LSA Report

Question 1

1.1

get_unigrams function splits the entire data into lines first and then by spaces and gets tokens.This returns a dictionary with the key as unigrams and value as count of the ungirams.

iunigrams→Input-word output→ index
runigrams→input=index output→word

Top words are considered to be the features and part of the dimesionsal vector.In this the stop words are exclueded.The top 100 frequently occruing words are considered to be stop words.
So dimensional vector is considered from 100:3100.

populate_cmatrix function:-
This function populates each row(correspodning to a unigram) in the cmatrix with the elements of the feature words if they are present in the window size considered to both the left and right of the ungiram considered.
Here first 5 window size is considered.

Then dimensional reduction is done for each row using svd and the feature space is reduced to 5 from 3000.

Then the distance between words is calculated by doing spatial distances like cosine distance.

1.2.1 Results

Window = 2

boy - boy, wood, tenderly, kissed, bonny, corn, hang, ower, green, sought, ane

sunday – sunday, morning, 11th, trinity, readings, pentecost, revelation, 4b-15, 22-27, 26-27, warrick

eat – eat, !, did, just, bit, think, lot, thing, 're, try, saying

good – good, but, way, yet, it, very, them, again, they, saw, got

slowly – slowly, fit, piloting, fact, properly, primarily, anymore, cliffs, people, enough, enhances

100 - 100, =, measurements, centimetre, metric, centimetres, millimetres, decimetres, decimeter, sq., ares

Window = 5

boy – boy, wood, kissed, tenderly, bonny, ower, corn, hang, sought, green, grandson

sunday – sunday, morning, 11th, trinity, readings, pentecost, 26-27, 4b-15, ezekiel, 22-27, warrick

eat – eat, !, just, did, lot, try, think, got, thing, again, 're

good – good, but, them, way, it, very, !, they, make, again, up

slowly – slowly, fact, ?, take, fit, real, try, it, takes, not, reflect

100 - 100, =, measurements, metric, centimetre, centimetres, millimetres, decimetres, metre, decimeter, decametre

Window = 10

Boy - boy, wood, sought, dear, hang, dream, bonny, vi, flood, green, dreamt

sunday - sunday, morning, 11th, trinity, pentecost, readings, 4b-15, 26-27, isaiah, warrick, ezekiel

eat - eat, just, !, lot, fun, think, bit, again, did, but, too

good - good, very, but, it, them, way, yet, well, up, !, there

slowly – slowly, fact, people, fit, feeling, real, take, difficulties, whole, inner, try

100 - 100, =, measurements, centimetre, centimetres, metric, millimetres, decametre, decimetres, decimeter, sq.

1.2.2 Separate Windows

For SVD=100 and considering total dimension size to be 3000 and vocabulory as 1500. So for left window updates given normally between range 0-1500 but for right window updates by adding 1500 to index resulting in range 1500:3000

Window = 2

boy - boy, bonny, kissed, tenderly, hang, wood, robin, sought, ower, green, grandson

sunday - sunday, morning, mark, trinity, readings, pentecost, 11th, ezekiel, revelation, 22-end, 4b-15

eat – eat, saying, just, ride, balls, winds, got, again, ivan, chef, good

good - good, again, yet, very, it, make, them, but, said, saw, got

slowly - slowly, pressure, farm, chances, properly, piloting, inherent, encouraging, primarily, rap, disquisition

100 - 100, =, centimetres, centimetre, millimetres, metric, metre, measurements, decimetres, decimeter, sq.

Window = 5

boy – boy, bonny, tenderly, kissed, wood, hang, robin, sought, ower, green, dear

sunday – sunday, morning, mark, trinity, 11th, readings, ezekiel, pentecost, 22-end, 4b-15, 26-27

eat – eat, just, ride, saying, got, but, again, me, balls, feel, watching

good – good, again, yet, it, very, but, make, them, got, said, well

slowly – slowly, pressure, properly, idea, chances, farm, grace, piloting, kids, fit, inherent

100 - 100, =, centimetre, centimetres, metric, millimetres, metre, measurements, decimetres, decimeter, sq.

