Anthony Colucci – MSIA 414 Homework 2

Part 1:

I created four sets of word2vec embeddings using the python genism package (all code in HW_Script2.py). I created the embeddings each time using the Newsgroups text corpus. The first version is a baseline using all of the default values for the word2vec parameters. The second version is the same except it uses the skip-gram embedding instead of the default continuous bag of words (CBOW) embedding. The third version uses the Phrase module to create embeddings for commonly occurring bigrams, allowing for a wider array of words to be compared, using the default CBOW algorithm. The final version uses both the bigram embeddings and the skip-gram algorithm to complete the comparison of differences. I decided to compare these parameter settings by taking a sample of words ('wisconsin', 'atheism', 'college', 'pizza', 'math', 'halloween', 'water', 'sunday', 'anarchy', 'train') and comparing what words are found to be most similar for each set of embeddings (full list of similar words at the bottom of this document).

One of the initial observations from these similar words was that adding bigrams changed the ordering of similarity for unigrams. This makes sense once I recognized that the bigrams introduced provide new ways for words to connect with each other in the embeddings and therefore may make other words more relatively distant. Using water as an example, the CBOW embeddings provided more confident predictions of similar words, additionally each set of embeddings included "heat" in the 5 most-similar words, except for the skip-gram model which included "heating", which matched with the conjugations of the other similar words, indicating possibly that it identified "water" primarily as a different part of speech than the other encodings. All of the encodings were most successful (by some measure) in encoding "Sunday" as most similar to five other days of the week, but all had very different interpretations of the word "anarchy" with no similar words across any of the encodings.

Part 2:

In a general sense, word2vec and BERT are meant to perform similar tasks of converting language into a numeric representation that can be analyzed with statistical models and methods. However, word2vec was created significantly earlier than BERT, and while word2vec has seen many optimizations throughout the years that can improve both the time it takes to process large amounts of text and the accuracy and flexibility of the representations, BERT's newer approach provides more flexibility in it's ability to capture information from different levels of language data.

Considering their impact on the NLP research community, word2vec seems to be the more influential paper, given its expansion on the state of the art and approximately 10x the number of citations on Google Scholar, but BERT, published only recently, shows a good deal of impact in further pushing the state of the art in a much more crowded space for NLP research.

BERT additionally takes on more steps than word2vec, suggesting a "one-size-fits-all" style pre-training step for all representations, while then fine-tuning based on the particular task in question. Word2vec only takes on a single step for training, requiring potentially longer times training on models while tuning hyperparameters.

	WORD2VEC	BERT
CREATED BY	Google (Mikolov, Sutskever, Chen, Corrado, Dean)	Google (Devlin, Chang, Lee, Toutanova)
PUBLISHED IN	2013	2019
GOOGLE SCHOLAR CITATIONS	15356	1894

