Distributed Representation of Documents with Explicit Explanatory Features

Han Kyul Kim, Hyunjoong Kim, Sungzoon Cho

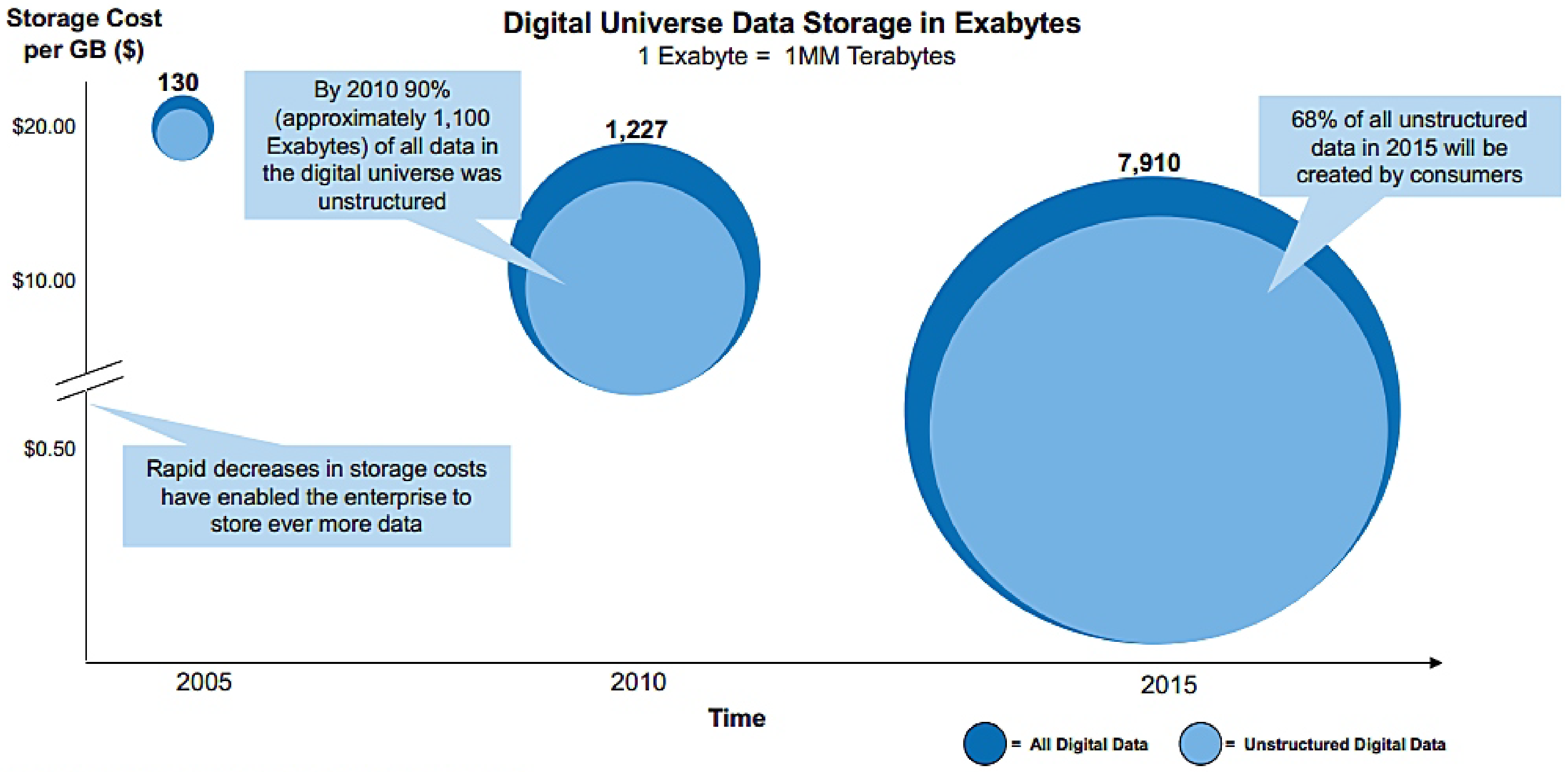
Dep. of Industrial Engineering

Seoul National University

Ascendance of "Unstructured Data"

- Due to the proliferation of smart phones and decreasing data storage cost, the amount of data being generated has drastically increased over the years

- Unstructured data is a realm of unexplored potential that provides an insight into consumers' psyche, as most of them are being generated directly from the consumers,



By 2010 90% (approximately 1,100 Exabytes) of all data in the digital universe was unstructured

