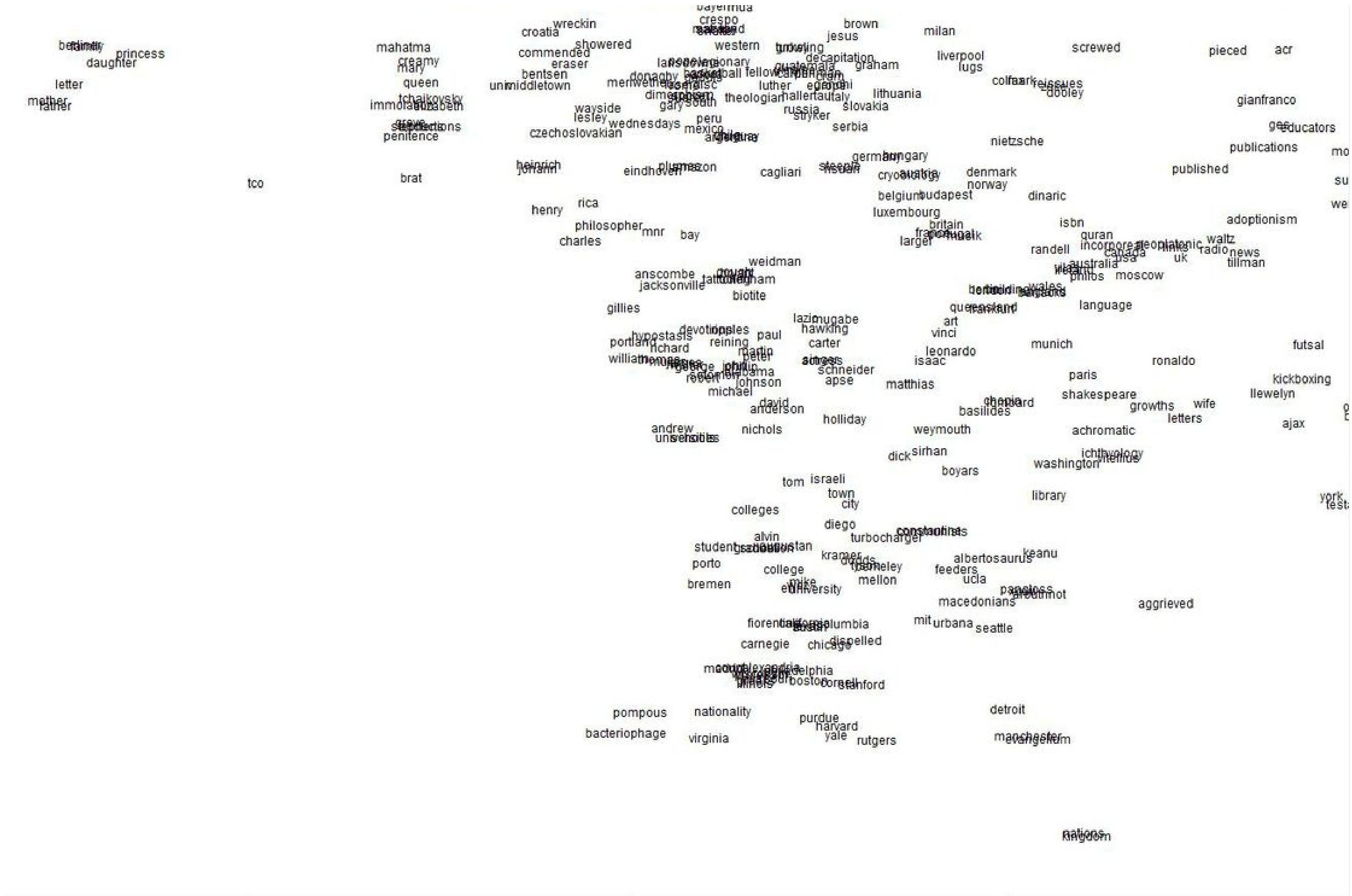
Distributed Representation of Documents with Explicit Explanatory Features: Pilot Test

