Natural Language Processing: Part II
Overview of Natural Language Processing (L90): Part III/ACS

Exercises for lecture 9

1 Lecture 9

Due to the nature of this lecture, these questions are open-ended, advanced and geared towards thinking about experiments rather than algorithms:

1. Read Lau and Baldwin (2016): `arxiv.org/abs/1607.05368`
2. Lau and Baldwin say (at the beginning of section 3):

   > For all tasks, we split the dataset into 2 partitions: development and test. The development set is used to optimise the hyper-parameters of doc2vec, and results are reported on the test set. We use all documents in the development and test set (and potentially more background documents, where explicitly mentioned) to train doc2vec. Our rationale for this is that the doc2vec training is completely unsupervised, i.e. the model takes only raw text and uses no supervised or annotated information, and thus there is no need to hold out the test data, as it is unlabelled.

   Question: is it justified in general with unsupervised approaches to train on the test data? Explain your reasoning.