ASSIGNMENT 1

SPRING 2019
BBM 497: INTRODUCTION TO NATURAL LANGUAGE PROCESSING LAB.

Author
burakemreozer@gmail.com

March 21, 2019

1 Introduction

Language Modeling is one of the most important and fundamental areas of natural language processing. In this assignment, we will use the Federalist Papers essays to identify authors and generate sentences similar to the style of these authors.

2 Task 1: Building Language Models

We will build unigram, bigram and trigram language models to predict the probability that a given piece of essay belongs to the work of a particular author.

```
def read_and_tokenize(path, num_list):
    ...
    text_block = f.readline().split( ": ", 1 )[1].lower()
        .replace( ",", "" ).replace( "(", "" ).replace( ")", "" )
        .replace( ";", "" ).replace( ":", "" ) .replace( ".", " <dot> " )
        .replace( "?", " <q> " ).replace( "!", "<ex> " )

        sentences = re.split( r"(?<!^)\s*[.\n]+\s*(?!\$)", text_block )
    ...
```

As seen above, different preprocessing steps (removing punctuation, identifying punctuation, marking markers in lowercase letters) will improve performance and accuracy in the given data set. We will use add-one (Laplace) smoothing to solve the out-of-vocabulary problem. The model will also be used in Task 1 and Task 2.

At the end of this step, we have obtained the necessary models and frequency tables.

3 Task 2: Automatically Generating Essays

In this task, we will generate 2 essays, each essay will have 30 words maximum, for each author using each model (unigram, bigram, trigram) and estimate the probability of the automatically generated essays and compare them.

```
def weighted_random_choice(choices):
    total = sum(probs[w] for w in choices)
    r = random.uniform(0, total)
    border = 0
    for pair in choices:
        pair_prob = probs[pair]
        if border + pair_prob > r:
            return pair, pair_prob
        border += pair_prob
```

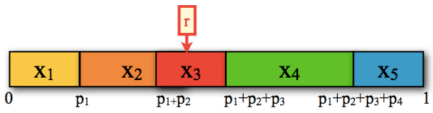

Figure 1: Weighted Random Choice

We use the models of the corresponding author when generating the essays, and also using the weighted random choice algorithm to decide the next chosen word when generating the sentences. You can see the outline of this algorithm in the figure above.

Sentences as probability models

More precisely, we can use n-gram models to derive a probability of the sentence , W, as the joint probability of each individual word in the sentence.

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k | w_{k-N+1}^{k-1}) \tag{1}$$

3.1 Generated Essays with Log Probability

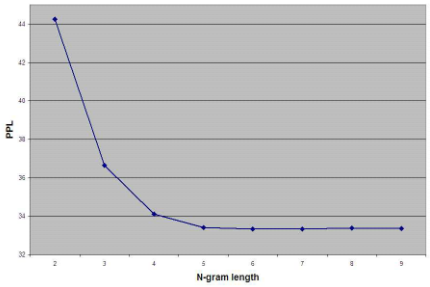

Figure 2: Plot of perplexity against order of n-gram

As expected, the perplexity from the unigram model to the trigram model decreases relative to the log probability.

3.1.1 Unigram

- **Hamilton**

  1. *"<s> apprehensions great the to lot to the </s> conducive guard both negociations where a of to levelled disinclined to numerous stated to in there experiment as inquirers whether"* **-123.83165820149695**

  2. *"preconceived revenue the if the to to true administration spectacle numerous afford of any mere to conjecture it business been keeping in and confederation formation </s> the"* **-118.91624192054488**

- **Madison**

  1. *"value founded were attainder particular encroachments and to carrying practice for affecting institutions the the wonderful that it union traffic the from subject is power of short well"* **-126.55683605815899**

  2. *"of in destroy these and can seems what treaties corruption marque places the its government universal number </s> whole be of which </s> provide material continuance the the"* **-108.83873542507168**

3.1.2 Bigram

- **Hamilton**

  1. *"<s> and stand ready to be maintained in many local institutions had a traitor to perform that plan which did not till it would permit them or reasoning on a"* **-97.63294886498781**

  2. *"<s> thus the result from which if opposition to be satisfactory to the administration of a regular authority kept in the power of friendship and reasons already explained they would"* **-99.83667602645218**