October 7th, 2015

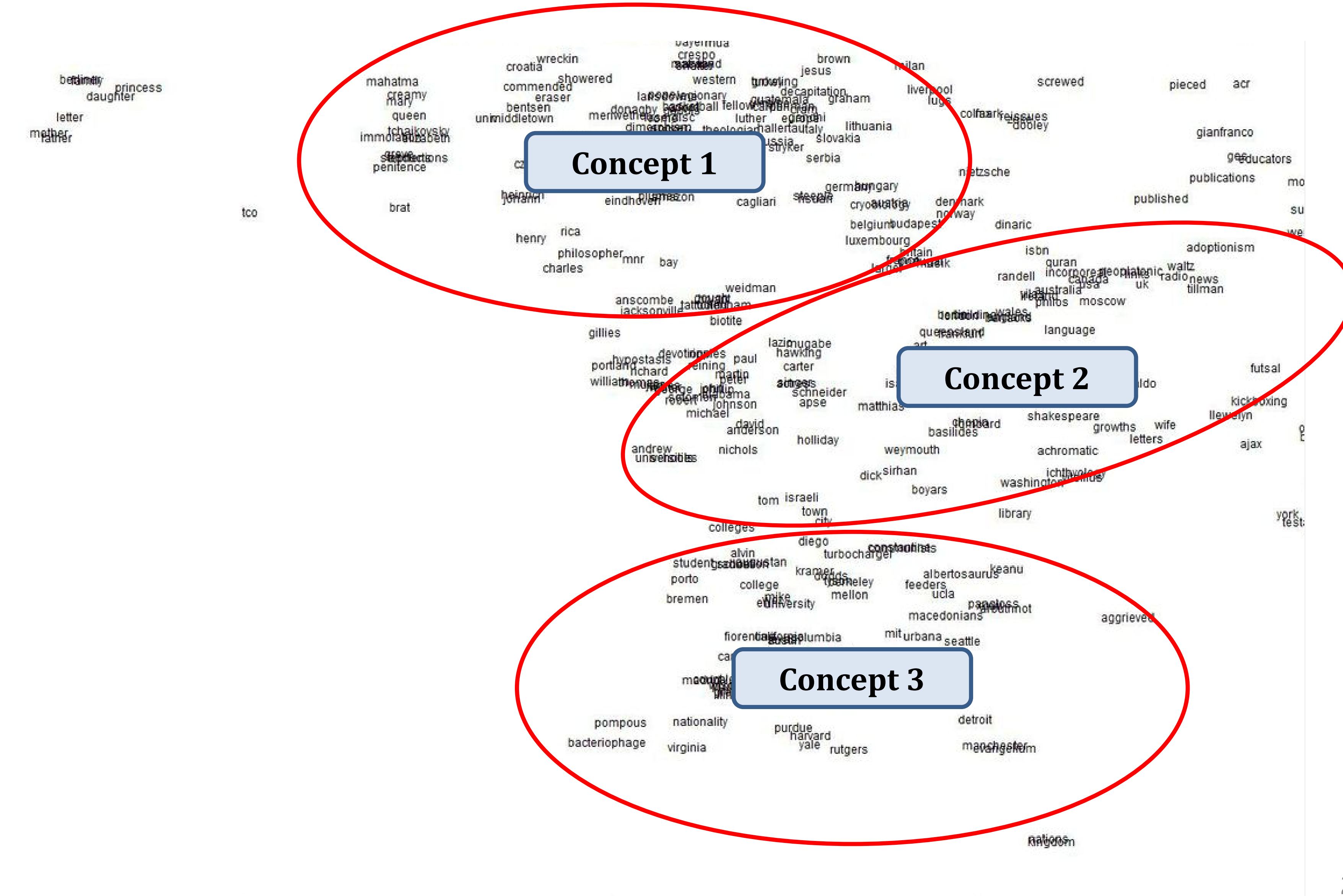
SNU Data Mining Center
Han Kyul Kim

Proposed Framework

1. Train Word2Vec with the collection of documents

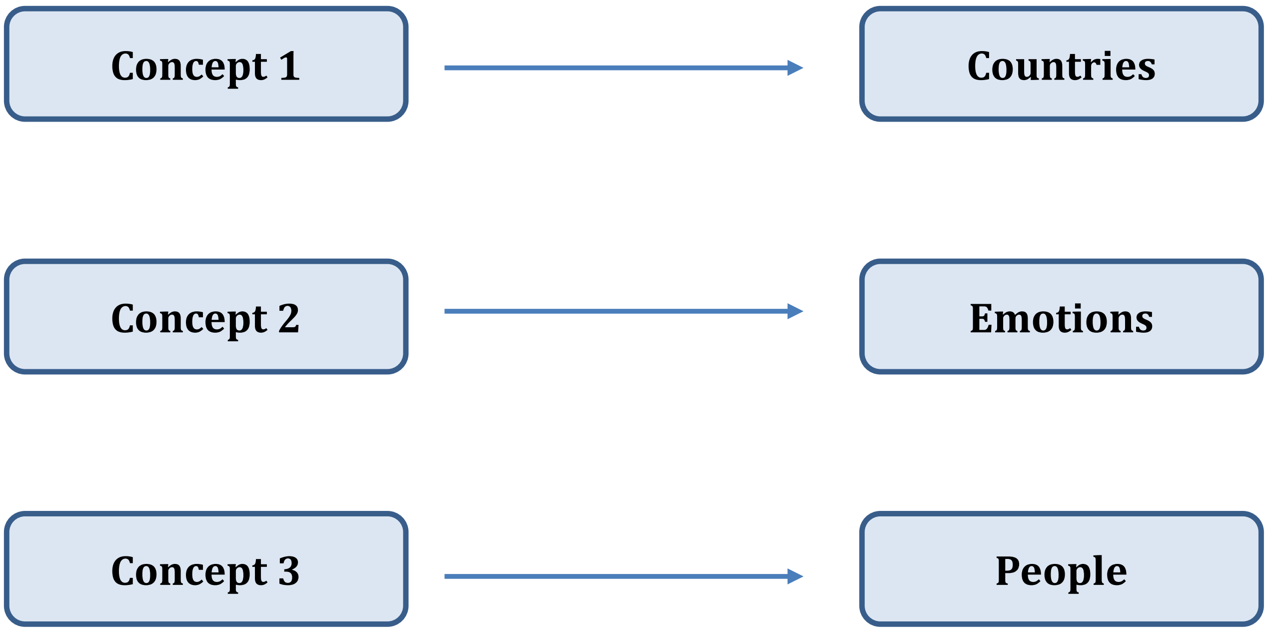


2. **Cluster word2vec generated vectors to create clusters of concepts**



Proposed Framework

3. **Label the concepts using the words associated with each cluster**



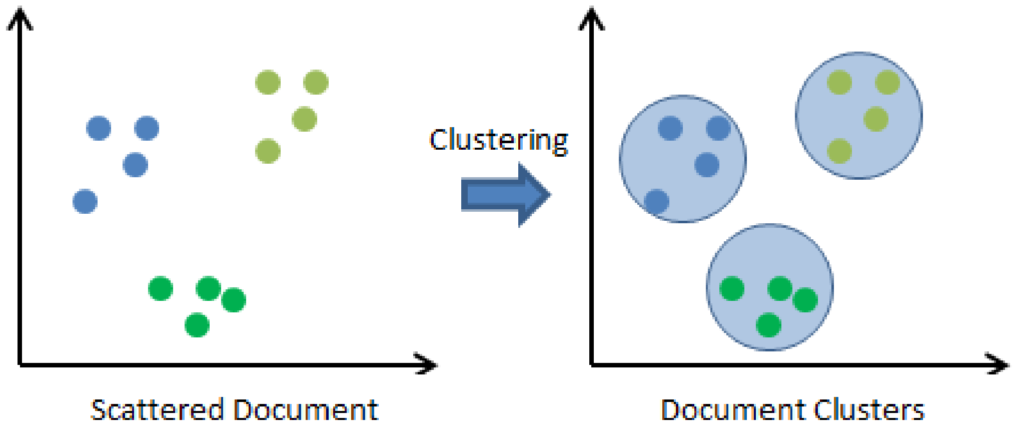
4. Represent the documents using the counts of these concepts

