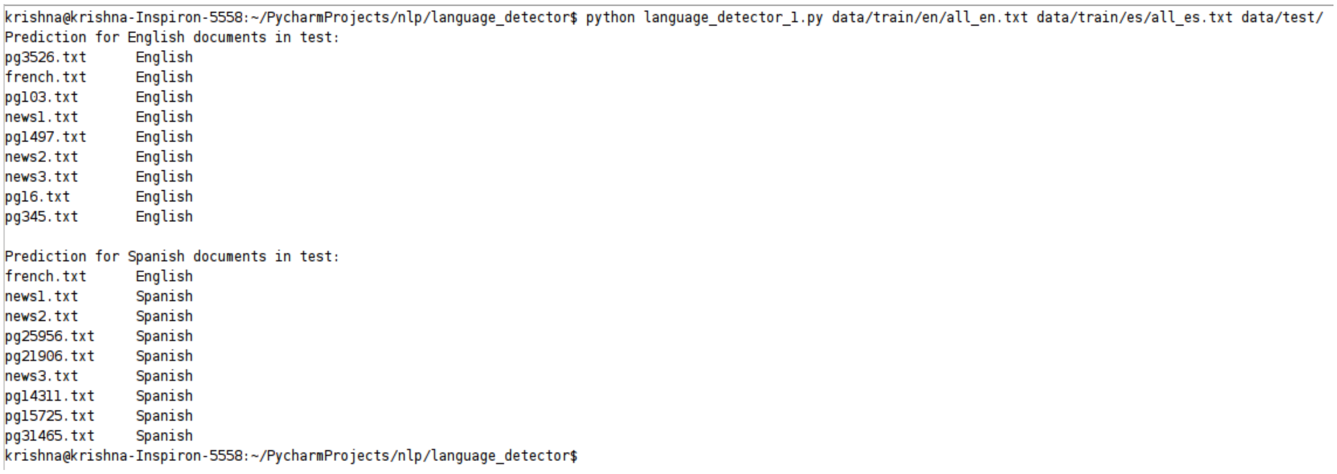
3)language model with trigrams:

- The trigram model implemented with Laplace smoothing predicts all English documents as English and Spanish documents as Spanish.
- This model identifies French document as English.
  - This could be due to bigrams in French being similar to the bigrams in English rather than Spanish



```
krishna@krishna-Inspiron-5558:~/PycharmProjects/nlp/language_detector$ python language_detector_trigram.py data/train/en/all_en.txt data/train/es/all_es.txt data/test/
Prediction for English documents in test:
pg3526.txt      English
french.txt      English
pg103.txt       English
news1.txt       English
pg1497.txt      English
news2.txt       English
news3.txt       English
pg16.txt        English
pg345.txt       English

Prediction for Spanish documents in test:
french.txt      English
news1.txt       Spanish
news2.txt       Spanish
pg25956.txt     Spanish
pg21906.txt     Spanish
news3.txt       Spanish
pg14311.txt     Spanish
pg15725.txt     Spanish
pg31465.txt     Spanish
krishna@krishna-Inspiron-5558:~/PycharmProjects/nlp/language_detector$
```

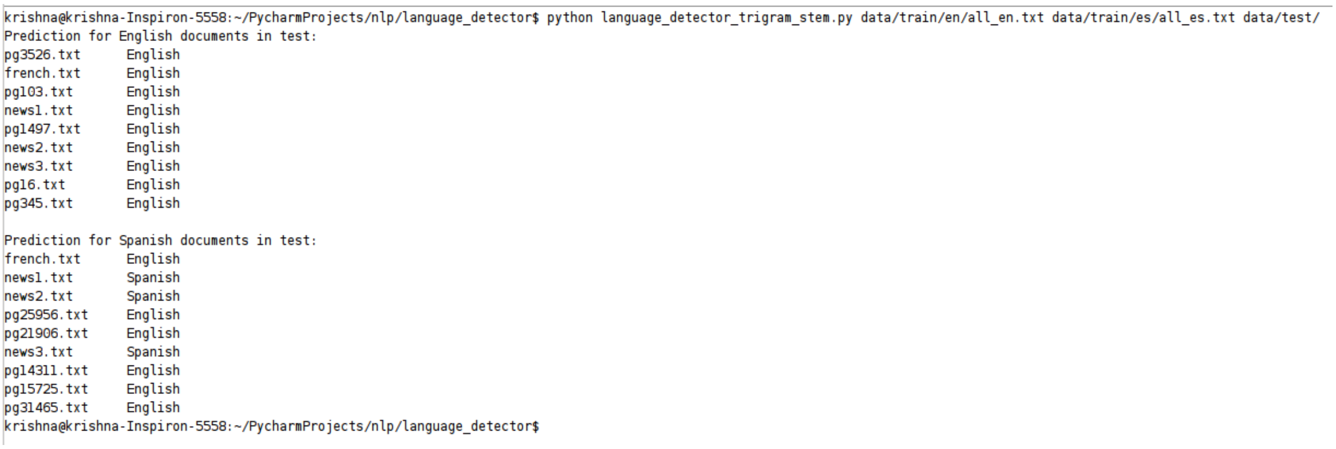


```
krishna@krishna-Inspiron-5558:~/PycharmProjects/nlp/language_detector$ python language_detector_1.py data/train/en/all_en.txt data/train/es/all_es.txt data/test/
Prediction for English documents in test:
pg3526.txt      English
french.txt      English
pg103.txt       English
news1.txt       English
pg1497.txt      English
news2.txt       English
news3.txt       English
pg16.txt        English
pg345.txt       English

Prediction for Spanish documents in test:
french.txt      English
news1.txt       Spanish
news2.txt       Spanish
pg25956.txt     Spanish
pg21906.txt     Spanish
news3.txt       Spanish
pg14311.txt     Spanish
pg15725.txt     Spanish
pg31465.txt     Spanish
krishna@krishna-Inspiron-5558:~/PycharmProjects/nlp/language_detector$
```