- **Madison**

  1. *"<s> this method of the same in political life of citizens the mean on which the imported manufactures must increase as they themselves an option and all this question <dot>"* **-98.58443649636531**

2. *"<s> the real inconveniency defeats one part of his life from achaeus and the earth and temptation are so many federal than in all possible changes in the states the "* **-105.24703063554844**

3.1.3 Trigram

- **Hamilton**

  1. *"<s> hence also those oppressive expedients for upholding its authority which are requisite in one case at the same energy of government <dot> </s> "* **-91.69488730043867**

  2. *" <s> on one side of questions of the dispute between connecticut and new jersey and rhode island of britain <dot> </s>"* **-80.09029257276642**

- **Madison**

  1. *"<s> although in most other cases there is perhaps no legislative act in which no particle of the preceding power of regulating the times and places of election <dot>"* **-93.5559620081007**

  2. *"<s> either the mode in which some have thought fit to comprehend and pursue great and critical object wholly foreign to their happiness but on the contrary we know that "* **-82.18721496758019**

4 Task 3: Classification and Evaluation

In this chapter, we will see how good our model is in determining the authors' essays by using the models and possibilities that we have obtained in previous chapters.

We will use bigram and trigram language models to estimate the perplexity distinctly for two models.

4.1 Test

Estimating the perplexity of each test essay using our language models, we will decide which essay was written by which author Madison or Hamilton.

Essay numbers to be tested: *49, 50, 51, 52, 53, 54, 55, 56, 57, 62, 63*

As can be seen below, all the test results of the unknown essays published by **MADISON**. The system classified the essays with 100% accuracy. The results were completely consistent and accurate for both model.

```
n-gram: 2
Author: h Perplexity: 5457.263929290164
Author: m Perplexity: 4951.07377616122
49.txt MADISON
########################################
n-gram: 3
Author: h Perplexity: 23132.18141897749
Author: m Perplexity: 22659.699198874336
49.txt MADISON
########################################
n-gram: 2
Author: h Perplexity: 6609.673504216697
Author: m Perplexity: 6105.903826718996
50.txt MADISON
########################################
n-gram: 3
Author: h Perplexity: 24834.31108471674
Author: m Perplexity: 24407.65899269936
50.txt MADISON
########################################
n-gram: 2
Author: h Perplexity: 5486.073061916661
Author: m Perplexity: 4840.931789578895
51.txt MADISON
########################################
...
...
...
```

In this test case, the system classified the actual authors with 100% accuracy for both models.

Essay numbers to be tested: *(9, 11, 12) of Hamilton, (47, 48, 58) of Madison*

```
n-gram: 2
Author: h Perplexity: 6142.177657843541
Author: m Perplexity: 6167.187986698916
9.txt HAMILTON
########################################
n-gram: 3
Author: h Perplexity: 23309.84800894795
Author: m Perplexity: 23860.40446677058
9.txt HAMILTON
########################################
n-gram: 2
Author: h Perplexity: 6721.5772214347635
Author: m Perplexity: 7303.077340578231
11.txt HAMILTON
########################################
n-gram: 3
Author: h Perplexity: 24065.231576882532
Author: m Perplexity: 25605.52679159736
11.txt HAMILTON
########################################
n-gram: 2
Author: h Perplexity: 6717.899866964864
Author: m Perplexity: 7238.61085036581
12.txt HAMILTON
########################################
n-gram: 3
Author: h Perplexity: 24159.461836796665
Author: m Perplexity: 25798.409658281485
12.txt HAMILTON
########################################
n-gram: 2
Author: h Perplexity: 6035.298966170517
Author: m Perplexity: 5129.943442350206
47.txt MADISON
########################################
n-gram: 3
Author: h Perplexity: 23955.51284462015
Author: m Perplexity: 21983.18928122665
47.txt MADISON
########################################
n-gram: 2
Author: h Perplexity: 6314.990084219555
Author: m Perplexity: 5673.72325561838
48.txt MADISON
########################################
n-gram: 3
Author: h Perplexity: 24064.45146190676
Author: m Perplexity: 23414.336490737274
48.txt MADISON
########################################
n-gram: 2
Author: h Perplexity: 5911.722150668606
Author: m Perplexity: 5181.755782690249
58.txt MADISON
########################################
n-gram: 3
Author: h Perplexity: 23820.667303802256
Author: m Perplexity: 22615.430948713205
58.txt MADISON
########################################
```