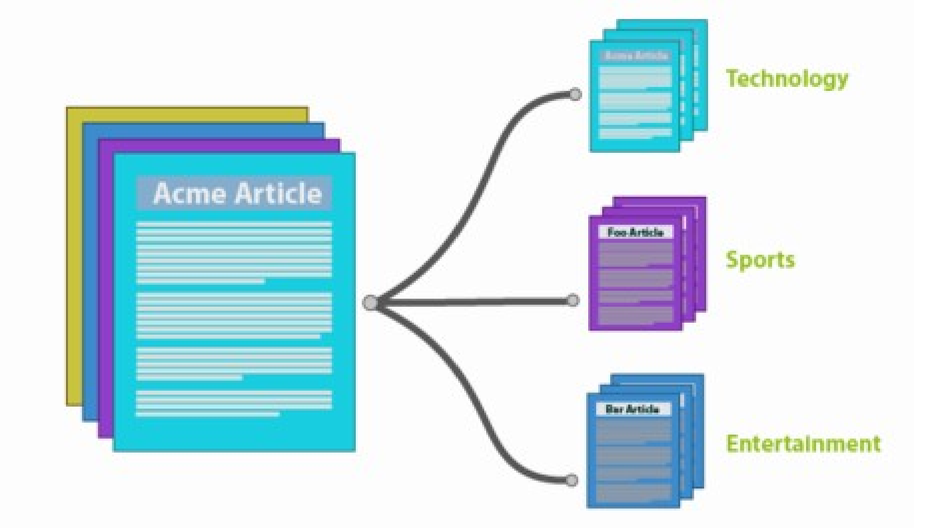
Doc 1 = [Countries, Emotions, People …]


```
[-0.08759557455778122, -0.04312118515372276, -0.08494572341442108, 0.024585919454693794, -0.05785191431641579, -0.02659076638519764, 0.04704275727272034, -0.03940117731690407, 0.005195754114538431, -0.
018994472920894623, -0.030896589159965515, -0.02599106915295124, -0.029802896082401276, -0.009517285041511059, -0.03624524921178818, 0.0029738633893430233, -0.04270448908209801, -0.0890769511461258, -0.
04064304754137993, 0.01777574978768825, 0.0910411849617958, 0.05333533510565758, -0.07692492008209229, 0.08628936856985092, -0.04232612252235412, -0.007681592833250761, 0.041417200118303, -0.
030358949676156044, 0.05717118829488754, 0.0396726056933403, -0.09482061862945557, 0.05382954701781273, -0.016189705580472946, 0.0013550696894526482, 0.004251557867974043, -0.1043981015682205, -0.
01734139770269394, 0.08733568340539932, -0.02014184184372425, 0.06905293464660645, -0.05219369381666835, -0.008379205130040646, 0.05078992992639541, -0.0521097406744957, 0.02524719014763832, -0.
09064795076847076, -0.01605154387652874, -0.08548879623413086, 0.09579522907733917, 0.07222563773393631, -0.01747663877904415, -0.07119490951299667, 0.04312814772129059, 0.006512957159429789, 0.
04662078991532326, 0.05369564145803451, 0.0017072528135031462, 0.0468018501996994, 0.03436211124062538, -0.05560864135622978, 0.04996718838810921, 0.09512156248092651, 0.004873079247772694, -0.
001331645064055919, 0.063605397939682, 0.024286553263664246, 0.009762034751474857, -0.04571126028895378, 0.032288745045661926, -0.06711595505475998, 0.03179262951016426, 0.0003801546408794803, 0.
013781636022031307, 0.049716588109731674, -0.04940832778811455, 0.04659050330519676, 0.08707352727651596, -0.10198234766721725, 0.0012964112684130669, 0.019826047122478485, 0.020798761397600174, 0.
03980609402060509, -0.01610562391579151, 0.09880314767360687, 0.03530235588550567, -0.03354038670659065, -0.06033261120319365, 0.009992017410695553, -0.07922962307929993, -0.04672875627875328, 0.
029249077662825584, 0.007386107929050922, 0.014454830437898636, 0.03679213672876358, 0.022595265880222709, -0.07544633001089096, 0.03752242028713226, -0.04298776760697365, 0.04293576627969742, -0.
04286639392375946, 0.05493065342307091, -0.010759958997368813, -0.02642848715186119, 0.05936839431524277, 0.010735556483268738, -0.022106561809778214, 0.005098363384604454, 0.02817825973033905, 0.
06781460344791412, 0.011524706147611141, 0.04529837146401405, -0.10414177924394608, -0.06333499401807785, -0.025369135662913322, 0.013805448077619076, 0.08876033872365952, -0.03237186372280121, 0.
08923118561506271, 0.035788971930742264, -0.0735849142074585, -0.11431705206632614, 0.004165078978985548, 0.03989437595009804, 0.011698578484356403, 0.009127369150519371, -0.007194945588707924, -0.
06279811263084412, -0.012036782689392567, 0.040613092482089996, -0.07241667807102203, 0.06803597509860992, 0.03180598095059395, 0.020111583173274994, 0.05419156327843666, -0.06192755699157715, 0.
03006706014275551, -0.015478908084332943, -0.05181756243109703, -0.027904510498046875, 0.01615394838154316, 0.0011802234221249819, -0.05243720859289169, -0.05809084326028824, 0.025693533942103386, -0.
04456381872296333, -0.05235142633318901, 0.002849023789167404, 0.02517843246459961, 0.07427450269460678, -0.004151491448283195, 0.06708431243896484, 0.010245034471154213, -0.03721074759960174, -0.
05194145068526268, 0.011150983162224293, -0.01369072962552309, -0.004474998451769352, 0.08407635986804962, 0.01842060498893261, 0.06935162097215652, -0.070859894156456, -0.04950730875134468, -0.
12054812163114548, -0.006143994629383087, 0.018353456631302834, 0.10588039457798004, 0.048505134880542755, -0.014145134948194027, -0.02343590557575226, -0.007739520631730556, -0.03350961208343506, 0.
02147112786769867, 0.07666371762752533, 0.06256308406591415, 0.012278590351343155, -0.017493827268481255, -0.025852585211396217, 0.02850286103785038, -0.04092925786972046, 0.1389494389295578, -0.
0005276211304590106, 0.019603941589593887, 0.047872498631477356, -0.02479643188416958, -0.03278869017958641, -0.03298942744731903, -0.0749734565615654, 0.017869295552372932, 0.014359974302351475, 0.
03260161727666855, -0.05898589268326759, 0.04998610913753509, -0.016536226496100426, -0.019838936626911163, -0.030345618724822998, -0.04567880555987358, -0.04636150225996971, -0.07739298790693283, 0.
05453772097826004, -0.01512290351092815, -0.05582341924309730, 0.02098872698843479, 0.0380556583404541, 0.041010051965713, 0.06500554829835892, -0.014411268755793571, 0.06324303895235062, 0.
12235700339078903, -0.003138007363304496, 0.04150061309337616, -0.05469806492328644, 0.028155071660876274, -0.02541017346084118, 0.02362082526087761, -0.055749375373125076, 0.034880928695201874, 0.
02925102412700653, -0.044924408197402954, -0.05796622857451439, 0.16091801226139069, -0.013727911747992039, -0.0028238031081855297, 0.005058911163359880, 0.009330375120043755, -0.040790800005197525, 0.
012631899677217007, -0.03498981520533562, -0.09075972437858582, -0.04068059101700783, 0.056468620896339417, 0.0793653130531311, -0.10637512058019638, 0.024917149916291237, 0.0766419768333435, 0.
