CMSC 409 Project 4 - Aaron Kim, Aplomb Huang

1. Feature Vector:

[autonom, sedan, travel, type, road, speed, up, mile, per, hour, futur, machin, learn, rai, kurzweil, predict, year, awai, singular, self-improv, artifici, super-intellig, far, exce, human, intellig, get, car, kilomet, second, newli, remodel, home, rent, bedroom, bath, live, room, larg, eat, kitchen, full, size, util, test, achiev, rang, around, charg, musk, wai, escap, obsolesc, have, sort, merger, biolog, entir, interior, freshli, paint, go, percent, befor, lap, possibl, lead, necessarili, sentienc, iot, devic, part, larger, concept, autom, includ, light, heat, air, condit, media, secur, long, term, benefit, energi, save, automat, ensur, electron, turn, applianc, well, system, tenant, respons, electr, water, ga, pet, negoti, base, anim, four, row, hous, come, washer, dryer, finish, basement, three, park, space, back, approv, owner, limit, public, us, driven, combin, two, paradigm, everyth, know, realiti, sens, knowledg, experi, via, five, on, suit, king, bed, veri, nice, queen, two-, four-door, design, built, normal, chassi, shorter, roof, club, avail, high-level, u , model, complet, over, drive, accident-fre, updat, renov, new, floor, john, mccarthi, inventor, program, languag, lisp, coin, “artifici, intelligence”, deal, number, situat, averag, gallon, pound, gener, involv, self-awar, comput, engag, common-sens, reason, attain, multipl, domain, feel, express, understand, emot, singl, famili, conveni, locat, near, major, rout, author, book, ag, spiritu, describ, spread, throughout, cosmo, cute, classi, open, area, great, decor, furnish, adorn, fashion, beauti, thing, went, round, minut, recent, work, fundament, techniqu, deep, lai, groundwork, through, increas, world, them two, bathroom, townhous, central, trash, sewag, vehicl, altern, fuel, such, ethanol, flexible-fuel, natur, gain, popular, countri, invent, earli, histori, began, becom, commerci, hollywood, movi, scienc, fiction, novel, depict, ai, human-lik, robot, take, current, evolut, technolog, isn’t, scari, quit, smart, art, peopl, behavior, creat, function, within, build, individual’, reflect, person, social, class, socioeconom, statu, uniqu, each, highli, skill, order, environ, safe, adher, code, regul, ada, requir, exampl, calcul, basic, recogn, text, method, optim, charact, recognit, longer, now, taken, grant, inher, small, desk, chair, upholst, chest, drawer, western, call, master, contain, closet, commonli, cloth, item, mirror, definit, sourc, feet, apart, elimin, shadow, face, modern, lot, varieti, help, motorist, find, unoccupi, guidanc, inform, retriev, improv, internet, action, initi, human-centr, mediat, role, primari, provid, framework, product, reproduct, passeng, pollut, contributor, produc, signific, amount, nitrogen, oxid, carbon, monoxid, self-driv, subject, controversi, tend, lowest, risk, casualti, realiz, seamless, integr, variou, manufactur, equip, identif, process, commun, actuat, network, capabl, trend, explos, growth, connect, control, internet]

Total of 387 Tokens

Contin.

2.1 Since there were 46 sentences (so 46 rows) and 387 feature vectors (so 387 columns), the Term Document matrix was too big, therefore can be viewed in a separate .txt file. The rows again correlate to which sentence. For the sanctity of spacing and viewing pleasure the columns despite not being labeled, are correlated to the index of an array called tokenList which contains all the list of the features vectors or what we called tokens. If you so wish to desire which token the sentence belongs, a simple get method from arrayList can be done ascertain the label. The TDM is a nice way to physically preemptively view any possible correlation.

2.2 A. Tokenize sentences - This is probably the first and most important step. This is ensure that the sentences (in this case) or documents have things in common. If sentences are not tokenized, there will no proper comparisons amongst the documents. There will be no information loss, only potential gain. A potential loss would certain key sentences that may be very nuanced in the topic or very active quotations.

B. Remove punctuation and special characters - This is used to help get better matching amongst the documents so that a character like a period will not get in the way of a match. Also to prevent false similarities or connections with documents, sort of an overinflation in a way. Information can be loss if the special characters pertain to the special topic of the document or if a certain array of characters means something special to the user.

C. Remove numbers - This is ensure similarities are not over inflated. Just because documents have similar numbers, does not mean that are related. For example, one could be advertisement and the other could be a scientific research paper. Information could be loss in a way that certain numbers like scientific constants are important and very relevant, but would not be used in the comparisons.

D. Convert upper-case to lower-case - This is to ensure better quality of comparison. So that certain key phrases are compared accurately, and would be judged as a match, instead of a simple capitalized letter preventing the match. Information could be loss as in proper nouns could lose their value and would not be weighted as highly.

E. Remove stop words. - Removing stop words which are very common words can prevent inflation in the comparisons, so that that documents would be more related to each other contextually. Information loss would most likely occur if there are certain phrases that involve stop words.

F. Perform stemming. - This is ensure quality of the comparisons between the sentences. Things like proper grammar like ‘ed’ or ‘ing’ would not prohibit matches between the stems of the words, and therefore is useful to use. Information would be loss if certain phrases are more useful and relevant to the topics of the document like ionized vs ‘ioni’. There more of a chance to get more matches due to stemming as well which depending on the quality of everything else can be a good or bad thing.

G. Combine stemmed words. - This is hard to implement, but very useful when done well. It helps with more relevant contextual matching and would solve the priors (F’s) problem of possible inflation of matches. Information is loss as combining can be more specific and narrow in its search, so a information that could be useful would be loss if the search was not more broadened.

3. The results maybe could’ve been better if it were a FCAN instead of a WTA. FCAN can be more optimal, but requires more effort to implement. FCAN would most likely obtain cleaner and more accurate results.

Result: cluster 0: 01, 01, 11, 21, 31, 41, 51, 61, 71, 81, 91, 101, 111, 121, 131, 141, 151, 161, 171, 181, 191, 201, 211, 221, 231, 241, 251, 261, 271, 281, 291, 301, 311, 321, 331, 341, 351, 361, 371, 381, 391, 401, 411, 421, 431, 441

BUILD SUCCESSFUL (total time: 0 seconds)