Comparison of most similar words across differently trained word2vec encodings:

```
{'wisconsin': {'baseline': [('madison', 0.8855159282684326),
   ('dame', 0.7922365069389343),
   ('notre', 0.7859771847724915),
   ('maryland', 0.7723612189292908),
   ('college', 0.7707038521766663)],
  'skipgram': [('eau', 0.8290741443634033),
   ('madison', 0.7979129552841187),
   ('claire', 0.7978003621101379),
   ('microbiology', 0.7886619567871094),
   ('marquette', 0.7702934741973877)],
  'withphrase': [('maryland', 0.8800849914550781),
   ('pennsylvania', 0.8678861260414124),
   ('electronic_engineering', 0.8497346043586731),
   ('delaware', 0.8428654670715332),
   ('organization_helsinki', 0.8387303948402405)],
  'sg_withphrase': [('chemical_engineering', 0.8272753953933716),
   ('maryland', 0.8244445323944092),
   ('american_academy', 0.8200943470001221),
   ('kentucky', 0.8170647621154785),
   ('saskatchewan', 0.8109601736068726)]},
 'atheism': {'baseline': [('autos', 0.7746827602386475),
   ('folklore', 0.7725124359130859),
   ('preciou', 0.7415283918380737),
   ('psychoactives', 0.7308033108711243),
   ('motorcycles', 0.7265267372131348)],
  'skipgram': [('alt', 0.7868323922157288),
   ('conspiracy', 0.7522424459457397),
   ('talk', 0.7471401691436768),
   ('moderated', 0.7465172410011292),
   ('pixutils', 0.7374867796897888)],
  'withphrase': [('clipper', 0.7358898520469666),
   ('morality', 0.7335233688354492),
   ('christianity', 0.7046993970870972),
   ('ignorance', 0.6955459117889404),
   ('islam', 0.6905561685562134)],
  'sg_withphrase': [('bizarre', 0.726189136505127),
   ('homosexuality', 0.7143049240112305),
   ('radical', 0.7096129655838013),
   ('mythology', 0.7095731496810913),
   ('origins', 0.70294588804245)]},
 'college': {'baseline': [('maryland', 0.7908498048782349),
   ('school', 0.7758386135101318),
   ('wisconsin', 0.7707038521766663),
   ('univ', 0.7442282438278198),
   ('florida', 0.7392972707748413)],
  'skipgram': [('osteopathic', 0.6907256841659546),
   ('saratoga', 0.6878799200057983),
   ('arkansas', 0.6579011082649231),
   ('mudd', 0.6565987467765808),
   ('scotia', 0.653168797492981)],
  'withphrase': [('computer_science', 0.8762494325637817),
   ('engineering', 0.8749523758888245),
   ('computing', 0.8579647541046143),
   ('institute', 0.8496163487434387),
   ('dept', 0.8455986976623535)],
  'sg_withphrase': [('osteopathic_medicine', 0.7349769473075867),
   ('delaware', 0.7096323370933533),
   ('maryland', 0.7073019742965698),
   ('portland', 0.70550537109375),
   ('microbiology_osu', 0.7046837210655212)]},
 'pizza': {'baseline': [('tlu', 0.7036414742469788),
   ('flaming', 0.660525918006897),
   ('gic', 0.6377788186073303),
   ('gslv', 0.5894471406936646),
   ('kirzioglu', 0.5826891660690308)],
  'skipgram': [('flaming', 0.6970493793487549),
   ('tlu', 0.6942031383514404),
   ('disposal', 0.6929289698600769),
   ('wings', 0.6547553539276123),
   ('ballyard', 0.6251239776611328)],
  'withphrase': [('elmbrook', 0.7228265404701233),
   ('provo_ut', 0.7187352180480957),
   ('prutchi', 0.7127741575241089),
   ('karlsruhe_germany', 0.7119694948196411),
   ('lincon', 0.7108801603317261)],
  'sg_withphrase': [('disposal', 0.8557425737380981),
   ('cincy', 0.8232797384262085),
   ('surfing', 0.8222557306289673),
   ('recently_upgraded', 0.8141143321990967),
   ('arkansas', 0.807761549949646)]},
 'math': {'baseline': [('dept', 0.704418420791626),
   ('nyikos', 0.6089462637901306),
   ('weedeater', 0.6068848371505737),
   ('department', 0.5940384864807129),
   ('milo', 0.5834436416625977)],
  'skipgram': [('undergrad', 0.7187871932983398),
   ('weedeater', 0.7096953392028809),
   ('harelb', 0.6795961856842041),
   ('papresco', 0.6763902306556702),
   ('rrmadiso', 0.6421548128128052)],
  'withphrase': [('dept', 0.7725290060043335),
   ('computer_science', 0.7591778635978699),
   ('tech', 0.7552310228347778),
   ('sci', 0.7466734647750854),
   ('univ', 0.7327686548233032)],
  'sg_withphrase': [('sci', 0.6712220907211304),
   ('denver_dept', 0.6269047856330872),