08899114280939102, -0.07655515521764755, 0.018688995391130447, -0.03665241599082947, -0.007174948696047068, -0.026764938607811928, 0.006813205778598785, 0.03434935212135315, 0.06390520185232162, 0.
005475071258842945, 0.006454255897551775, 0.01241873949766159, -0.048750586807727814, -0.025715982541441917, -0.0013290401548147202, -0.0036538743879646063, -0.03493126481771469, 0.07465724647045135, 0.
04718988761305809, -0.027499066665768623, 0.011664669029414654, 0.020739786326885223, -0.0011999694397673001, 0.02283419854938984, -0.039574239403009415, 0.027619406580924988, 0.013819732703268528, 0.
00996309332549572, 0.01770842634141445, -0.013031989336013794, -0.007582968100905418, -0.0025542087387293577, 0.03596331179141998, 0.0165165513753891, -0.053140003234148026, -0.03133388236165047, 0.
07702472805976868, -0.06405989825725555, -0.05113033577799797, 0.11122194677591324, -0.0431693010032177, -0.0351727195084095, -0.06082572788000107, -0.05721887946128845, -0.04940216988325119, -0.
04693160206079483, -0.04560410603880882, -0.06396282464265823, -0.08668574690818787, -0.008351934142410755, 0.08004070073366165, 0.020836271345615387, 0.03663250431418419, 0.01230071671307087, -0.
042455870658159256, 0.006077070720493793, 0.0036929140333086252, 0.004990659188479185, -0.0027848768513649, -0.03410743921995163, -0.05878137424588203, 0.05231849849224090, 0.018009591847658157, 0.
05601579323410988, 0.06174979358911514, 0.009586947038769722, 0.10309688746929169, -0.0013724834425374866, -0.03734941035509109, -0.06568935513496399, -0.0019719060510396957, -0.02522912435233593, -0.
006819620728492737, -0.08099189400672913, 0.14139385521411896, 0.025998227298259735, 0.044450853019952774, 0.09344274550676346, -0.022272149100899696, -0.019415834918618202, 0.05529139563441276, -0.
03372661769390106, 0.07417678833007812, 0.0070419879630208015, -0.053876567631959915, 0.10055916011333466, -0.11293848603963852, -0.03376157209277153, 0.060266170650720596, 0.03926900401711464, -0.
035960644483566284, 0.008666090667247772, 0.027477947995066643, 0.03710789978504181, -0.023510152474045753, 0.03532886877655983, -0.002786519005894661, -0.021855372935533524, -0.028223032131791115, 0.
042905956506729126, -0.02157396450638771, -0.0291085597127676, 0.039419323205947876, -0.06341256201267242, -0.04983802139759064, 0.04110337421298027, -0.08459826558828354, -0.0303522776812315, 0.
03796708956360817, 0.17562396824359894, 0.05129672959446907, 0.005838119424879551, -0.05645650997757912, 0.02716675214469433, -0.024479210376739502, -0.05037922412157059, 0.04924848675727844, 0.
016620146110653877, -0.04153639078140259, 0.02921335957944393, 0.040044404566287994, 0.012756554409861565, -0.007722984999418259, -0.0386621430516243, 0.017636995762586594, -0.04599433019757271, -0.
05585790053009987, -0.00897784624248743, -0.07755934447050095, 0.02797551639378071, 0.05918818712234497, 0.009610161185264587, 0.09371864050626755, 0.006441071629524231, -0.015960507094860077, 0.
10366759449243546, 0.00836758129298687, 0.05804479494690895, 0.04077683761715889, -0.0030101246666163206, -0.007653041277080774, 0.12586210668087006, 0.02910740114748478, 0.0408298559486866, -0.
08000585436820984, -0.11633670330047607, -0.1095055490732193, 0.06872736662626266, -0.015693403780460358, 0.045604776591062546, 0.04468008130788803, 0.07531865686178207, 0.03042691200971603, -0.
0503707230091095, 0.009182695299386978, -0.06680766493082047, 0.032497283071279526, -0.03883403912186626, 0.028378712013363838, -0.06321480125188828, -0.005385537166148424, 0.11055602878332138, -0.
1312117725610733, -0.08003625273704529, -0.0033477982506155968, -0.08428100496530533, 0.020551182329654694, -0.027252040803432465, 0.04237489402294159, 0.03843776136636734, -0.015112183056771755, 0.
017832273617386818, 0.07243572175502777, 0.03400840610265732, 0.027588292956352234, -0.0224411990493536, -0.029101211577653885, 0.008950938470661640, -0.013047640211880207, 0.05460863187909126, -0.
0017075719079002738, -0.031082144007086754, -0.06595993787050247, -0.013669811189174652, -0.019473319873213768, -0.07515253126621246, 0.05773814767599106, 0.005721567664295435, 0.02362120896577835, 0.
040131740272045135, 0.07627321779727936, -0.04665115848183632, -0.043376557528972626, 0.061202943325042725, 0.004489867947995663, -0.020421821624040604, -0.044818829745054245, -0.0038449352141469717, -0.
05663484334945679, 0.03718886151909828, -0.02901551127433777, -0.03713458776473999, -0.05943121761083603, -0.06698095798492432, 0.012730694375932217, -0.021541433408856392, -0.018205733969807625, -0.
014323082752525806, -0.07051131874322891, 0.11712339520454407, 0.0007166709401644766, -0.06536946445703506, -0.015314971096813679, 0.03457680717110634, 0.061441387981176376, 0.02478879317641258, 0.
01627667434513569, -0.06662845611572266, -0.015765206888318062, -0.0037493400741368532, -0.004942044615745544, -0.042015254497528076, -0.07548461109399796, -0.03183511644601822, 0.061288982629776, -0.
06246870756149292, -0.0013078112388029695, -0.06325135380029678, -0.12265635281801224, 0.020725131034851074, 0.028045836836099625, 0.0168102215975523, -0.014592664316296577, 0.04328052327036857, -0.
```

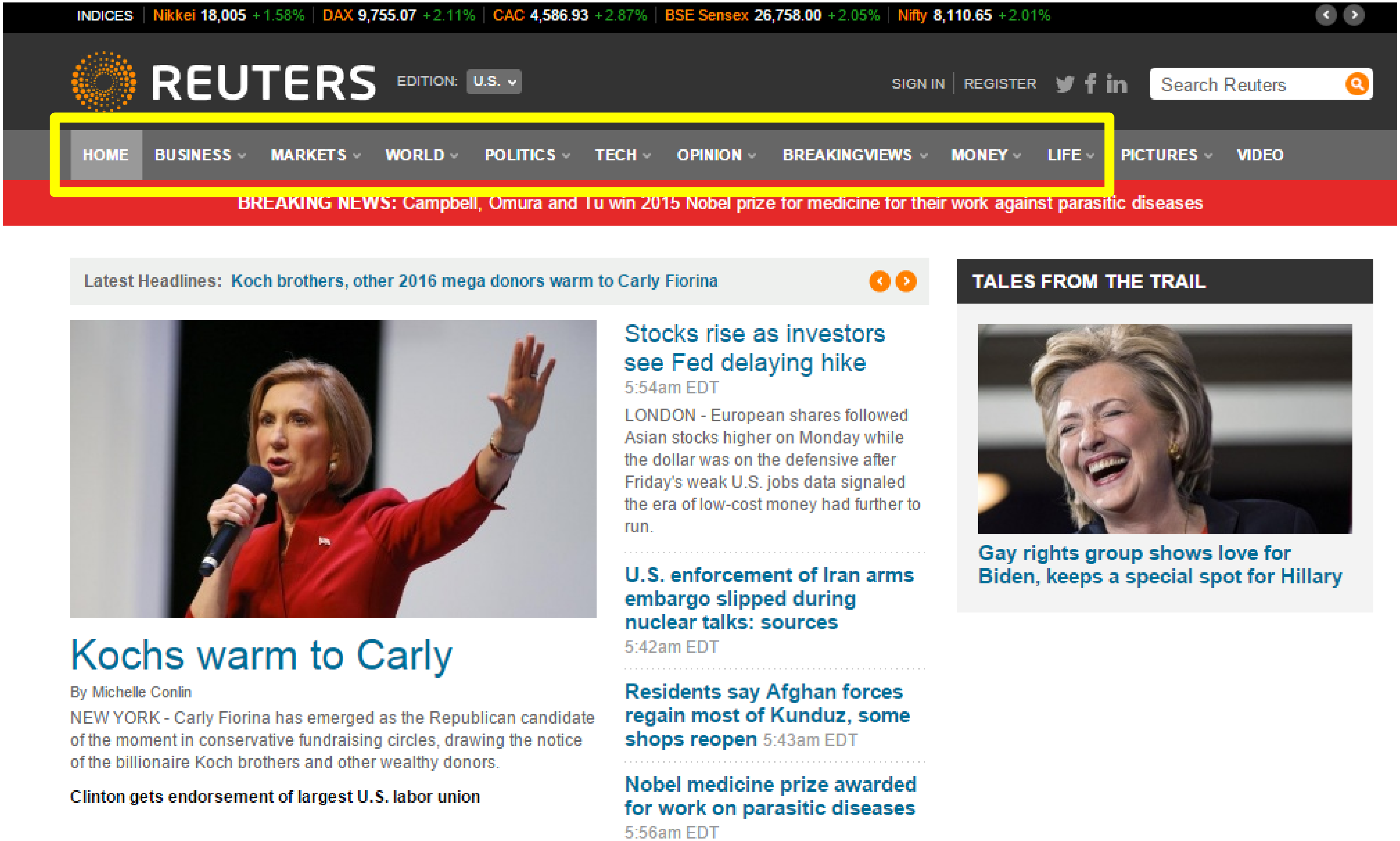
Proposed Framework

5. **Test the effectiveness of the document representation through document clustering and classification**





