Topic Clustering

- Enron dataset (Lay, Skilling, Forney, Delainey)

- Training:Test in 9:1

Fech Scen Khoo
Universität des Saarlandes

25 November 2020

Models considered for the test documents

PVTM model: paragraph vector + Gaussian mixture model

Ref.: Lenz and Winker, "Measuring the diffusion of innovations with paragraph vector topic models", PLOS ONE (2020)

▷ We use the metric: BIC values (the lower the better)

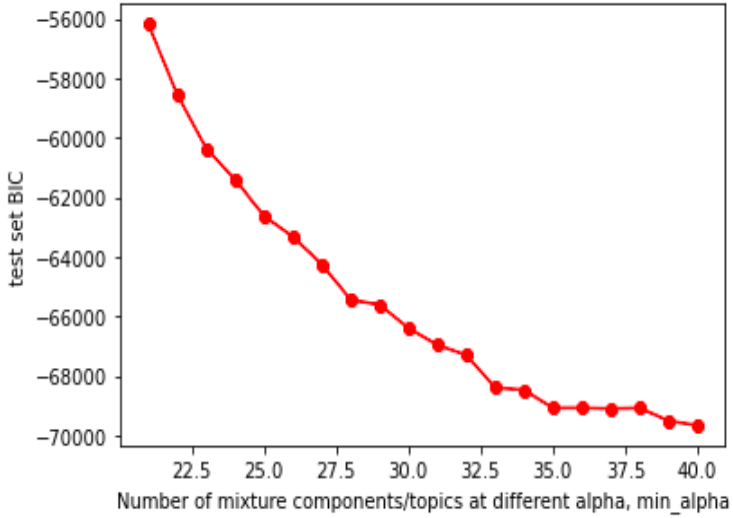
LDA model

Ref.: David M. Blei, Andrew Y. Ng and Michael I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research 3 (2003) 993-1022

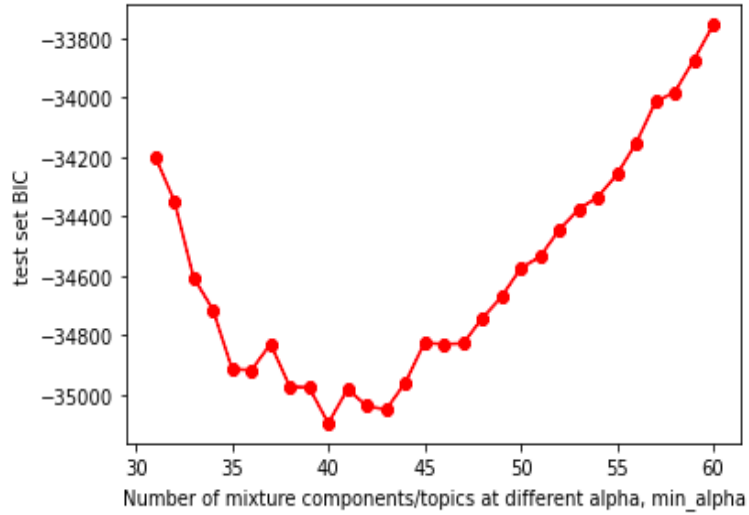
▷ We use the metric: Coherence (the higher the better)

Remark: Used PVTM in training and inference

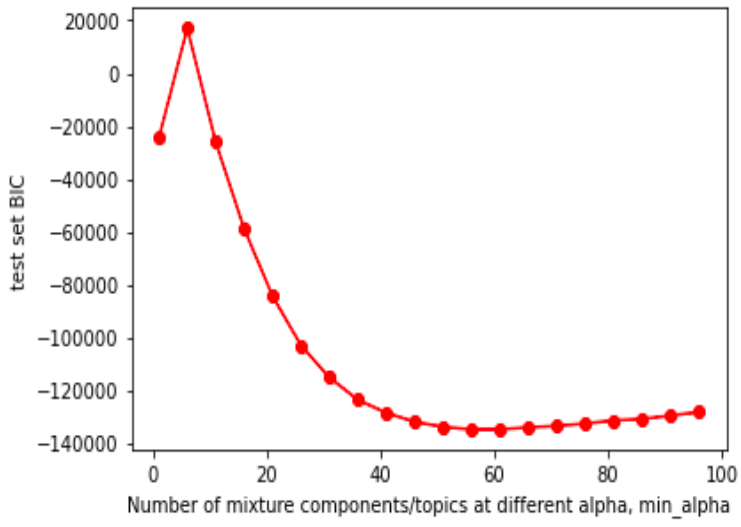
PVTM: BIC of 4 groupings of test documents



