SNLP 2014 - Exercise 5

**Backing-Off Language Models**

This exercise comes with additional text files containing training data \( D \), validation data \( D_{\text{val}} \) and test data \( D_{\text{test}} \) for both English and Finnish.

1. Implement a count tree to store counts for sequences of arbitrary length.
2. Implement a trigram language model with *add-epsilon smoothing*.
   What is the backing-off distribution here?
3. Implement a trigram language model with *linear discounting*.
   Please use the formula where \( \epsilon \) depends on the history \( h \).
4. Implement a trigram language model with *absolute discounting*.
5. For both languages separately, determine the \( \epsilon \)-parameter for add-epsilon smoothing by tuning perplexity on \( D_{\text{val}} \).
6. For both languages separately, determine the \( d \)-parameter for absolute discounting using the approximate leaving-one-out solution.
7. Interpret the differences in \( d, \epsilon \) between the two languages.
8. Report 3 \( \times \) 2 perplexities for each model and each language using the respective parameters from (5) or (6). Discuss the results briefly.

* When backing-off, if \( w_{i-|h|} \ldots w_{i-1}w \notin D \) make sure that you first shorten \( h \) to \( k < |h| \) such that \( w_{i-k} \ldots w_{i-1}w \in D \) before you continue. Remember that the back-off distribution for the unigram distribution \( p(w) \) is the uniform distribution.

---

**Note on Submission**

Please use PDF as a document format. If you need to compress files, use ZIP or GZIP.

If you attend the **Mon. tutorial**, the **deadline** is **Friday 13th June, 23:59**. In that case send the solutions to **s9tsback@stud.uni-saarland.de**.
If you attend the **Wed. tutorial**, the **deadline** is **Sunday 15th June, 23:59**.
In that case send the solutions to *kalofoli@ceid.upatras.gr*.

Please indicate each group member's name in the filename:

`Ex05-MemberName1(MN)-MemberName2(MN)-MemberName3(MN).zip`