   ('infinity_so', 0.6237177848815918),
   ('comp', 0.621103823184967),
   ('res', 0.6184713244438171)]},
 'halloween': {'baseline': [('facto', 0.769126296043396),
   ('nxirt', 0.7562804222106934),
   ('ruiter', 0.7342511415481567),
   ('innjkv', 0.720191478729248),
   ('dkfz', 0.7184090614318848)],
  'skipgram': [('sorcerors', 0.9573749899864197),
   ('eckersly', 0.9523353576660156),
   ('_millions_', 0.9512154459953308),
   ('friggin', 0.9502686262130737),
   ('parapsychologist', 0.9494603872299194)],
  'withphrase': [('ietf', 0.714362382888794),
   ('bettern', 0.7116301655769348),
   ('oyb', 0.7094466686248779),
   ('anterior', 0.701113224029541),
   ('zce', 0.6968926191329956)],
  'sg_withphrase': [('blossom', 0.9753079414367676),
   ('sustainings', 0.9727280735969543),
   ('offi', 0.9726210832595825),
   ('linen', 0.9723285436630249),
   ('haroun', 0.9722879528999329)]},
 'water': {'baseline': [('fuel', 0.7125285863876343),
   ('pressure', 0.697189211845398),
   ('heat', 0.6966814994812012),
   ('metal', 0.6899901032447815),
   ('ground', 0.6899632215499878)],
  'skipgram': [('diverting', 0.665228009223938),
   ('plants', 0.6615215539932251),
   ('steam', 0.6606500148773193),
   ('heating', 0.6594173908233643),
   ('flowing', 0.655773401260376)],
  'withphrase': [('oil', 0.8179678916931152),
   ('ground', 0.810444712638855),
   ('heat', 0.80181884765625),
   ('pressure', 0.7972385883331299),
   ('glass', 0.7950130701065063)],
  'sg_withphrase': [('heat', 0.684272050857544),
   ('plastic', 0.6826778650283813),
   ('steam', 0.6784791350364685),
   ('wind', 0.6738631725311279),
   ('gas', 0.6711673736572266)]},
 'sunday': {'baseline': [('saturday', 0.8939638733863831),
   ('tuesday', 0.880969762802124),
   ('monday', 0.8703751564025879),
   ('thursday', 0.8636084794998169),
   ('wednesday', 0.8520613312721252)],
  'skipgram': [('saturday', 0.9074181914329529),
   ('monday', 0.838702380657196),
   ('friday', 0.8377469778060913),
   ('thursday', 0.8230875730514526),
   ('tuesday', 0.8223088383674622)],
  'withphrase': [('saturday', 0.8956931829452515),
   ('thursday', 0.8526354432106018),
   ('friday', 0.8393689393997192),
   ('vacation', 0.8055986762046814),
   ('march', 0.8033523559570312)],
  'sg_withphrase': [('saturday', 0.8826751708984375),
   ('friday', 0.8329971432685852),
   ('monday', 0.8213541507720947),
   ('on_monday', 0.8133141994476318),
   ('tuesday', 0.812187671661377)]},
 'anarchy': {'baseline': [('drifted', 0.6736441850662231),
   ('integrate', 0.6067550778388977),
   ('oxalic', 0.6031420826911926),
   ('tread', 0.5946148633956909),
   ('dive', 0.5781358480453491)],
  'skipgram': [('drifted', 0.7524865865707397),
   ('monopoly', 0.7374759912490845),
   ('commision', 0.7335785031318665),
   ('totality', 0.7313978672027588),
   ('automata', 0.7307014465332031)],
  'withphrase': [('rumkovsky', 0.7212339639663696),
   ('repo_man', 0.7159184813499451),
   ('had_televison', 0.706023097038269),
   ('dave_hung', 0.6894766092300415),
   ('ancient_mayans', 0.6802935004234314)],
  'sg_withphrase': [('reasoning_behind', 0.9137635827064514),
   ('newbies', 0.913034200668335),
   ('dictionary_definition', 0.9118955135345459),
   ('preexisting', 0.9102602005004883),
   ('clinton_cripple', 0.9079402685165405)]},
 'train': {'baseline': [('finish', 0.7562305927276611),
   ('duty', 0.7505533695220947),
   ('floor', 0.7443188428878784),
   ('walk', 0.7408044338226318),
   ('hook', 0.7327516078948975)],
  'skipgram': [('unlocked', 0.7946608066558838),
   ('courthouse', 0.7939974069595337),
   ('bargain', 0.7931797504425049),
   ('shouts', 0.7855324745178223),
   ('sweat', 0.7839381694793701)],
  'withphrase': [('trip', 0.8584544658660889),
   ('lift', 0.8313921689987183),
   ('corner', 0.8234850168228149),
   ('seat', 0.8166527152061462),
   ('bat', 0.8150397539138794)],
  'sg_withphrase': [('bus_station', 0.8370068073272705),
   ('bargain', 0.8317971229553223),
   ('radiator', 0.8274469971656799),
   ('pavement', 0.8267682194709778),
   ('upstairs', 0.8212249279022217)]}}
```