Homework 2 Notebook 1

```
import nltk
from nltk import word_tokenize
from nltk.util import ngrams
nltk.download('punkt')

import pickle

def pickle_dictionary(filename, ngram_dict):
    with open(filename, "wb") as pickle_file:
        pickle.dump(ngram_dict, pickle_file)

def read_file(filename):
    with open(filename, "r+") as input_file:
        input_file_contents = input_file.read()
    return input_file_contents

def get_ngrams(filename):
    file_content = read_file(filename)

    unigrams = word_tokenize(file_content)
    bigrams =  list(ngrams(unigrams, 2))

    print(filename, "word count :", len(unigrams))

    unigram_count = { unigram : unigrams.count(unigram) for unigram in list(set(unigrams)) }
    bigram_count = { bigram : bigrams.count(bigram) for bigram in list(set(bigrams)) }

    return (unigram_count, bigram_count)

def main():
    train_english_file = "LangId.train.English.txt"
    train_french_file = "LangId.train.French.txt"
    train_italian_file = "LangId.train.Italian.txt"

    print("\n")
    en_unigram_count, en_bigram_count = get_ngrams(train_english_file)
    fr_unigram_count, fr_bigram_count = get_ngrams(train_french_file)
    it_unigram_count, it_bigram_count = get_ngrams(train_italian_file)

    pickle_dictionary("english_train_unigram.pkl", en_unigram_count)
    pickle_dictionary("english_train_bigram.pkl", en_bigram_count)
    pickle_dictionary("french_train_unigram.pkl", fr_unigram_count)
    pickle_dictionary("french_train_bigram.pkl", fr_bigram_count)
    pickle_dictionary("italian_train_unigram.pkl", it_unigram_count)
    pickle_dictionary("italian_train_bigram.pkl", it_bigram_count)

main()
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!


LangId.train.English.txt word count : 83309
LangId.train.French.txt word count : 95535
LangId.train.Italian.txt word count : 83803
```