68% of all unstructured data in 2015 will be created by consumers

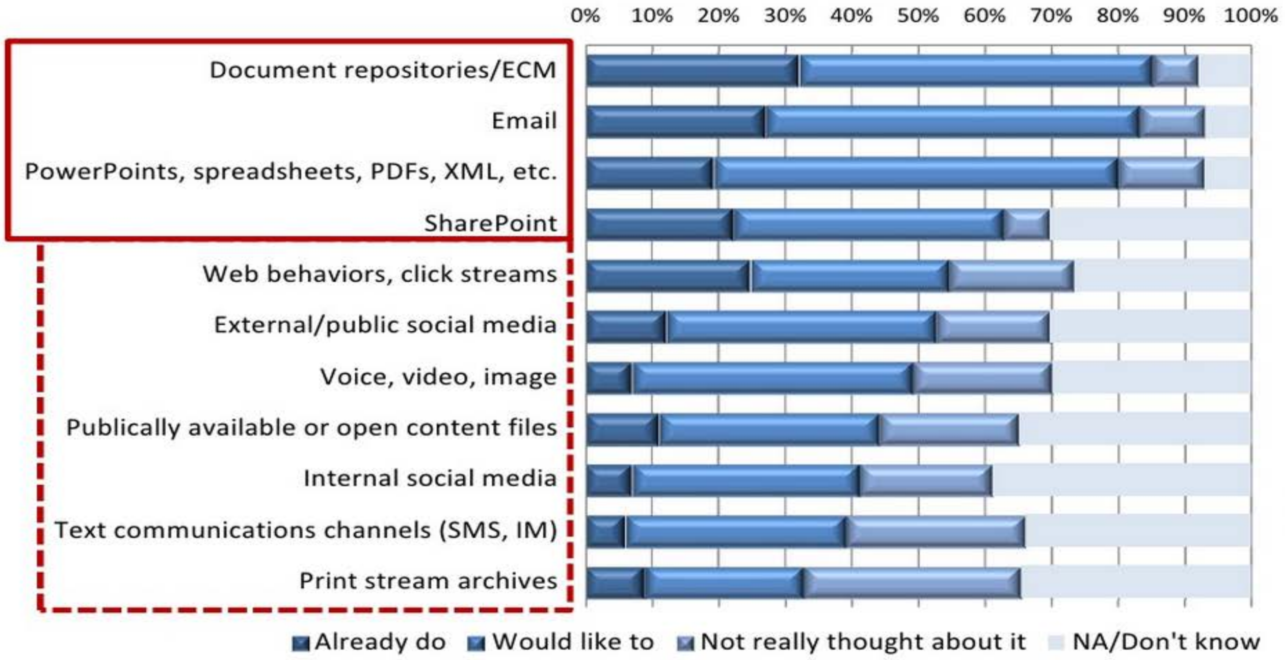
Rapid decreases in storage costs have enabled the enterprise to store ever more data

Source: IDC's Digital Universe Study, sponsored by EMC, June 2011

What kind of unstructured data are companies interested in?

- Text data, as one of the most common form of unstructured data, has generated great interest amongst companies
- Extracting insight through text mining essentially begins with analyzing documents

Are there large unstructured or semi-structured data repositories (ie, text, rich media, etc.) in your business that you would like to analyze, monitor or query - as opposed to search/retrieve?



©AIIM 2012

N=300

Then, how should we represent a document?

- Words
- Paragraphs
- Themes
- Main arguments

……

Doc 1: [1, 4, 3, 0, …]
Doc 2: [5, 2, 8 0, …]
………

- A document is a complex system composed with words, paragraphs and many explicit and implicit features

- In order for computers to extract insights from such complex system, documents need to be represented in some kind of numerical format

Bag of Words Approach

- Bag of Word Hypothesis: frequencies of words in a document tend to indicate the relevance of the documents

- Basically, BOW uses token frequencies as features for representing a document

[Document 1]:

Arsenal's annual injury problem is underway. Their thin squad will be put to the test by a Swansea City team looking to build on a vital win at Aston Villa.

Arsenal have a whole host of injury problems to contend with. The Gunners currently sit top of the Premier League's infamous injury table. Eight senior players will be unable to take part at the Liberty Stadium

	Arsenal's	**Annual**	**Injury**	**Problem**	**is**	**…**
Document 1	1	1	3	1	1	…

Bag of Words Approach

- Pros: It’s intuitive!
  - A feature of document vectors uniquely correspond to a specific token
  - Users can directly understand the components of the documents
  - It can provide clear explanation for why and how certain documents are different from each other

- Cons: Dimension of a vector increases quickly
  - A number of words in a document is NOT small
  - Words with similar meanings (not necessarily synonyms) and identical words with different grammatical structures are considered as individual tokens

[Document 1]:

Arsenal’s annual injury **problem** is underway. **Their** thin squad will be put to the test by a Swansea City team looking to build on a vital win at Aston Villa.

