# Accelerated Natura Language Processing

# Assignment-1

1) Preprocess line:

def preprocess\_line(line): #Function for preprocessing and cleansing data from special characters

char\_removed = re.sub('[^a-zA-Z0-9. ]', "", line) #This line removes the characters other than alphabets, numbers , dot and space

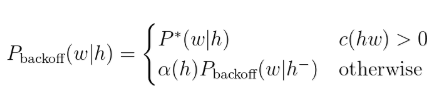
number\_formated = re.sub('[0-9]', "0", char\_removed) #Replaces each digit with zero

string\_lower = number\_formated.lower() #Converts the string from the line above to lower case

2) This probability distribution uses Katz backoff model for trigram probability calculation.

For example: If we take the trigram ' cq', there are no words in english language that start with "cq" and the probability of finding a word with " cq" is zero. As per Katz backoff model, when the highest order n-gram (trigram in this case) is not present, the algorithm goes back to calculate the probability of the bigram " c" in the corpus. So the probability of " cq" is taken as the probability of the bigram " c". By initial validation we have observed from the model file that " cq" and " c " have the same probability. As the probabilities of bigrams are not provided in the model file, we have come to an assumption that since " cq" and " c " are unlikely words in English language, they are given the probability of " c" which is the reason for both trigrams " cq" and " c " to have the same probability (probability of " c" assumed). This has also been observed with other trigrams like " cp" -has the probability of " c"," "," 0[a-z]" - has the probability of " 0" - which are less likely to appear in the language.

Formula for calculation of Probability with Katz backoff model:



3) The model that we have used for calculating the Probability distributions for Language model is Add-alpha smoothing with Maximum Likelihood Estimation (MLE). We have chosen this model as the training data is limited and the test set may have some unknown trigrams which are not encountered in training data. We have chosen Add-alpha smoothing to regularize the language model.

Method used for Language model probability generation:

Step-1: Generate all the possible trigrams with the given vocabulary. The vocabulary is [a-z0.<space>]. These are all stored in the dictionary.

Step-2: Calculate the count of all Trigrams from the training corpus generated after pre-processing of the data. The corresponding count of the trigram is updated in the dictionary created in step-1

Step-3: Calculate the count of all bigrams from the training corpus generated after pre-processing of the data.

Step-4: Calculate the probability of the trigram by using the formula:

probability(trigram)=(count of Trigram+alpha)/(count of bigram + (alpha\*v))

where “v” is the vocabulary (a to z),0,., <space> - 29 characters

and “alpha” is the discount factor which reduces the probability of high frequency trigram by a factor and distributes it to the low/no frequency trigrams.

We have assumed the value of alpha as 0.4 for calculating the probabilities.

Step-5: We have split the corpus into 3 sets for validation of the language model. 80% of the training corpus is used as traning set, 10% of it is used as validation set, the remaining 10% is used as test set.

Step-6: The output is written to a file with system parameter "outfile". The name of the file can be given at runtime.

The output is written in the following format

trigram"\t"probability(in exponential notation)

Ex: esu 1.312e-02

sum 3.312e-02

As per the language model, we have calculated the probabilities for all the trigrams. The trigrams like "ng0" which are not present in the training data, are expected to have a value of alpha/count(ng)+(alpha\*vocabulary). This result is as expected and it is shown in the ng<> probability values below:

ng 8.049e-01

ng. 2.625e-02

ng0 5.148e-04

nga 3.089e-03

ngb 5.148e-04

ngc 5.148e-04

ngd 4.376e-03

nge 8.674e-02

ngf 1.802e-03

ngg 5.148e-04

ngh 5.148e-04

ngi 1.802e-03

ngj 5.148e-04

ngk 5.148e-04

ngl 3.089e-03

ngm 5.148e-04

ngn 1.802e-03

ngo 6.950e-03

ngp 5.148e-04

ngq 5.148e-04

ngr 1.210e-02

ngs 2.111e-02

ngt 1.338e-02

ngu 3.089e-03

ngv 5.148e-04

ngw 5.148e-04

ngx 5.148e-04

ngy 5.148e-04

ngz 5.148e-04

4) generate\_from\_LM is the function to generate random sequences of 300 characters from the language model. The function is coded as follows:

Step-1: Get all the trigrams, probabilities of the training data provided.

Step-2: The string is initialized to ## to get a start of the sequence.

Step-3: All the trigrams having the first two characters of the trigram are identified from the language model and a random trigram from those is chosen using the function np.random.choice. Then the last character of the trigram is taken and appended to the string.

Step-4: The pointer moves to the newly generated character and a trigram is taken from the newly added character which is the new character along with 2 previous characters in the sequence. Then repeat step-3 This loop continues until we reach 300 characters.

Step-5: When we encounter a trigram from the sequence where there is no match in the dictionary, we are using space (“ “) to continue the sequence.

Step-6: # appended from Step-2 are removed.

Random sequence for English language model:

mr me for is and oving the pay guld hounal meent beluding if so di witur votiong itice ocation schnotion parlichissmend fort unt the as preporitiodis re progy com oupposages.# a mississibut istratir ke usimuslat th plaresinal eu.# pork of of to knot unds ing ackin a veralmons nor prom nerred and t

With model-br.en as input:

lice.# und that for.# vor this to you gon hund to itticup abbirror facke bunce.# vallo you there.# up.# ellown.# ge.# eys.# fee this the ithery pullook.# up.# ze.# dow may thats bight chy wanny.# right a wand.# his tucks that you sme ther flit.# uppin ch has going.# whour my at i kno you falloores s

The sequence generated from the model file has a proper start with ## and end with #. With the 30 characters [a-z0.# ] that are used in the model file, it contains 98% of the possible trigram combinations. So it provides better language due to the availability of data.

5) We have observed that the perplexity calculated for test file with the 3 language models generated as per training data is as follows:

Perplexity of test file given:

English language model: 9.361087775314987

Spanish language model: 25.49567901945792

German language model: 25.90632067379816

The perplexity of the test file with English language model is lower when compared to Spanish and German. So, we can safely assume that the test file provided is in English language because of the higher probability compared to the other two language models.

If the language model is checked for any other test file, we can compare the perplexity of the test files with the language model generated from training data for 3 languages and we can conclude the language of the file where there is lower perplexity.

While this is often not the case, as we created a sample file in English and calculated the perplexities with all the 3 language models and observed that the perplexity lies in the same limits. Though the file is more likely to be in English language due to the lower perplexity, but when there are common words between the language models then it becomes difficult to assume.

Test file used:

*“A great way to learn Spanish vocabulary is by reading texts, stories or articles that are completely in the language. That is why we have written are own short reading passages in Spanish about different topics.*

*Remember not to worry about trying to understand all of the details and the grammar rules that appear. Just try and get the gist or general idea of the text. As your Spanish improves, return to each passage and you will be surprised by just how much more you have learned.*

*Spanish Reading Passages*

*Here is the list of original reading passages that we have created. Below each passage we have also included a list of vocabulary associated with the topic along with a definition of each word.”*

Perplexity of test file in 3 language models:

English: 849.4740892131646

German: 860.7246355703909

Spanish: 865.930479474695