Latent Dirichlet Allocation vs. Non-Negative Matrix Factorization

Spenser Johnson
Mentor: Koyuki Nakamori

Introduction

We're back again with 29,752 Jeopardy! questions and answers. Based on the text, we will compare two Unsupervised Learning topic generation models: Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). These methods will cluster our questions into ten different topics. Then, we will use these generated features to create a Supervised regression model to predict if a question belongs to a certain topic.

This particular data set was found on Kaggle. During the cleaning phase we excluded questions from categories that appeared fewer than 100 times.

Latent Dirichlet Allocation Topics

- Topic 0: Miscellaneous
  [('derived', 48.765983969795826), ('late', 43.748010347653953), ('using', 40.690556329113754), ('40', 39.615694081092002), ('gas', 36.839162477405011), ('wild', 36.264318116004894), ('jersey', 36.127885122826811), ('free', 34.78341105323252), ('divided', 34.620428959619915), ('metal', 34.413687974552097)]
- Topic 1: Miscellaneous - Compare use of Dutch to NMF Topic 4
  [('spoken', 56.795920474411581), ('dutch', 51.802380804478403), ('gulf', 46.475726256528212), ('street', 43.400453628951595), ('formed', 42.117864013156257), ('dog', 41.864878320078965), ('ship', 40.005207903012135), ('mark', 39.674446158105553), ('coast', 39.247481130956594), ('moon', 37.459316776122407)]
- Topic 2: Miscellaneous - Politics, but contains Jefferson and Japanese
  [('takes', 46.621455684590238), ('member', 42.628964466279356), ('cities', 41.780903544413142), ('foot', 40.936432620922091), ('elected', 38.982990689831851), ('2009', 37.631029019798056), ('jefferson', 36.76631356468463), ('painting', 36.760347706155599), ('sir', 36.350601145399764), ('japanese', 36.146023581777271)]
- Topic 3: Some Geographical content
  [('peak', 48.559084540664394), ('types', 46.312764360718958), ('contains', 44.353922374784759), ('india', 42.929462853437272), ('food', 42.219749306896468), ('port', 40.467064719771969), ('plant', 40.301640716953791), ('100', 39.164764590731693), ('right', 38.744907850167422), ('statue', 38.695403929691125)]

Latent Dirichlet Allocation Topics

- Topic 4: Misc. - Contains mother and female, but also mount and Roosevelt
  [('order', 47.311121113684983), ('mother', 45.850914849711437), ('female', 45.302898206839174), ('mount', 45.07834621097367), ('according', 39.999835580281321), ('didn', 39.583126201865092), ('30', 39.054883978040202), ('roosevelt', 37.786463179664807), ('invented', 36.264621860925459), ('square', 35.281783080646356)]
- Topic 5: Misc. - Maybe Shakespeare related
  [('given', 41.850127907654006), ('johnson', 40.835023066260682), ('birds', 39.997145816500421), ('sound', 39.181800464096902), ('modern', 39.024928475943298), ('game', 38.67906032462755), ('emperor', 37.961389776342649), ('devoted', 36.855464649127022), ('peter', 36.672334153400236), ('shakespeare', 36.048488684912392)]
- Topic 6: Misc.
  [('africa', 50.510445737254514), ('army', 42.797661708261963), ('pacific', 41.506019653675288), ('animals', 38.845621467526364), ('festival', 38.541738189133383), ('actress', 37.479899640257443), ('military', 36.397916696641893), ('thisthe', 34.988320825776078), ('prime', 34.582963775400593), ('jesus', 34.374317749168156)]

Latent Dirichlet Allocation Topics

- Topic 7: Misc.
  [('variety', 45.891068673689446), ('ocean', 44.998901483030707), ('fish', 41.943576472540435), ('golden', 41.318255372733788), ('books', 40.991259533320381), ('nickname', 40.464242046303717), ('grand', 40.449631025906811), ('run', 39.115324379830554), ('stone', 37.712768740231077), ('dr', 36.668702839843199)]
- Topic 8: Some Historical content
  [('history', 45.73401142555128), ('building', 45.329823302807455), ('original', 43.547367629856275), ('spain', 39.039559499308531), ('tell', 37.74265600702121), ('continent', 37.23095857382355), ('hero', 36.713073392543279), ('brown', 36.48949667906988), ('power', 35.12822281869137), ('plays', 34.066469523868633)]
- Topic 9: Misc.
  [('mean', 50.501714096980777), ('phrase', 46.441025305169681), ('royal', 44.557530391633875), ('horse', 43.764494881627861), ('northern', 40.722835649887891), ('mountains', 40.230672899423062), ('florida', 38.286697417132011), ('dance', 37.577283214575729), ('tree', 36.972814669331768), ('adams', 35.65452393161037)]

