Name1: Roei Herzig, id1: 300360310, mail: roeiherzig@mail.tau.ac.il
Name2: Moshe Raboh, id2: 300611878, mail: shikorab@gmail.com

Advanced Methods in NLP – Assignment 3

1.

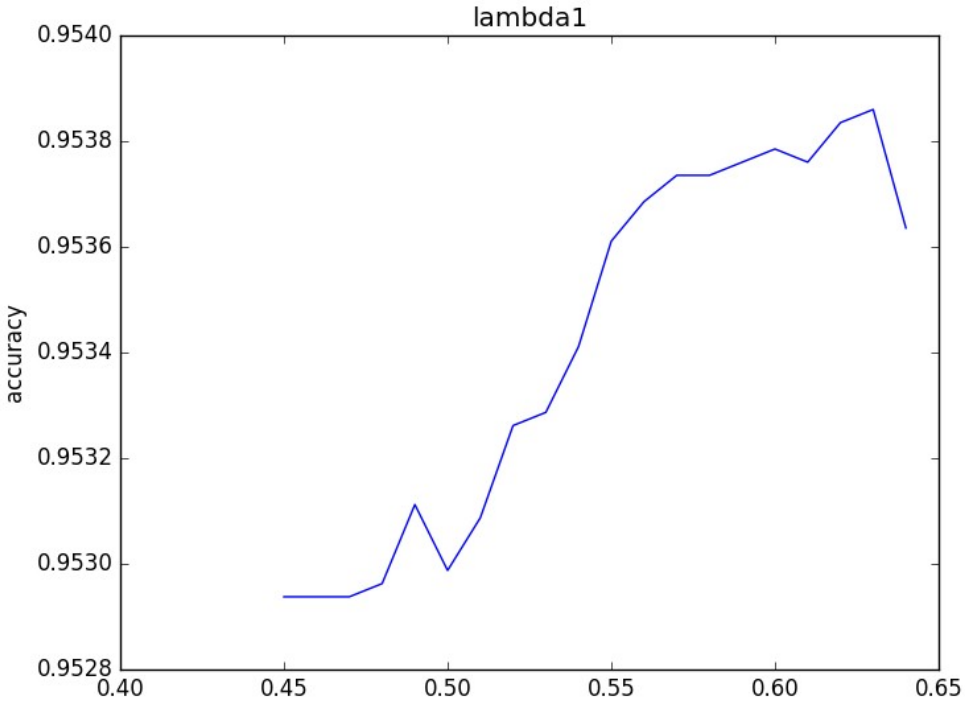
a. Implemented “replace_word” in file **data.py**

2.

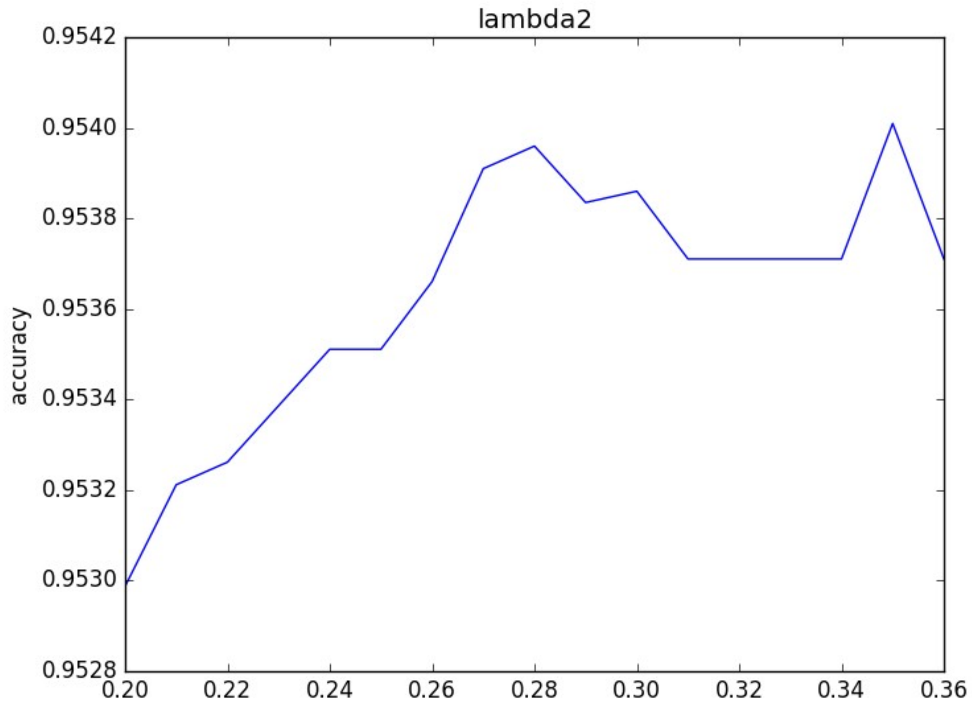
a. Implemented in file **most_frequent.py**
b. Implemented evaluation in file **most_frequent.py**
The accuracy of the development set: 0.9248.

3.

a. Implement “hmm_train” in **hmm.py**
b. Implement “hmm_viterbi” in **hmm.py**
c. Implement “hmm_eval” in **hmm.py**. The accuracy of development set is : 0.954
We preformed a grid search to tune the hyper parameters lambda1 and lambda2. (lambda1 is the coefficient of the trigram, lambda2 is the coefficient of the bigram and lambda3 is the coefficient of the unigram)
(We got lambda1=0.63, lambda2=0.35, lambda3=0.02)



Name1: Roei Herzig, id1: 300360310, mail: roeiherzig@mail.tau.ac.il
Name2: Moshe Raboh, id2: 300611878, mail: shikorab@gmail.com



4.
a. Implemented “extract_features_base” in file memm.py
b. Implemented “memm_greeedy” in file memm.py
c. Implemented “memm_viterbi” in file memm.py
We’ve added the following time optimization to the algorithm:
- Caching:
  The relevant inputs to each prediction in our algorithm are:
  (curr_word, next_word, prev_word, prevprev_word, prev_tag, prevprev_tag).
  We stored the probability prediction for each set of inputs we encountered during the algorithm and reuse it.
- Pruning:
  We added a pruning, skipping any prediction chain with an element with
  probability smaller then 0.1.
  This optimization is of course not exact. However the algorithm ran much faster with this optimization and the affect on the results wasn’t

significant.

d. Implemented in "memm_eval” in memm.py
Greedy accuracy: 0.9615
Viterbi accuracy: 0.9627

e. The common failures in our models were in cases of two popular tags for a given word, especially in cases of potentials ambiguity. In English language verbs might be used as nouns which makes difficult for our model.

For example:
- “ why **limit** the practice”
‘limit’ tagged as VB instead of NN.
- “Average shared outstanding **dropped** to”
‘dropped’ tagged as noun instead as verb.