Introduction:

This project analyzes the usage of swear words from those speaking Quebec french. It uses the conversational Quebec french corpus for analysis. In general, swear words in Quebec were derived from their anger against the Catholic Church. As a way to protest them, the people started taking sacred words from the Church and using them as swear words. This project is interested in learning about how often these swear words are used in conversations between Quebec french speaking persons. The list contains Quebec french swear words, as well as traditional french swear words. This is to see if whether Quebec french swear words are more common when compared to tradtional french swear words. The data is also split up by age to further analyze which age groups are cursing more, as well as if the age groups are split in terms of which swear words they're using. Further analysis can be done via sentiment dictonaries or new corpus separations via gender and educational level.

Code:

In [ ]:

```
##from PyPDF2 import PdfFileReader, PdfFileWriter #needed to pip instal PyPDF2

#First give the path to the file, name of file was the specific souscorpus number

##file_path = 'final_project_data/souscorpus#.pdf'
#read in file
##pdf = PdfFileReader(file_path)

#make new txt file
##with open('souscorpus9.txt', 'w', encoding = "utf-8") as f:
    #loop through each page
##    for page_num in range(pdf.numPages):
##        pageObj = pdf.getPage(page_num)

        #get text from each page
##        try:
##            txt = pageObj.extractText()
##        except:
##            pass
##        else:
##            f.write(txt)
##f.close()
#copy code and repeat another 18 times for each souscorpus pdf
#Big thank you Edith and Jenna for sharing this code with me!!!
```

In [ ]:

```
#make list with file names
##filenames = ['souscorpus9.txt', 'souscorpus25.txt', 'souscorpus19.txt']

#create new txt file
##with open('corpus15_25.txt', 'w') as outfile:
```

```
#concatenate each file with write function
##    for fname in filenames:
##        with open(fname) as infile:
##           outfile.write(infile.read())

#repeat 6 more times to get the 6 corpus for each age group
```

In [19]:

```
import re

from collections import Counter, OrderedDict #grab our counter

#setup universal lists,strings
ages= ['15_25','25_35','35_45','45_55', '55_65','65+']
bad_words=['putain','merde','bordel','foutre','pute','salaud','salope','zut','tabarnak','câlice','baptême','sacrament','

# create a new list for only words with no numbers/characters/punctuation
pattern = r'[^a-zA-Z\s]'
data= open("corpus15_25.txt").read()

#initial loop through ages
for age in ages:
    print("For ages "+age+":")
    corpus=age.replace(age,"corpus"+age+".txt") #get text file name
    data=open(corpus).read()
    stripped_corpus = re.sub(pattern,'', data) #replace everything with nothing
    corpus_split = stripped_corpus.split()
    content_word_freq = Counter() #create instance of counter
    for word in corpus_split: #start word loop
        if word in bad_words:
            content_word_freq[word] +=1 #update counter
    print(content_word_freq)
```

```
For age 15_25:
Counter({'ostie': 28, 'marde': 18})
For age 25_35:
Counter({'merde': 3, 'putain': 1, 'ostie': 1})
For age 35_45:
Counter({'ostie': 73, 'marde': 4, 'merde': 1, 'pute': 1})
For age 45_55:
Counter({'ostie': 24, 'marde': 2, 'foutre': 1, 'salaud': 1, 'merde': 1})
For age 55_65:
Counter({'marde': 7, 'ostie': 2})
For age 65+:
Counter({'vidange': 1})
```

Results:

- Swear words were used 180 times throughout the conversations
- Out of the 180 times here is the breakdown:
  - Ostie was used 136 times
  - Marde was used 31 times
  - Merde was used 5 times
  - Vidange was used 4 times
  - Putain, Pute, Salaud, and Foutre were used 1 time
- This table analyzes this usage as percentages

Word	% Used by 15-25	% Used by 25-35	% Used by 35-45	% Used by 45-55	% Used by 55-65	% Used by 65+	% Total Usage
Ostie	22.8%	1.5%	56.6%	17.6%	1.5%	0%	75.6%
Marde	58.0%	0%	12.9%	6.5%	22.6%	0%	17.2%
Merde	0%	60%	20%	20%	0%	0%	2.8%
Vidange	0%	0%	25%	0%	25%	50%	2.2%
Putain	0%	100%	0%	0%	0%	0%	0.55%
Pute	0%	0%	100%	0%	0%	0%	0.55%
Salaud	0%	0%	0%	100%	0%	0%	0.55%
Foutre	0%	0%	0%	100%	0%	0%	0.55%

Observations:

- Conversations from people in the 35-45 as well as the 15-25 age range account for the majority of the swear word usage, both age groups used more swear words than any other group.
- In general people above the age of 45 didn't use swear words too often
- The data also suggests that the Quebec swear words are used more frequently than their traditional french counterparts for every age group

Follow-ups:

- Analyze word usage via gender, and educational level
- Build dictionaries to perform sentiment analysis and can look at counts for sentiments separated by gender, educational level, and age group

In [ ]: