Baum-Welch Implementation Report

I implemented the Baum Welch algorithm on the untagged Brown corpus, for tagging the tokens with PoS tags.

The tags given were:

- { tag1, tag2, tag3, tag4, tag5, tag6, tag7, tag8, tag9, tag10 }

Initially, A and B were initialised with random start parameters, by using the random library from Python's in-built libraries. The random.random() function outputs a floating no. between 0 and 1. Thus, the values in the A and B matrices in each row, are to be normalised as the probabilities should sum upto 1.

The corpus had 57000 words and a tokeniser was used on the data before training the algorithm.

Observations:

It was observed that on implementing the algorithm and running it on the data, that after each sentence iteration :

- The probabilities for eta obtained were so small, that they resulted in an underflow (were below the lower bounds for floating point ranges = 10^{-320}). *[Zero Division error obtained in python]*
- On implementing the forward algorithm with log values for alpha, we cannot obtain the summation of the (t -1) log values for the t step, as log(x+y) is not ((log x) + (log y)).

Discoveries:

To fix the problem we observed on the *zero error*, we normalised the b matrix such that, the normalisation was done for a given row(tag) with - the sum of probabilities of words for that given tag(row).

The solution to this underflow problem is to scale the numbers.[1] We normalise each $\alpha t(i)$ by dividing by the sum (over j) of $\alpha t(j)$.

$$c_t = \frac{1}{\sum_{j=0}^{N-1} \tilde{\alpha}_t(j)}.$$

$$\hat{\alpha}_t(i) = \frac{\alpha_t(i)}{\sum_{j=0}^{N-1} \alpha_t(j)}$$

From the above description of scale value *ct* , and $\alpha t(j)$, the summation of all alpha is 1 and the values are thus scaled and normalised.

On executing the code, the top 100 words for each tag are(for the best available segregation):

t1: child, roos, aggression, felix, sales, eligible, gone, crises, strongly, offenses, caldwells, under, partys, limited, worth, trenton, supported, teaching, placed, legislators, lao, modest, obtained, risk, aug, violate, aided, controversial, voter, captain, asian, facilities, four, votes, sway, paris, edward, saved, rise, voted, exception, handling, scholarship, every, commented, machines, estimated, likely, harriet, castro, eligio, colonialist, attended, school, construed, budget, 29th, force, enforce, companies, 295, skills, blue, religious, sabbath, expanded, asia, established, endorse, leaders, miller, new, told, charter, simultaneously, direct, increasing, reported, wednesday, here, allowances, spokesman, never, announced, ever, fall, selected, reconstruction, items, elaborate, reports, hyannis, county, decried, schooling, opponents, changed, vicepresident, portugal, fulltime

[1] http://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf

t2: child, wednesday, 1962, 295, aggression, aug, eligible, trenton, paris, facilities, under, partys, teaching, sway, voted, offenses, legislators, risk, rise, supported, placed, four, edward, captain, voter, worth, aided, limited, violate, votes, controversial, saved, caldwells, asian, crises, exception, handling, every, scholarship, commented, selected, estimated, new, reported, fall, charter, force, castro, reconstruction, budget, skills, eligio, religious, spokesman, simultaneously, harriet, here, construed, leaders, asia, allowances, companies, machines, direct, endorse, announced, colonialist, school, ever, increasing, miller, blue, sabbath, expanded, established, told, 29th, roos, attended, enforce, likely, never, items, elaborate, vicepresident, reports, schooling, daughter, highly, military, hyannis, diagnostic, campaign, county, removal, opponents, sanantonio, would, army, confronts

t3: child, voter, votes, caldwells, four, under, partys, confronts, worth, crises, aided, paris, facilities, offenses, placed, legislators, risk, sway, teaching, aggression, supported, limited, saved, voted, aug, trenton, controversial, rise, captain, asian, violate, eligible, edward, exception, handling, skills, commented, budget, religious, attended, blue, selected, companies, school, charter, roos, 29th, direct, miller, new, reported, castro, every, allowances, announced, harriet, established, expanded, ever, endorse, told, scholarship, machines, leaders, likely, here, eligio, spokesman, construed, reconstruction, asia, colonialist, force, 295, enforce, simultaneously, increasing, fall, wednesday, estimated, never, sabbath, items, elaborate, reports, military, county, army, removal, diagnostic, schooling, sanantonio, permit, credit, highly, program, tell, hospital, type, precincts

t4: child, crises, controversial, under, subpenaed, teaching, repay, partys, disapprove, lao, ai, placed, vicious, sway, supported, president, votes, saved, legislators, rise, risk, balance, description, voted, aug, caldwells, aided, paris, trenton, limited, violate, worth, offenses, aggression, captain, edward, eligible, four, facilities, voter, asian, exception, handling, scholarship, every, commented, expanded, 295, castro, machines, direct, sabbath, new, told, roos, companies, wednesday, fall, religious, eligio, colonialist, construed, attended, allowances, increasing, 29th, simultaneously, likely, selected, skills, blue, ever, here, established, enforce, miller, force, harriet, school, reconstruction, spokesman, endorse, charter, budget, leaders, never, reported, announced, asia, estimated, items, elaborate, reports, military, county, removal, arms, schooling, explained, army

t5: child, resolution, known, distance, obtained, aggression, votes, aided, limited, under, sway, crises, placed, edward, controversial, caldwells, voter, worth, captain, aug, legislators, rise, risk, teaching, trenton, four, eligible, saved, offenses, paris, violate, partys, voted, facilities, supported, asian, exception, handling, scholarship, every, commented, endorse, miller, announced, new, ever, selected, increasing, machines, sabbath, 29th, colonialist, leaders, never, direct, charter, skills, school, spokesman, castro, religious, roos, here, asia, established, reconstruction, enforce, wednesday, reported, eligio, attended, likely, construed, simultaneously, estimated, 295, told, force, fall, budget, allowances, blue, expanded, harriet, companies, items, elaborate, reports, fulltime, precincts, launched, military, changes, diagnostic, vicepresident, county, program, recommend, hospital, credit

