CS 4501: INFORMATION RETRIEVAL

MP2— Statistical Language Model

This assignment is designed to help you get familiar with statistical language models. You will get the basic ideas of maximum likelihood estimation, smoothing, generate text documents from language models, and language model evaluation.

This assignment has in total **115** points, among which 15 points are for the last bonus question. The deadline and submission guidance are explicitly described here.

Data Set

We will use the same data set as used in our MP1. The data set is located at

```
"http://www.cs.virginia.edu/~hw5x/Course/IR2015/_site/docs/codes/data/yelp.zip"
```

In the following discussion, this data set is referred as **train folder**.

A separate (smaller) Yelp review set is provided for this assignment in addition. It can be downloaded from

```
"http://www.cs.virginia.edu/~hw5x/Course/IR2015/_site/docs/codes/data/yelp_test.zip"
```

And this data set will be referred as **test folder** in the following discussions.

Following the same manner as in MP1, we will refer to each individual user review as a **document** (e.g., as in computing document frequency). You should reuse your JSON parser in this assignment.

The same pre-processing steps you have developed in MP1 will be used in this assignment, i.e., tokenization, stemming and normalization. Note: **NO** stopword removal is needed in this assignment.

Statistical Language Models

1 Maximum likelihood estimation for statistical language models with proper smoothing (40pts)

Use all the review documents in the **train folder** to estimate a unigram language model $p(w)$ and two bigram language models (with different smoothing methods specified below). Note those language models are corpus-level models, i.e., aggregating all the words across different documents.

When estimating the bigram language models, using linear interpolation smoothing and absolute discount smoothing based on the unigram language model $p_u(w)$ to get two different bigram language models accordingly, i.e., $p^L(w_i|w_{i-1})$ and $p^A(w_i|w_{i-1})$. In linear interpolation smoothing, set the parameter $\lambda = 0.9$; and in absolute discount smoothing, set the parameter $\delta = 0.1$.

Specifically, when estimating $p^L(w_i|w_{i-1})$ and $p^A(w_i|w_{i-1})$, you should use the unigram language model $p(w_i)$ as the reference language model in smoothing. For example, in linear interpolation smoothing, the resulting smoothing formula looks like this,

$$p^L(w_i|w_{i-1}) = (1-\lambda)\frac{c(w_{i-1}w_i)}{c(w_{i-1})} + \lambda p(w_i)$$

where $c(w_{i-1}w_i)$ is the frequency of bigram $w_{i-1}w_i$ in the whole corpus.

From the resulting two bigram language models, find the top 10 words that are most likely to follow the word "good", i.e., rank the words in a descending order by $p^L(w|\text{“}good\text{”})$ and $p^A(w|\text{“}good\text{”})$ and output the top 10 words. Are those top 10 words the same from these two bigram language models? Explain your observation.

HINT: to reduce space complexity, you do not need to actually maintain a $V \times V$ array to store the counts and probabilities for the bigram language models. You can use a sparse data structure, e.g., hash map, to store the seen words/bigrams, and perform the smoothing on the fly, i.e., evoke some function calls to return the value of $p^L(w|\text{“}good\text{”})$ and $p^A(w|\text{“}good\text{”})$.

What to submit:

1. Paste your implementation of the linear interpolation smoothing and absolute discount smoothing.
2. The top 10 words selected from the corresponding two bigram language models.
3. Your explanation of the observations about the top words under those two bigram language models.

2 Generate text documents from a language model (30pts)

Fixing the document length to 20, generate 10 documents by sampling words from $p(w)$, $p^L(w_i|w_{i-1})$ and $p^A(w_i|w_{i-1})$ respectively.

HINT: you can use $p(w)$ to generate the first word of a document and then *sampling from the corresponding bigram language model when generating from $p^L(w_i|w_{i-1})$ and $p^A(w_i|w_{i-1})$.*

What to submit:

1. Paste your implementation of the sampling procedure from a language model.
2. The 10 documents generated from $p(w)$, $p^L(w_i|w_{i-1})$ and $p^A(w_i|w_{i-1})$ accordingly, and the corresponding likelihood given by the used language model.

3 Language model evaluation (30pts)

Perplexity is an important metric used for evaluating the predictive power of a statistical language model. More detailed discussion about it can be found in this wiki article. In general, lower perplexity indicates better predictive power of a language model over the unseen documents.

Compute perplexity of the previously estimated language models, i.e., $p(w)$, $p^L(w_i|w_{i-1})$ and $p^A(w_i|w_{i-1})$, on all the review documents in the **test folder**.

The perplexity for one testing document is defined below. Follow this definition to compute perplexity for every review document in the test folder and compute the mean and standard deviation of the resulting perplexities.

$$PP(d) = \sqrt[n]{\frac{1}{\prod_{i=1}^{n} p(w_i|w_{i-1},\ldots,w_{i-N+1})}}$$

where $d = w_1, w_2, \ldots, w_n$, i.e., a text sequence in testing document d of length n; and the likelihood is computed under an N-gram language model.

NOTE: to smooth the unseen words in the test folder, use additive smoothing to smooth the unigram language model $p(w)$ by setting the parameter $\delta = 0.1$. Then use the smoothed $\hat{p}_u(w)$ to smooth $p^L(w_i|w_{i-1})$ and $p^A(w_i|w_{i-1})$. The resulting smoothed language models are denoted as $\hat{p}^L(w_i|w_{i-1})$ and

$\hat{p}^A(w_i|w_{i-1})$

What to submit:

1. Paste your implementation of the perplexity calculation of a language model.
2. Report the mean and standard deviation of the perplexities for $\hat{p}(w)$, $\hat{p}^L(w_i|w_{i-1})$ and $\hat{p}^A(w_i|w_{i-1})$ on the test folder.
3. Can you conclude which language model predicts the data in test folder better? And why?

4 Bonus question (15pts)

Repeat Problem 3 with Dirichlet Prior smoothing to get a new bigram language model $\hat{p}^D(w_i|w_{i-1})$, and vary the smoothing parameter μ to verify if we can have an improved perplexity from this smoothing method comparing to your best bigram language model in Problem 3.

Deadlines & How to submit

This assignment is due on **Nov. 6th 11:55pm**. Therefore, you have in total 11 days to finish this MP. The late policy for all our homework has been carefully discussed in the course syllabus.

The collab assignment page has been created for this MP. Please submit your written report strictly following the requirement specified above. The report **must be in PDF** format.

Share on Twitter Share on Facebook Share on Google+

Updated October 26, 2015

Department of Computer Science **University of Virginia**