CS 276 Programming Assignment 2 Project Report

Jiaji Hu, Xuening Liu

1 System Design

1.1 Code Structure

The classes used in this programming assignment are as follows:

Index and *Query* are used to handle indexing and querying, respectively.

BaseIndex is the interface, and *BasicIndex*, *VBIndex* and *GammaIndex* do the encoding and decoding for writing to (reading from) output files.

pairComparator and *postingListComparator* are used to help sort the posting lists.

1.2 Model Building

1.3 Error Correction

2 Methods

2.1 Smoothing for Edit Probabilities

2.2 Smoothing for Language Model

2.2.1 Original method

2.2.2 Extra: Back-off method

[2]

2.2.3 Extra: Bucketing lambda

[1]

3 Candidate Generation

4 Parameter Tuning

Algorithm	index time(s)	index size (MB)	index size - including dicts (MB)	average retrieval time (s)
Basic	67	58	72	1
VB	149	18	33	1
Gamma	178	13	29	1

References

[1] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics, 1996.

[2] Slava Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(3):400–401, 1987.