Dataset: <Reuters>



Dataset: <Reuters>

Total Number of Documents: 612,374 (2006. 09. 01 ~ 2015. 06. 06)

- Random sample of 10,000 documents from 5 distinctive categories (Sports, Market, Politics, Business, World, Health)

Categories	Total Number of Documents	Number of Sampled Document
Entertainment	25,764	-
Sports	**49,883**	**10,000**
Technology	26,899	-
Market	**189,399**	**10,000**
Oddly Enough	5,864	-
Politics	**42,319**	**10,000**
Business	**96,611**	**10,000**
Art	4,792	-
World	**138,852**	**10,000**
Health	**31,991**	**10,000**

Word2Vec & Doc2Vec Training

- During training, words that occur less than 20 times are discarded for stable result
  - Number of unique tokens: 67,390
- **Hyperparameters** to consider while training word2vec / doc2vec models:
  - How many epochs to iterate over the documents?
  - Averaging or concatenating the vectors in the hidden nodes?
  - Number of nodes in hidden layers (**dimension of embedding vectors**)
  - Number of windows (**number of words to use as contexts**)
- No universal hyperparameter settings exist as they are highly dependent on the characteristics and the amounts of the training corpus
- Two types of evaluation methods for word2vec methods:
  1. <u>Extrinsic Evaluation</u>
     - Since the embedded vectors are used as ingredients for building more complex task-specific language model(usually as a pre-training step), evaluation on actual real task
     - Take a long time to compute accuracy
     - Unclear if the subsystem is the problem or its interaction or other subsystems are the problems
  2. <u>Intrinsic Evaluation</u>
     - Evaluation on a specific/intermediate subtask
     - Fast to compute
     - Not helpful unless correlation to real task is established

Word2Vec & Doc2Vec Training

- Trained 100 different Word2Vec models
  - Dimension: 200 ~ 800
  - Context Window: 6 ~ 10
- Chose the model with the best accuracy in intrinsic evaluation criteria provided by Tomas Mikolov
  - A list of four words with specific relationship is given (19,558 analogies)
    - Capital-Country, Opposing words, Nationalities
  - Given only three words out of four words, test if the model can produce a correct answer
- Chosen hyperparameters:
  - Dimension: 550 & Window Size: 9

```
: capital-common-countries
Athens Greece Baghdad Iraq
Athens Greece Bangkok Thailand
Athens Greece Beijing China
Athens Greece Berlin Germany
Athens Greece Bern Switzerland
Athens Greece Cairo Egypt
Athens Greece Canberra Australia
Athens Greece Hanoi Vietnam
Athens Greece Havana Cuba
Athens Greece Helsinki Finland
Athens Greece Islamabad Pakistan
Athens Greece Kabul Afghanistan
Athens Greece London England
```

```
find finds generate generates
find finds go goes
find finds implement implements
find finds increase increases
find finds listen listens
find finds play plays
find finds predict predicts
find finds provide provides
find finds say says
find finds scream screams
find finds search searches
find finds see sees
find finds shuffle shuffles
```

Issues with High Dimensional Clustering

- As the dimension of vectors (data points) grow, normal clustering method doesn't work due to following reasons:

1. **Distance metric becomes useless (no sense of proximity)**
   - Under various data distribution and distance function, ratio of distances of the nearest and farthest neighbors to a given target in high dimensional space is almost 1.
   - Meaningfulness of L_k norm worsens faster with increasing dimensionality for higher values of k

2. **Problems associated with local feature relevance or local feature correlation**
   - Presence of irrelevant features or of correlations among subsets of features heavily influences the appearance of clusters in the full dimensional space
   - Dimension reduction cannot be applied as it only considers one subspace of the original data space in which the clustering can be performed

Spherical K Means

- To deal with these issues of high dimensional clustering, spherical k means algorithm has been used
- Essentially same as k means algorithm but with cosine similarity as a measure of proximity instead of Euclidean distance

Algorithm: spherical k-means (SPKM)
Input: A set of N *unit-length* data vectors $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ in $\mathbb{R}^d$ and the number of clusters K.
Output: A partition of the data vectors given by the cluster identity vector $\mathcal{Y} = \{y_1, ...y_N\}$, $y_n \in \{1, ..., K\}$.
Steps:

1. Initialization: initialize the *unit-length* cluster centroid vectors $\{\mu_1, ..., \mu_K\}$;
2. Data assignment: for each data vector $\mathbf{x}_n$, set $y_n = \arg\max_k \mathbf{x}_n^T \mu_k$;
3. Centroid estimation: for cluster k, let $\mathcal{X}_k = \{\mathbf{x}_n | y_n = k\}$, the centroid is estimated as $\mu_k = \sum_{\mathbf{x} \in \mathcal{X}_k} \mathbf{x} / \| \sum_{\mathbf{x} \in \mathcal{X}_k} \mathbf{x} \|$;
4. Stop if $\mathcal{Y}$ does not change, otherwise go back to Step 2a.

Fig. 1. Spherical k-means algorithm.