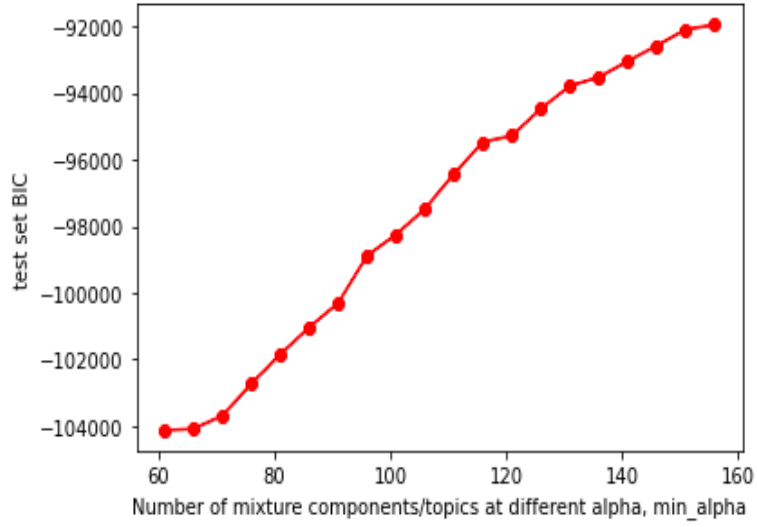
(1,1)



(1,2)

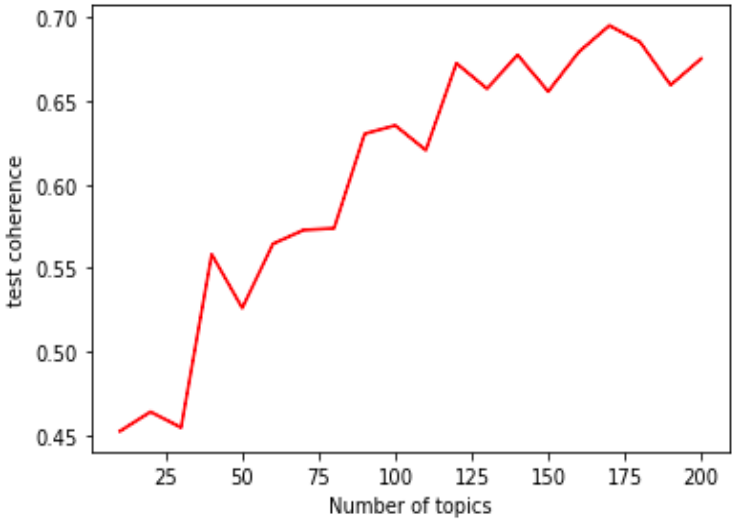


(2,1)

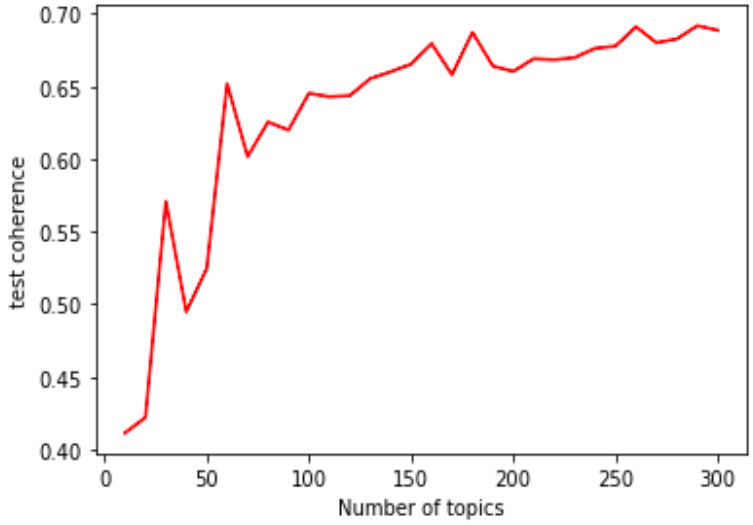


(2,2)

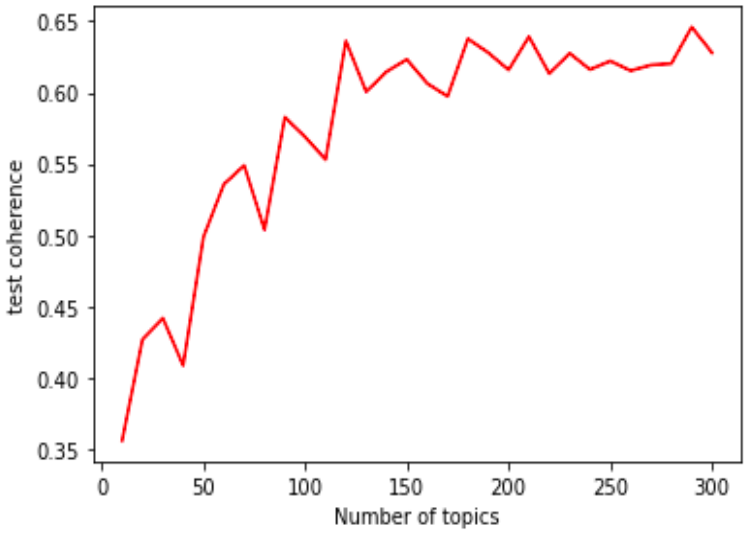
LDA: Coherence of the same 4 groupings of test documents