Arsenal have a whole host of injury **problems** to contend with. **The Gunners** currently sit top of the Premier League’s infamous injury table. Eight senior players will be unable to take part at the Liberty Stadium

Distributed Representation: Word2Vec & Doc2Vec

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

- Distributed Hypothesis: words that occur in similar contexts (neighboring words) tend to have similar meanings

- Based on this assumption, simple neural network is used to embed words into continuous vector space

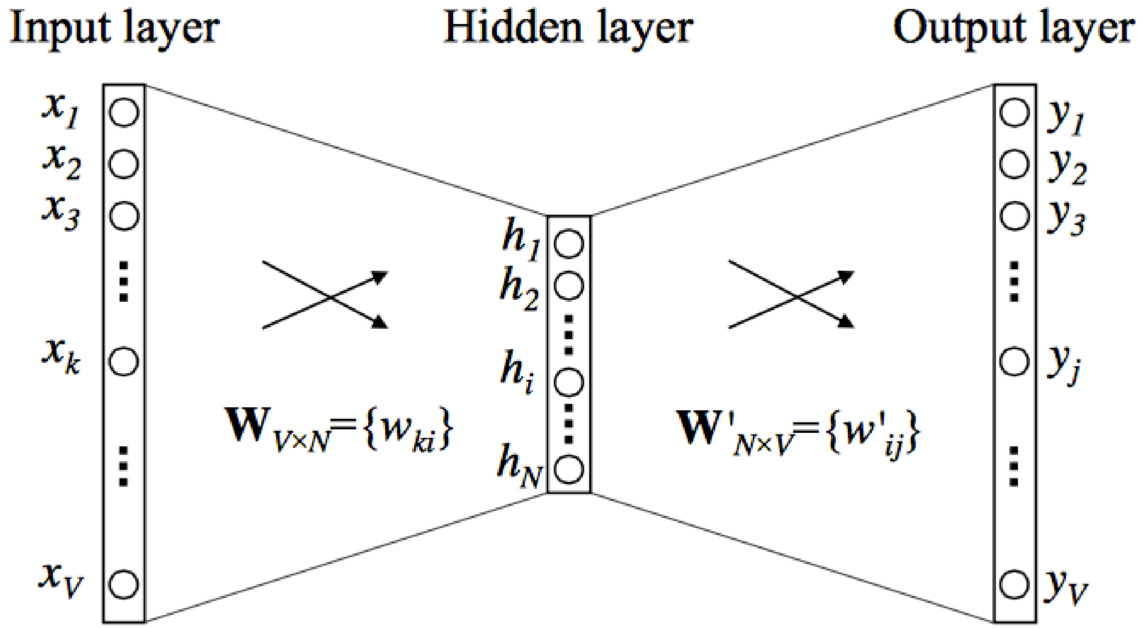
- Train the weights of the network so that an input word to the network can predict its neighboring words within certain window size

Simple Word2Vec Architecture

- Amongst vocabulary of size V, let's say we want to predict one target word(output) when we are given one context word (input) → bigram structure

- Input vector is an one-hot encoded vector (only one node with for designated context word will be 1)

- Weight matrix W_{VXN} and W'_{NXV} embed context and word into continuous vector space, respectively



Simple Word2Vec Architecture

- Non-linearity (soft-max) is incorporated only for computing the output

- v_{w_I}: vector representation of the input context word w_I ($x^T W$, k-th row of W)