Choosing K

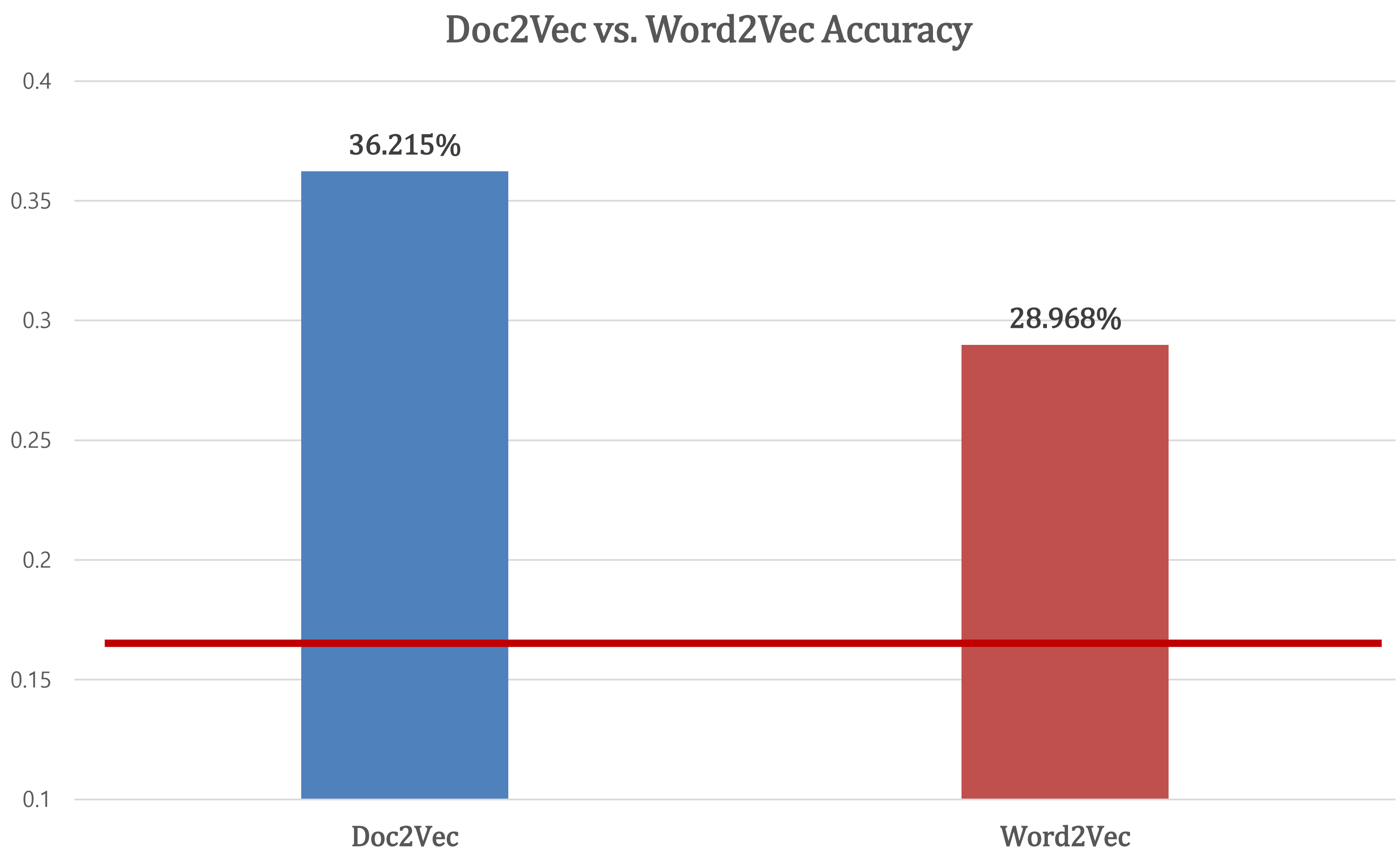
- Concept of inertia was used for selecting K
    - Inertia: Within-cluster difference with the centroid

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_j - \mu_i||^2)$$

- Cluster the embedded word vectors for various values of K and select the one with the lowest average inertia per cluster
- For each experiment with designated k value, best result out of 300 different trials (different initial points) was compared
- **K = 120 selected**
- Document clustering also used same algorithm except the value of K was fixed at 6

Clustering Result

- **Accuracy**
  - Perhaps need another representation method for comparison instead of random selection baseline



Cluster Labels

- Name of companies

delphi's	2
navistar's	2
telstra	2
agrium	2
elan's	2
terra's	2
tiscali	2
nonvoting	2
break-up	2
avon's	2
genzyme's	2
mosaic's	2
discovery's	2
wynn	2
elan	2
firstgroup	2
sprint.	2
drahi	2
sterlite	2
unitymedia	2

- People

kids	4
olds	4
mentors	4
academic	4
academia	4
cooks	4
enroll	4
professors	4
rehab	4
motivational	4
workplace	4
disabilities	4
telehealth	4
courses	4
navigators	4
underemployme	4
neediest	4
educational	4
bilingual	4
vaccinations	4
physicians	4
autistic	4
undergraduate	4
nurses	4

- Negative Words

belligerence	9
brainwashed	9
unjust	9
sickening	9
committing	9
oaths	9
tyranny	9
obscenity	9
gypsies	9
inhuman	9
rapes	9
intent	9
betraying	9
revulsion	9
neglect	9
punishable	9
instigator	9
deplore	9
mindless	9
abuses	9

Cluster Labels

- Name of countries

bratislava	73
wales	73
czech	73
energy-rich	73
catalonia	73
angola.	73
overlord	73
uae.	73
singapore	73
thailand	73
lesotho	73
lanka	73
jordan.	73
fiji	73
malta.	73
indonesia.	73
andorra	73