4)language model with trigrams and stemming:

- The trigram model implemented with Laplace smoothing and stemming predicts all English documents as English and 3 out of 8 Spanish documents as Spanish.
- This model identifies French document as English.
    - This could be due to trigrams in French after stemming are similar to the trigrams in English rather than Spanish
- Snowball stemmer with both English and Spanish is used, but gives poor results



```
krishna@krishna-Inspiron-5558:~/PycharmProjects/nlp/language_detector$ python language_detector_trigram_stem.py data/train/en/all_en.txt data/train/es/all_es.txt data/test/
Prediction for English documents in test:
pg3526.txt      English
french.txt      English
pg103.txt       English
news1.txt       English
pg1497.txt      English
news2.txt       English
news3.txt       English
pg16.txt        English
pg345.txt       English

Prediction for Spanish documents in test:
french.txt      English
news1.txt       Spanish
news2.txt       Spanish
pg25956.txt     English
pg21906.txt     English
news3.txt       Spanish
pg14311.txt     English
pg15725.txt     English
pg31465.txt     English
krishna@krishna-Inspiron-5558:~/PycharmProjects/nlp/language_detector$
```

Observations:

Q: Take a look at the test documents for English and Spanish. Are documents written in only one language?

No, some text in Spanish documents is written in English (copyright metadata). However due to higher Spanish content the models predict the text as Spanish.

Q: What is the minimum number of tokens you need to process to always make the right prediction when testing? You can try with 100 tokens, 200 tokens, etc.; you do not need an exact number.

With the experiments performed, we need about 50 – 60 % of the data always to make a right prediction.

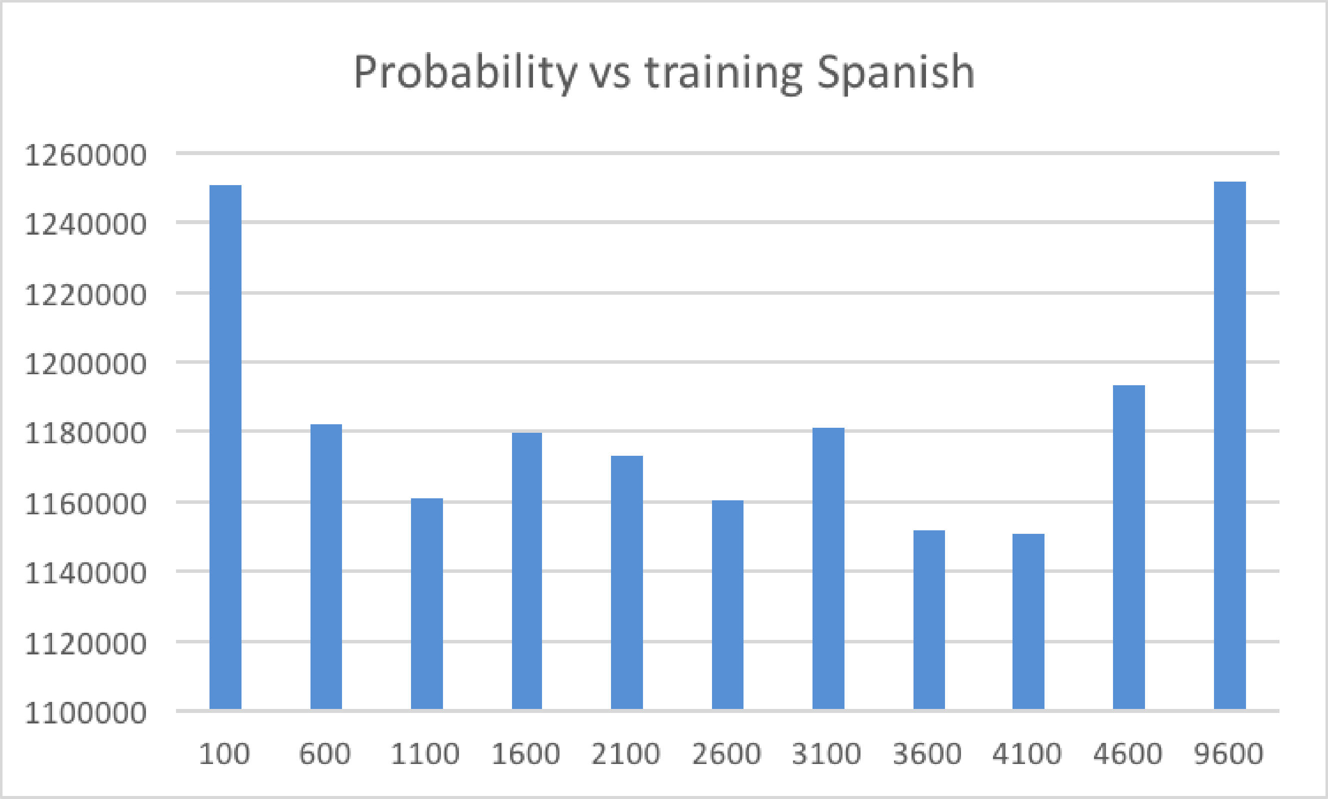
Q: If you create several models for English and Spanish (e.g., using different training data), how can you compare them?