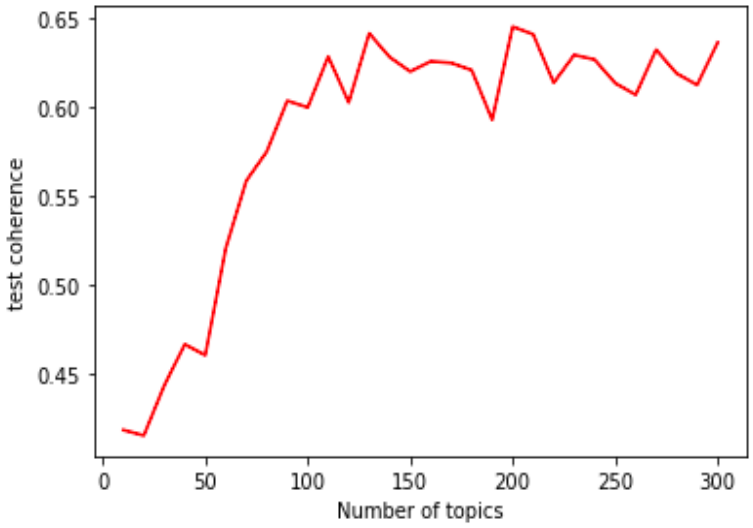
(a,a)



(a,b)



(b,a)



(b,b)

PVTM: Topics of plot (1,1)

139 docs [trained model n160], BIC ≈ -68828, $n = 36$

Topic A: credit, billion, stock, investor, paper, enron, commercial, bank, line, financial, company, fall, share, partnership, begin

Topic B: incredible, reunion, everyone, future, forget, big, consider, tent, crowd, connect, pick, stay, year, far, ago

Topic C: asset, equity, percent, york, dollar, shareholder, decision, limit, ljm, partner, approval, security, news, lose, quarter

Topic D: width, spin, td, tr, table, class, br, border, cellspacing, cellpadding, normaltext, top, tag, information, right

LDA: Topics of plot (a,a)

139 docs [trained model n160],
coherence ≈ 0.6725, $num\_topics = 120$

Topic A: enron, company, billion, investor, credit, financial, know, stock, partnership, week, business, line, paper, commercial, share

Topic B: years, beta, homecoming, reunion, pick, forget, ago, class, everyone, seem, brother, caldwell, guy, crowd, big

Topic C: meeting, daily, new, edition, conference, york, chief, president, wto, newspaper, assembly, world, economic, issue, publish

Topic D: harassment, training, session, date, workplace, enron, learn, avoidance, violate, values, hotel, understanding, contribute, respectful, tree

PVTM: Topics of plot (2,1)

248 docs [trained model n260], BIC ≈ -134886, $n = 60$

Topic A: tr, td, width, spin, table, class, border, cellpadding, cellspacing, normaltext, br, **gif**, member, information, top

Topic B: residential, llc, jimmie, show, loss, retail, hydrocarbon, gray, perform, propose, direction, high, entry, create

Topic C: story, oct, caracas, petroleumworld, nov, oil, venezuela, opec, previous, president, full, war, production, cut, ohep

Topic D: market, lehman, buy, company, expect, brother, old, inc, perform, trade, offer, estimate, business, strong, percentage

Topic E: natural, gas, business, news, power, electric, com, utility, co, trade, customer, new, chief, president, focus

Glossary: llc: limited liability company; opec: org. petroleum exporting countries; ohep: office of home energy programs

LDA: Topics of plot (b,a)

248 docs [trained model n260],
coherence ≈ 0.6361, $num\_topics = 120$

Topic A: tr, td, width, span, table, br, class, normaltext, cellspacing, border, cellpadding, li, wpo, tag, align

Topic B: company, market, trading, new, panel, brother, believe, beta, bandwidth, lehman, jones, information, buy, dow, rate

Topic C: energy, power, gas, customer, newpower, new, electric, plan, consumer, business, utility, electricity, texas, state, natural

Topic C': energy, power, new, consumer, newpower, customer, plan, gas, percent, texas, business, txu, price, please, deal

Topic E: minister, prime, sponsor, event, vajpayee, commercial, reception, washington, member, government, dignitary, indian, commerce, congress, company