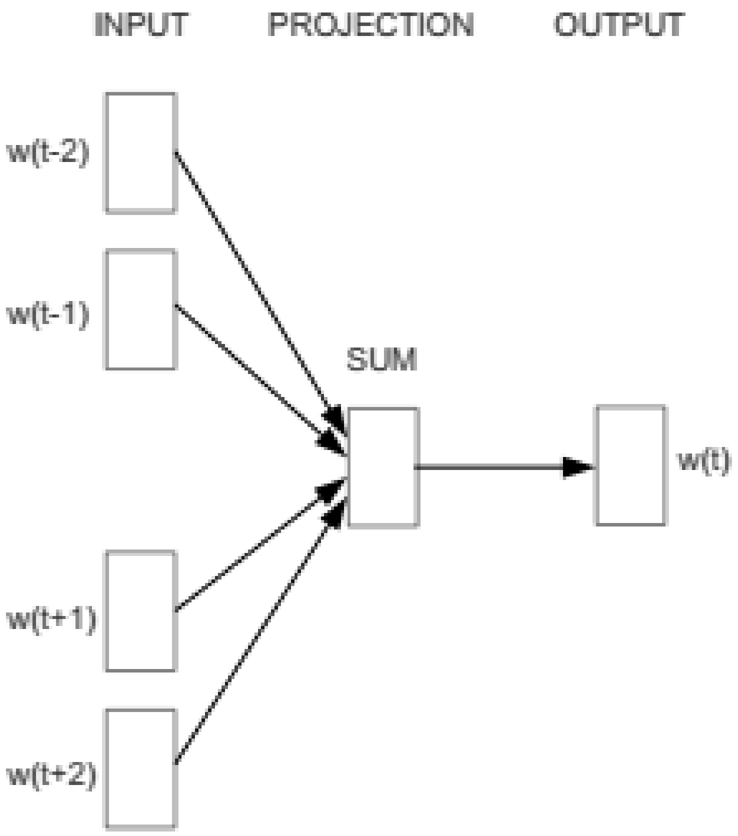
- $v'_{w'_j}$: vector representation of the output word (j-th column of W’)

- Train the weights via gradient descent to maximize conditional (log) probability of observing the actual output word w_O given the input context word w_I

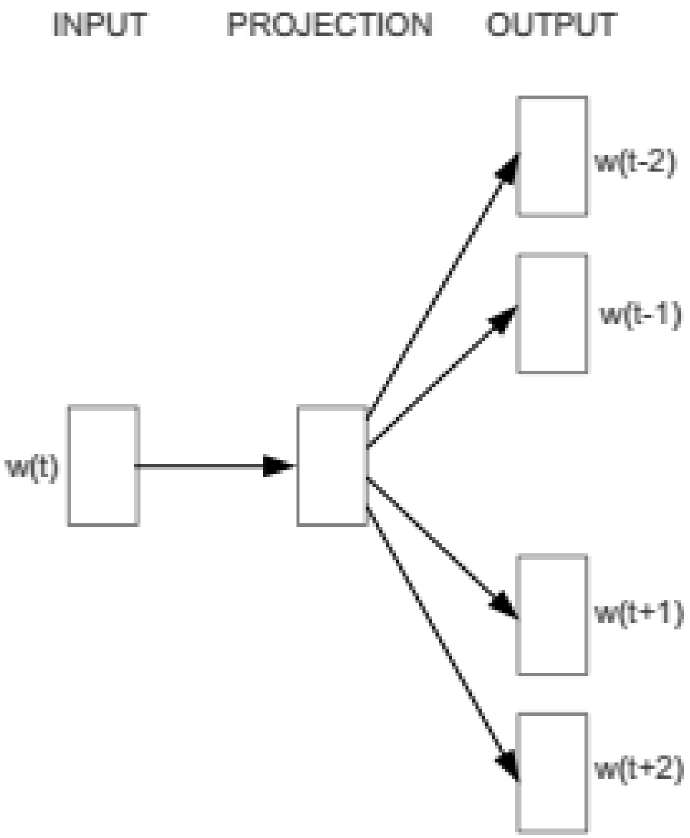
- As a result, word vectors with similar context (thus, meaning) will be located close to each other in the embedding space

$$p(w_j|w_I) = \frac{\exp\left({\mathbf{v}'_{w_O}}^T \mathbf{v}_{w_I}\right)}{\sum_{j'=1}^{V} \exp\left({\mathbf{v}'_{w'_j}}^T \mathbf{v}_{w_I}\right)}$$