Window = 10

boy - boy, bonny, tenderly, hang, kissed, wood, sought, fore, robin, dear, ower

sunday - sunday, morning, mark, trinity, 11th, pentecost, ezekiel, 22-end, readings, 4b-15, 26-27

eat - eat, just, but, saying, got, again, so, me, balls, long, 're

good - good, again, yet, but, it, very, make, well, said, got, a

slowly - slowly, pressure, properly, piloting, chances, farm, necessary, temperature, people, fact, primarily

100 - 100, =, centimetres, centimetre, millimetres, decimetres, decimeter, metric, metre, sq., decametre

1.2.3 Verb Context

Window = 2

boy - boy, sought, daughter, wood, bonny, fore, hang, tenderly, kissed, grandson, ower
sunday - sunday, morning, june, trinity, pentecost, during, romans, warrick, 1-17, 1-21, 12-17

eat - eat, bit, fame, gloria, raunds, watching, 'd, precious, baxter, crowd, matt
good - good, it, but, they, way, them, him, make, though, a, that
slowly - slowly, thinks, anonymity, authorial, shepheardes, callers, rise, determines, inherent, statements, fossil
100 - 100, =, 1.50, -1.5, 120.41, 1234.56, 16-12, 183.19, 19, 876.2e+20, 1943., 1e10

Window = 5

Boy - boy, hang, wood, daughter, robin, sought, hood, bonny, fore, vow, dreary
Sunday - sunday, morning, june, revelation, during, trinity, pentecost, romans, 1-21, warrick, journey
eat - eat, watching, did, definitely, reminded, florida, couple, seemed, bit, sometimes, talked
good - good, but, it, they, them, that, way, so, there, a, ,
slowly - slowly, statements, soon, heading, 1915, truce, aspect, interact, one, bulb, selling
100 - 100, =, -1.5, 1234.56, 16-12, 19, 876.2e+20, 1943., 1e10, const, dk/ds, gpg

Window = 10

boy - boy, hang, daughter, wood, sought, hood, bonny, robin, tenderly, kissed, fore
sunday - sunday, morning, june, romans, during, pentecost, trinity, 1-21, 22-27, warrick, revelation
eat - eat, often, much, did, watching, then, but, bit, getting, so, one
good – good, it, but, them, they, that, very, there, way, ‘, ’ , a
slowly - slowly, prevent, bulb, hoping, naked, happens, dominated, heading, thinks, ?, statements
100 – 100, =, centimetre, centimetres, ares, decametre, decametres, metres, hectare, hectometre, -1.5

1.2.4 SVD_changed

1.2.4.3 Verb context

SVD 200:
boy - boy, hang, hood, bonny, fore, tenderly, kissed, robin, daughter, ower, sought
sunday - sunday, morning, trinity, 1-21, warrick, 1-17, 12-17, 22-27, 26-27, 4b-15, isaiah
eat - eat, florida, did, watching, seemed, curries, would, !, den, then, fullness
good - good, it, but, they, that, them, a, , , to, way, and

slowly - slowly, 1915, truce, thinks, surely, dominated, one, katherine, soon, dizziness, statements
100 – 100, =, decimetres, ares, decametre, decametres, hectare, hectometre, metre, metres, sq.

SVD 50:
boy - boy, green, daughter, hang, sought, flood, dreamt, dreary, vow, drown, fore
sunday - sunday, morning, journey, june, during, revelation, his, visits, took, roy, 11th
eat - eat, definitely, reminded, did, couple, seemed, watching, game, behind, kind, then
good - good, but, it, them, way, might, they, around, so, very, just
slowly - slowly, dominated, prevent, wait, situation, soon, majority, however, takes, happen, ?
100 - 100, =, measurements, metric, centimetres, -1.5, 1234.56, 16-12, 19, 876.2e+20, 1943., 1e10

SVD 10:

boy - boy, ordered, rare, salt, fairtrade, obvious, media, targeting, meeting, sold, green
sunday - sunday, 17th, saturday, produced, speaker, rice, british, designer, particularly, keynes, county
eat - eat, voted, anyway, reminded, gear, thank, emnm, new-media, try, prayer, uncluttered
good - good, see, better, getting, very, rule, might, there, then, my, music
slowly - slowly, idea, speak, recruit, independence, horribly, attempt, months, above, –, denied
100 – 100, mm, =, ni-affinity, ni-sepharose, slurry, 0.137.725, tep, cm, 9.34, pcb

1.2.4.2: Separate Windows

SVD 200:

boy - boy, bonny, kissed, tenderly, hang, ower, wood, grandson, ane, robin, sings
sunday - sunday, morning, trinity, pentecost, 26-27, 22-27, ezekiel, 4b-15, isaiah, warrick, 1-21

eat - eat, but, saying, got, did, again, great, really, watching, my, just
good - good, it, but, a, again, said, got, they, that, saw, true
slowly – slowly, piloting, pressure, encouraging, enhances, surrounding, juan, primarily, controversies, romn, psychotherapeutic
100 – 100, =, centimetres, centimetre, millimetres, decimetres, metric, decimeter, measurements, sq., decametre