Non-Negative Matrix Factorization Topics

- Topic 0: Horse Racing
  [('horse', 3.2482111937393352), ('race', 0.22403469654060248), ('racing', 0.16814883533728442), ('nile', 0.14291646486177598), ('crown', 0.13662653261471722), ('winged', 0.11750627784939911), ('triple', 0.11704734444098196), ('drawn', 0.088758738396920972), ('lee', 0.086461189776939396), ('wild', 0.084097009739033779)]
- Topic 1: Languages
  [('spoken', 3.3403620019198379), ('languages', 0.92757637191391762), ('widely', 0.29467359115899316), ('dialect', 0.10853632261427593), ('united', 0.097903432557003608), ('morning', 0.080150318934918002), ('continentafrica', 0.064494877377648155), ('guinea', 0.049588259753491751), ('portuguese', 0.043083190231663453), ('continent', 0.040499163236133687)]
- Topic 2: Mountains
  [('peak', 2.582106579368824), ('mount', 2.2593353947847419), ('mountains', 0.98001854128311061), ('range', 0.67812349189283871), ('foot', 0.53927526981259366), ('everest', 0.28831995094196572), ('colorado', 0.26061232876312351), ('border', 0.17959748699966641), ('14', 0.16164233690501589), ('mt', 0.13763971264958125)]
- Topic 3: Misc. - Parts of speech and Baseball
  [('mean', 3.4026288681259733), ('adjective', 0.23201931575208518), ('verb', 0.21335282503047318), ('face', 0.13620562271552764), ('fifth', 0.093773205124551645), ('worn', 0.072255741037261748), ('card', 0.060347324393678167), ('baseball', 0.059549811106706749), ('tissue', 0.054697533963548733), ('doesn', 0.052092832422346656)]

Non-Negative Matrix Factorization Topics

- Topic 4: Dutch Culture

  [('dutch', 3.3419556739962748), ('orange', 0.20556091930541984), ('cake', 0.19492057788780637), ('cape', 0.10496924638570156), ('royal', 0.079050704791982063), ('colony', 0.072719713649597298), ('gogh', 0.071580083856695662), ('netherlands', 0.070024872492022341), ('treaty', 0.068448287090149371), ('kind', 0.063675483602507299)]
- Topic 5: Oceanography

  [('ocean', 2.6404071793767012), ('pacific', 2.2271103657457862), ('atlantic', 0.48772279702842214), ('port', 0.40483693787367603), ('half', 0.17863933549873839), ('arm', 0.12842660703896602), ('territory', 0.096488329476895257), ('coastline', 0.089793737654656022), ('size', 0.073758348517651826), ('arctic', 0.063880267285881837)]
- Topic 6: Misc. - Some Science content

  [('types', 3.3109535103060859), ('basic', 0.16194543738032111), ('different', 0.094418982739042909), ('sweet', 0.073187714705687218), ('cells', 0.063097417795312735), ('organisms', 0.060092996358508559), ('particle', 0.056248491083389363), ('spot', 0.053538107186909406), ('liquor', 0.045780944154156809), ('matter', 0.042648900861595569)]

Non-Negative Matrix Factorization Topics

- Topic 7: African Culture / Geography
  [('africa', 3.3286000867574002), ('cape', 0.29318309857746699), ('populous', 0.19463343438108527), ('countries', 0.10809239911951668), ('desert', 0.10135933614312169), ('mandela', 0.089285546743635924), ('horn', 0.052769080364747652), ('kenya', 0.051655864405832953), ('asia', 0.050826847105200378), ('continent', 0.048736866091171595)]
- Topic 8: Misc.
  [('derived', 3.3085688897332308), ('colorwhite', 0.074926536242787167), ('probably', 0.073958583679734721), ('andes', 0.066608103097461277), ('worn', 0.0383636813712721), ('metal', 0.037474926377671414), ('bones', 0.037435791118517352), ('synonym', 0.037398661652159208), ('daniel', 0.037382680745170989), ('partly', 0.036644865722071387)]
- Topic 9: Geography
  [('gulf', 3.217966285331209), ('persian', 0.52547498266552506), ('arabian', 0.23590134326190229), ('suez', 0.15043186618482507), ('finland', 0.13690049527953826), ('connects', 0.11838739539972372), ('alabama', 0.11244782219682406), ('arm', 0.095446183468845813), ('coastline', 0.093453413186103534), ('channel', 0.093335233042847326)]

Preparation for Supervised Learning

- NMF creates more meaningful and clustered topics than LDA, so we'll use that to assign the topics as unsupervised feature generation.
- The problem is that NMF is currently assigning most questions to topic 9.
- Topic 9 Questions 27370
  2 650
  5 404
  7 249
  0 235
  3 226
  4 212
  1 174
  6 145
  8 87

Logistic Regression: 91%

Adjusted R-Squared for ten topics...at fifty topics is drops to around 60%...could look into it more but the focus here is not Supervised Learning!

Further Steps and Research

- Right now word2vec is returning topics that have only years as key words, so fix that
- Example: Topic 0:
  [('1931', 0.10000103601623145), ('1994', 0.10000103407644656), ('1855', 0.10000101271067602), ('2012', 0.10000099080677116), ('30', 0.10000099014785155), ('1835', 0.10000099006047798), ('1919', 0.10000098524159437), ('1982', 0.10000098119254308), ('20s', 0.10000098006706748), ('1783', 0.10000097693297465)]
- Fix class imbalance on Topic 9 from our NMF model
- Create a Jeopardy! question generative neural network in my Thinkful specialization