Topic F: worth, enron, executive, editor, profile, magazine, select, interview, quality, high, economy, list, manager, top, company

Topic G: deepwater, production, devil, jv, tower, assets, structure, lease, capacity, project, financial, note, sell, spv, rfp

Topic H: photographer, wedding, might, try, linda, florist, basic, love, help, machine, know, number, decorate, church, phenomenal

Topic I: petroleumworld, broadband, us, story, oil, enron, opec, venezuela, caracas, regard, oct, nov, interest, telion, previous

Topic J: walk, jdrf, join, sign, entertainment, donation, sneaker, us, walker, collect, great, lunch, shirt, sale, fun

Topic K: alumnus, association, october, ag, homecoming, register, visit, mizzou, festivity, mu, notification, furnish, tent, directory, muaa

Glossary:

jv: joint venture?
spv: special purpose vehicle
rfp: request for proposal?
jdrf: org. diabetes research
mizzou: Uni. of Missouri

PVTM: Topics of plot (1,2)

75 docs [trained model n310], BIC ≈ -35056, $n = 40$

Topic A: billion, credit, line, commercial, paper, investor, stock, bank, low, bond, financial, enron, company, tap, cent

Topic B: story, oct, petroleumworld, caracas, nov, oil, venezuela, opec, previous, president, full, elio, law, war, ohep

Topic C: expect, market, brother, lehman, point, percentage, buy, perform, estimate, inc, old, strong, street, eps, rate

Topic D: panel, jones, minute, dow, panelist, bandwidth, allocate, representative, follow, remark, session, energy, speaker, newswires, associate

Topic E: alliance, datum, sound, utility, online, energy, industry, form, puget, service, marketplace, base, team, product, internet

Glossary: eps: earnings per share; puget sound energy company

LDA: Topics of plot (a,b)

75 docs [trained model n310],
coherence ≈ 0.6553, $num\_topics = 130$

Topic A: story, oct, petroleumworld, caracas, oil, nov, venezuela, opec, previous, president, us, elio, com, ohep, law

Topic B: panel, bandwidth, trading, jones, dow, panelist, company, energy, ceo, remark, allocate, representative, follow, minutes, keynote

Topic B': panel, ebusiness, vp, conference, industry, member, event, draw, meeting, culture, well, move, anxious, steve, dr

Topic D: core, business, company, non, enron, value, know, assets, make, decision, maintain, million, believe, future, focus

Topic D': core, company, business, enron, non, sell, plan, time, future, trading, valuable, know, message, decision, value

Topic F: market, demand, estimate, lehman, believe, new, strong, expect, buy, eps, brother, year, security, information, cable

Topic F': market, company, believe, new, brother, lehman, expect, buy, strong, power, rate, information, trading, old, year

Topic H: investor, financial, company, enron, week, stock, billion, paper, cash, line, share, debt, share, partnership, commercial

Topic I: trading, alliance, energy, online, full, power, product, fuel, story, data, utility, click, new, week, tradespark

Topic I': energy, power, cell, click, full, tradespark, data, alliance, new, product, online, story, trading, fuel, puget

PVTM: Topics of plot (2,2)

200 docs [trained model n460], BIC ≈ -104898, $n = 61$

Topic A: estimate, buy, expect, strong, market, point, percentage, perform, eps, rate, usd, reduce, underperform, outperform, continue

Topic B: billion, enron, credit, stock, financial, commercial, paper, investor, bank, line, trade, corp, chief, company, fall

Topic C: gas, power, natural, customer, energy, electric, co, houston, plan, electricity, bill, texas, consumer, state, newpower

Topic D: brother, lehman, expect, estimate, inc, point, market, strong, percentage, director, buy, underperform, reduce, ercot, perform

Glossary: ercot: electric reliability council of texas

LDA: Topics of plot (b,b)

200 docs [trained model n460],
coherence ≈ 0.6417, $num\_topics = 130$

Topic A: market, expect, brother, lehman, believe, new, estimate, eps, year, strong, recent, business, company, buy, share

Topic B: enron, company, energy, power, vacation, price, gas, natural, co, take, billion, utility, state, year, time

Topic B': energy, power, new, consumer, customer, gas, newpower, ampo, plan, city, texas, news, electric, co, utility

Topic D: beta, years, homecoming, yahoo, wall, reunion, member, everyone, group, seem, caldwell, incredible, future, ago, tent

Topic D': years, beta, homecoming, class, brother, guy, reunion, ago, forget, pick, left, tent, future, incredible, see

Topic F: defense, civil, shelter, american, vanguard, one, even, well, america, world, people, protect, attack, state, terrorist

Topic G: skybar, mixer, parking, two, us, club, bring, hope, entire, information, please, kroger, montrose, valet, outdoor