SVD 50:

boy – boy, bonny, wood, tenderly, kissed, corn, hang, robin, green, sought, dream
sunday – sunday, morning, 11th, mark, trinity, revelation, 26-34, readings, 1-13, 14-17, 1st
eat – eat, ride, just, watching, discovered, moment, saying, did, winds, 're, 'll
good – good, very, but, way, one, it, them, again, then, yet, great
slowly – slowly, properly, xbox, pressure, colour, nightmare, wire, agonisingly, impossible, receiver, stuck
100 – 100, =, 1000, centimetres, metre, centimetre, measurements, sq., millimetres, decimetres, 000

SVD 10:

boy – boy, sought, seat, mere, painted, bags, blackwell, singing, treat, birds, reviews
sunday – sunday, morning, bassett, bertie, embossed, pc133, amjed0403, 1st, 04/07/06, 05:42, mrs
eat – eat, ride, begin, happen, hello, her, bob, stop, mother, saying, shake
good – good, he, him, father, god, getting, right, though, them, feet, might
slowly - slowly, barnabas, button, affect, prius, appellant, constitutional, faulty, deferred, prospective, predicted

100 – 100, mm, reverence, threatenings, lira, controllable, 130x150, 130x130x50, coriander, gl-100, projector

1.2.4.1: Normal

SVD 200:

boy – boy, wood, tenderly, kissed, bonny, hang, ower, grandson, dear, sought, sings

sunday – sunday, morning, pentecost, readings, trinity, warrick, 4b-15, 26-27, 22-27, isaiah, 1-21
eat – eat, !, saying, did, again, thing, but, fun, got, think, just
good – good, it, but, again, !, them, idea, 'll, they, a, got
slowly – slowly, enough, take, broken, not, it, so, people, however, one, try
100 – 100, =, centimetre, centimetres, millimetres, decimetres, metric, measurements, decimeter, decametre, metre

SVD 50:

boy – boy, wood, bonny, tenderly, corn, kissed, green, hang, ower, sought, painted
sunday – sunday, morning, 11th, trinity, 1st, readings, pentecost, 26-34, june, ezekiel, 14-17
eat – eat, !, 'll, just, try, pull, saying, again, moment, 're, lot
good – good, very, but, way, them, it, getting, really, idea, because, even
slowly – slowly, stuck, atmosphere, probably, dilemma, unfortunately, means, takes, fabulous, simply, point
100 – 100, =, 1000, measurements, centimetres, centimetre, metric, millimetres, decimetres, sq., metre

SVD 10:

boy – boy, birds, gases, studios, bags, forced, ending, miles, tall, glorious, fold
sunday – sunday, 1912, fieldwork, 1st, autumn, morning, 11:12:52, eugenic, 18:55:24, manville, saturday
eat – eat, saying, me, 've, wanted, worry, i, 're, sit, 'll, think
good – good, him, though, my, very, but, went, he, leonardo, fear, right
slowly – slowly, individually, difficulty, criticism, realize, freed, retreat, meant, dress, by-pass, asks
100 – 100, =, scaffolding, 85, minnesota, reservations, 1682, holyhead, measurements, 1000, p(x

2. POS

Bin 0: 41799

Bin 1: 1(egdon)

BIN2: 2(puzzlements, diffused)

BIN3: 1(78)

BIN4: 2(blocked, progresive)

BIN5: 2(preventing, giotto)

BIN6: 1(uncompetitive)

BIN7: 17 (monopoly, mainland, wither, lessing, unmasked, e3.7m, chinaman, criticism., people[1, invoice, hunton, induce, it, 6, initializer, newhaven, birkhead, dodge)

BIN8: 3(removal, wallhacks, prevails)

BIN9: 6(norm, automating, astonishing, france, bysmorfullum, rearin)

BIN10: 1(delighted)

BIN11: 26(george.overton@uhl-tr.nhs.uk, affords, ploughshares, seize, yeah, row, misao, plateoffood, ability, subunits, bodies, 5mm, healthcare, affect, messrs, burk, greenock, lingaii, imperialism, loader., knows, p25, walt, 7:7, 8, need, relaxation)