t6: child, roos, supported, earlier, caldwells, violate, under, worth, sway, votes, voter, legislators, ad, risk, offenses, limited, captain, aggression, placed, facilities, edward, asian, saved, crises, four, partys, eligible, controversial, voted, aug, trenton, rise, paris, teaching, aided, exception, handling, commented, reconstruction, new, fall, eligio, construed, 295, scholarship, charter, sabbath, colonialist, wednesday, allowances, skills, blue, told, attended, every, machines, direct, ever, enforce, established, simultaneously, harriet, companies, budget, religious, selected, likely, spokesman, never, 29th, school, castro, asia, miller, reported, announced, expanded, endorse, increasing, estimated, leaders, here, force, items, elaborate, reports, county, sanantonio, 100, schooling, launched, permit, explained, confronts, program, recommend, campaign, tolerated, decried, forum

t7: child, crises, confronts, eligible, votes, four, under, sway, 1963, teaching, caldwells, trenton, aggression, limited, legislators, risk, placed, aug, rise, offenses, edward, partys, voter, paris, captain, aided, violate, voted, asian, controversial, facilities, worth, saved, supported, exception, handling, scholarship, every, commented, estimated, established, companies, sabbath, miller, new, ever, simultaneously, asia, reconstruction, direct, charter, colonialist, 29th, attended, reported, castro, religious, school, eligio, announced, blue, leaders, never, selected, force, endorse, enforce, increasing, budget, skills, 295, spokesman, construed, allowances, told, fall, wednesday, expanded, roos, here, harriet, machines, likely, items, elaborate, reports, county, sanantonio, opponents, would, obtained, launched, vicepresident, unit, campaign, military, program, tell, portugal, arms

t8: child, asian, supported, sound, edward, under, placed, eligible, votes, trenton, worth, voter, controversial, violate, aided, captain, facilities, paris, caldwells, offenses, aggression, saved, teaching, legislators, rise, partys, voted, crises, risk, aug, limited, four, sway, exception, handling, scholarship, every, commented, expanded, announced, miller, religious, selected, new, ever, established, attended, 29th, castro, force, construed, roos, school, machines, likely, blue, endorse, told, enforce, estimated, leaders, never, wednesday, skills, companies, sabbath, harriet, allowances, asia, colonialist, fall, charter, eligio, reported, here, 295, reconstruction, direct, budget, simultaneously, spokesman, increasing, items, elaborate, reports, permit, schooling, highly, county, program, type, tell, 1000, arms, recommend, confronts, controversy, military, precincts, credit

t9: child, eligio, four, under, 1963, certain, sway, gone, partys, eligible, teaching, vehicles, am, captain, voter, placed, facilities, aug, trenton, paris, legislators, rise, voted, aggression, caldwells, offenses, risk, violate, controversial, edward, limited, votes, supported, asian, saved, aided, worth, crises, exception, handling, every, scholarship, commented, estimated, colonialist, wednesday, skills, religious, new, increasing, castro, budget, here, fall, expanded, construed, harriet, likely, companies, miller, 295, direct, roos, selected, ever, told, established, asia, machines, announced, 29th, school, allowances, simultaneously, force, enforce, endorse, never, spokesman, attended, leaders, charter, sabbath, reconstruction, reported, blue, items, elaborate, reports, changes, schooling, obtained, forum, 100, county, hospital, army, program, type, hyannis

t10: child, edward, trenton, limited, captain, votes, saved, under, partys, teaching, facilities, eligible, violate, four, voted, sway, legislators, risk, rise, controversial, aided, voter, crises, caldwells, aggression, asian, placed, supported, offenses, paris, worth, aug, exception, handling, every, scholarship, commented, miller, construed, wednesday, direct, 295, eligio, charter, new, fall, leaders, never, ever, asia, harriet, roos, likely, increasing, sabbath, enforce, told, endorse, companies, reconstruction, blue, religious, attended, here, skills, simultaneously, 29th, established, expanded, colonialist, castro, budget, allowances, force, spokesman, announced, estimated, reported, school, selected, machines, items, elaborate, reports, county, unit, would, fulltime, program, tolerated, schooling, permit, ore, sanantonio, obtained, changed, opponents, highly, hyannis, confronts

Comments:

- When we were cross-validating the data, we found a stark observation that if we trained for a small corpus, the frequency of words mattered a lot. They were gaining huge probabilities compared to the others and this resulted in all the tags getting these words as their top 100(in some order, not the same). Thus , it was essential that we train it on a bigger corpus. Hence, we incrementally increased the no. of sentences that we were using from the Brown corpus and voila, we were getting varied words in our classes indicating our previous assumption was true.

- 1000_10, we could classify adjectives to some extent.

- Also, a peculiar thing to notice was that if we run the entire corpus for 1 iteration, we got all tags emitting the same words in the 'same order'.

- Our 10-fold cross validation files are given in the same folder, with x_y.txt representing x=corpus size and y=iteration size.

- Also, why didn't we use batch training instead of online training?

  Pretty clear, because the algorithm used for Baum-Welch expects us to converge A and B until convergence for each observation sequence. So using a batch training would in fact not be Baum-Welch at all :)

  Thus the updated A and B values for each observation sequence are used for the next observation sequence until we reach the end of the corpus.