- Numbers

272	14
275	14
excl	14
2025s	14
393	14
395	14
398	14
29nov	14
270	14
273	14
279	14
305	14
796	14
795	14
790	14
830	14
911	14
199	14
198	14
195	14
194	14

- Food

cooking	18
intake	18
concoction	18
smithfield	18
kft	18
wholesome	18
rice	18
fruits	18
lunches	18
breakfast	18
h.j.	18
one-cup	18
hallucinogenic	18
shoots	18
mango	18
salads	18
buffet	18
tortilla	18
steamed	18

Final Experiment Setup

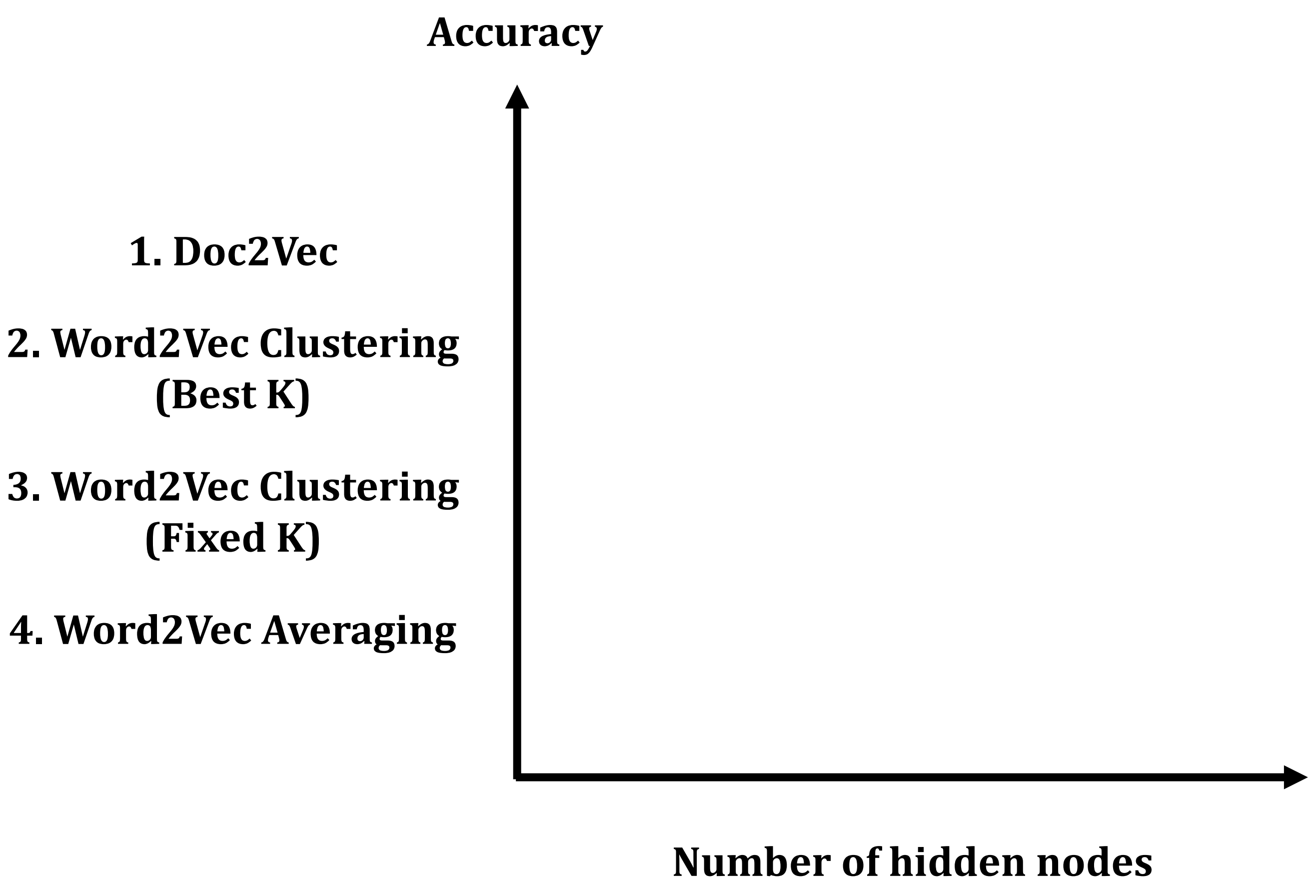
- Total number of documents used: 204,000

Categories	Total Number of Documents	Number of Selected Document
Entertainment	25,764	25,500
Sports	49,883	25,500
Technology	26,899	25,500
Market	189,399	25,500
Oddly Enough	5,864	-
Politics	42,319	25,500
Business	96,611	25,500
Art	4,792	-
World	138,852	25,500
Health	31,991	25,500

To Do & Issues

1. Need for another representation method for comparison as a baseline
    - Possible candidates: averaging word embedding vectors, tf-idf

2. Need to compare the accuracy of clustering & classification task given different values of K and the number of hidden nodes



To Do & Issues

3. Need to come up with a method that can automatically create labels for each word2vec clusters
   - Probably based on hypernym found by Wordnet (or using the counts)
   - But the issues still remain with pronouns and stemming
   - More profound background research is needed

4. Find examples of misclassifications in Doc2Vec representation that can be explained by word2vec clusters
   - Qualitatively substantiating the explanatory power of the suggest method

5. Finish the experiment for the final test settings and submit a preliminary results for the conference in November

Reference

Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. Springer Berlin Heidelberg, 2001.

Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek. "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering." ACM Transactions on Knowledge Discovery from Data (TKDD) 3.1 (2009): 1.

Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." arXiv preprint arXiv:1405.4053 (2014).

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in Neural Information Processing Systems. 2013..

Zhong, Shi. "Efficient online spherical k-means clustering." Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on. Vol. 5. IEEE, 2005.