BIN12: 2(gms, nation-wide)

BIN13: 991

BIN14: 6(ireland.com, 8rq, chop, iris, machinist, one-)

BIN15: 4(al-malik, inter-faith, thevisorshop.com, ten-year)

BIN16: 32(tilting, tff, uncovered, suu, leftmost, fetus, ratcliffe, apt, satellites, racer, ctbt, motivations, 785, 780, similar, borneo, anti-islamic, winger, part, doctors, carnage, supposes, mahal, namesake, optometrists, returns, cover-up, cobbly, three, waterboys, 0.00, hypogranular)

BIN17: 120(mcnamara, tundra, troubles, rover, customer-contractor, 199, november, exchequer, olly, downloadable, iain, scullers, 668‑685, polices, http://www.chm.davidson.edu/chemistryapplets/calorimetry, griffiths, special, isotropic,

kumasi, forgot, kevin, firewall, goodhart, ramble, biometric, imaginable, bulk-buy, decent, alloys, valuer, draughtsmanship, upto, babymatic, towns, 3, 808, clearance, practiced, tithe, charisma, hurry, bromeliad, 2, 118, belgian, whistleblowers, networked, newsuk, corris, post-experience, hargreaves, tortoise, maturing, dispensation, 15/12/99, oranges, unspectacular, icy, weird, thermostatic, om, tem, socindex, spate, open, ellesmere, te, 15, 152, 000, up, 1880's, licenses, bad-credit, whitewashed, dodgson, chrysocome, extensible, wheat, 15:45:20, luhon, spotted, terrain, exegetical, lubricate, sofer, 13, corner, climbing, 2b, intra-abdominal, replaced, levye, hilcrest, /etc/default/rccustom, attested, school-based, donat, 5s, council`s, contributed, sumiko, md-, swiveling, c-sap, class-based, jumpcross, lender, einsatzgruppe, aspirin, www.designersblock.org.uk, joint, originate, www.customer-projects.co.uk/downloads.htm, sift, euclid, popular-but, 135, 000, scaremongers, lizzie, cuore, predicated, ahistorical, bar.)

BIN18: 107(genesis, betacart/lms, bagging, atttention, tightrope, ex-students, tica, maori, challenger, mammal(animal):-, insulate, sweltering, implementations, '05, tells, assigned, consisted, 48x, karaoke, geoff, 4aw, turin, 1926, yippee, duct, promise, 255, articulated, newcastles, braga, ants, y'day, 4x100, hall, rudimentary, martinus, howarth, processing, lakes, pickett, foxes, resize, attributed, mohau, core, tez, design, 3.6, ferguson, painted, worked, headsets, transgressed, city-districts, http://www.prospects.ac.uk,drink,sibling,galatasaray,kerr,rankin,name(ns,qualms,people-watching,raiii,therapy,macos,/without,newsreader,199.67.205.136,mpc,believe,289.099.8744,18:27:11,recording,4050,negative,melksham,ceche,predominantly,delays,97.6fm,forbade,feel,feedback,gerrards,ate,remedial,remainder,consisting,athtar,angular,ccw,strauss,importance,trouble-shooter,blackmail,powerboating,retailers,populism,7.4,2450,temptation,panton,spoons,inseparable,superstar,1.a)

BIN19: 1(tracon)

Observations(1&2)

1) Length of the window captures the context where the word is used. So more the length of the window more is the knowledge about of the context but when the window size is increased a lot then there will be more terms added to cmatrix and only those words with all the terms are marked closer. Only those words which occur in the same

context with all those words in it are marked similar. So we get more similar terms restricted to that context only.

2) When we use separate left, right windows then we mark the context much clearly so the words which are very similar are those with same syntactic representation. But when the word order is not important and words can be used in different orders then this would fail to mark similar words in same context. But with the order importance it marks the similarity.

3) Only when verbs are considered in context, marking similarity between words is more accurate because only certain words are used with certain verbs and other words like adj, adv mark the word similarity less compared verbs because only certain words occur specific to verbs.

4) SVD when dimensionality is reduce

5) Once we placed those words into their respective bins via our program, we found no big correlation between the words in the bins.

6) Bin 0 contained the most amount of words by a big margin i.e. 41799 words.

7) Bin 17 contained some interesting words, but we did not find POS categorization of the same.

8) Also, most of the bins did not contain more than 10 words. We have bins which contain only 1 word each, etc. which show that the data plays a big role in deciding how our bins are formed. In this case, it was not very helpful.