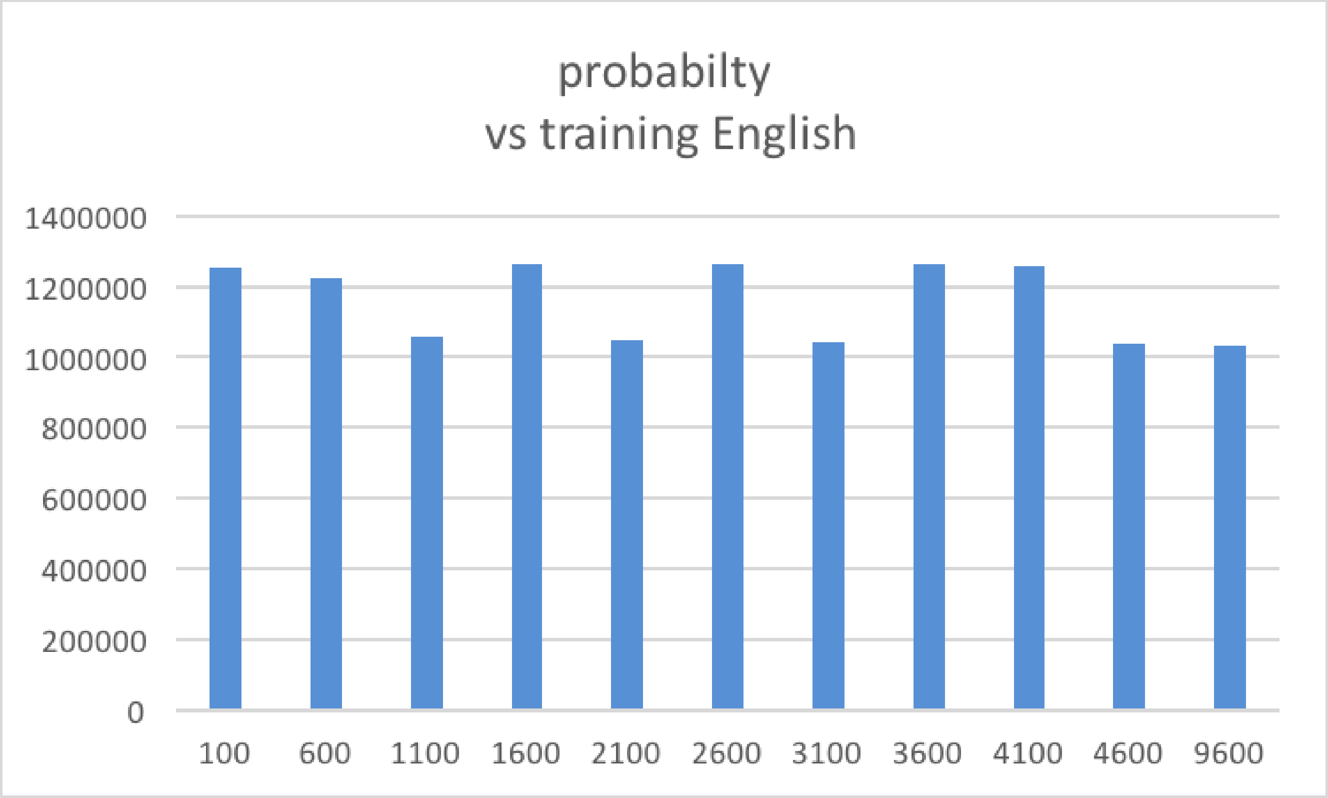
We can compare different models built by seeing the prediction probabilities and comparing the values. We can decide by calculating the difference between the probability values. The bigger the difference the better the model for the predictions. Because if the difference between the probability scores gives an evidence that the model can be used to predict particular language and not for other languages.

Q: Can you train with less training data and still get the right predictions? How does training size affect predictions during testing? A graph showing how probabilities change is the best way to answer this questions (along with an interpretation of the graph).

Graphs:

Below are the graphs plotted for prababilities vs training for both spanish and english models.




probabilty
vs training English
1400000
1200000
1000000
800000
600000
400000
200000
0
100
600
1100
1600
2100
2600
3100
3600
4100
4600
9600