Complete Word2Vec Architecture



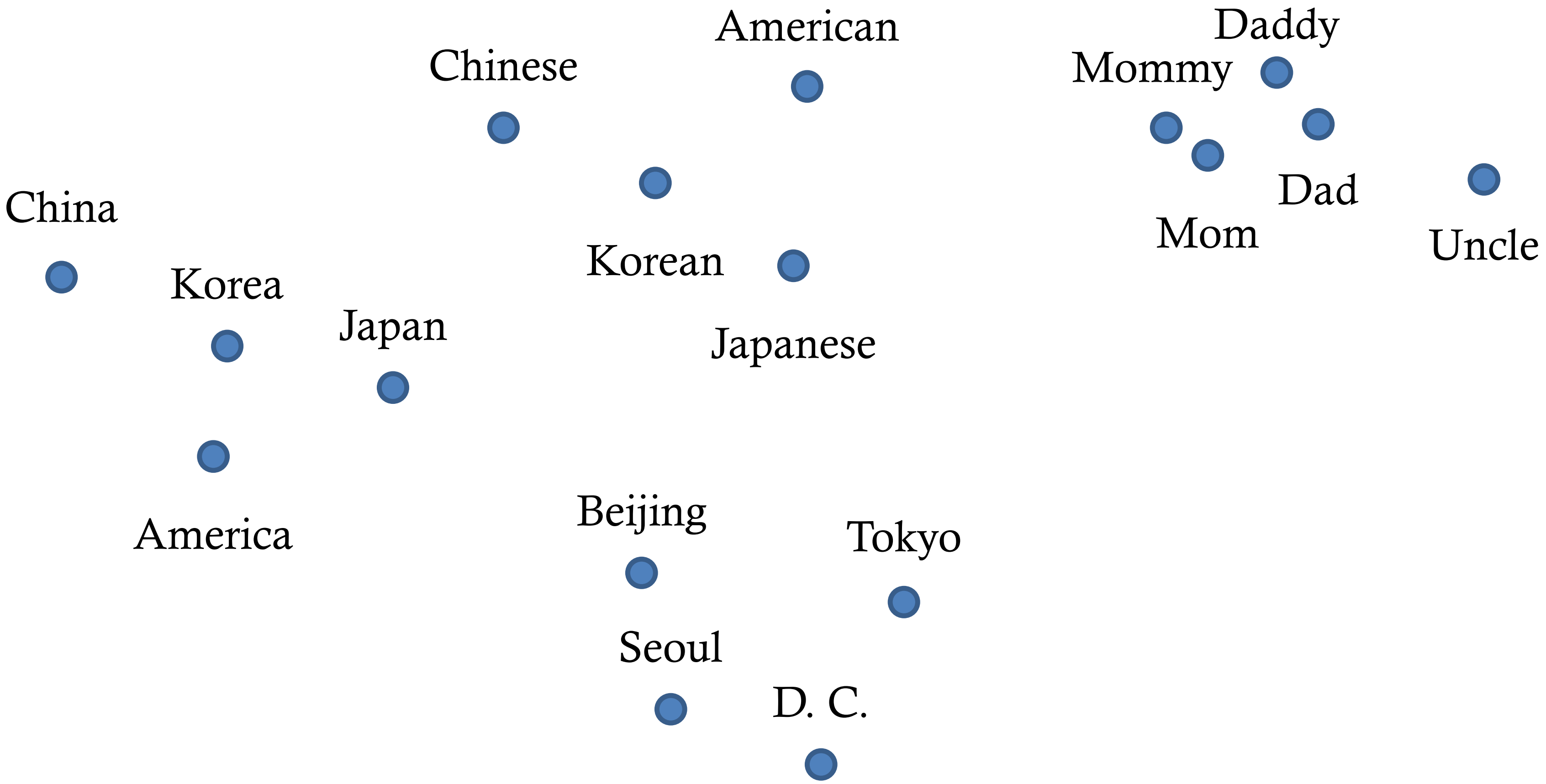
CBOW



Skip-gram

Result of Word2Vec

Example



- Word vectors with similar context (thus, belonging to a similar concept) will be located close to each other in the resulting embedding space

Using Word2Vec for Document Representation

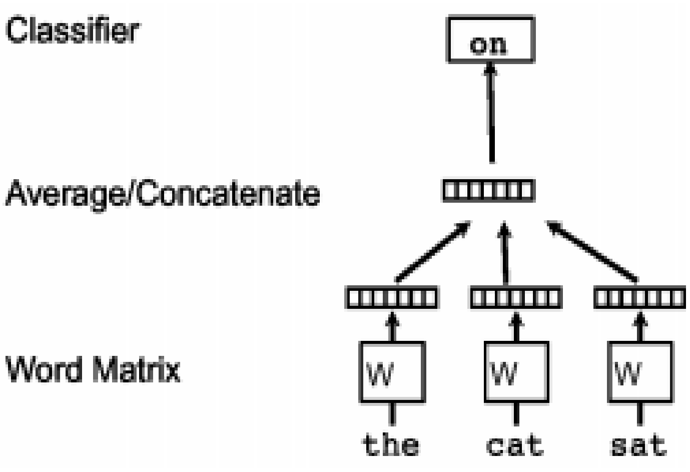
Xing, Chao, et al. "Document classification with distributions of word vectors." Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA). IEEE, 2014.

- **Simple average pooling approach:**

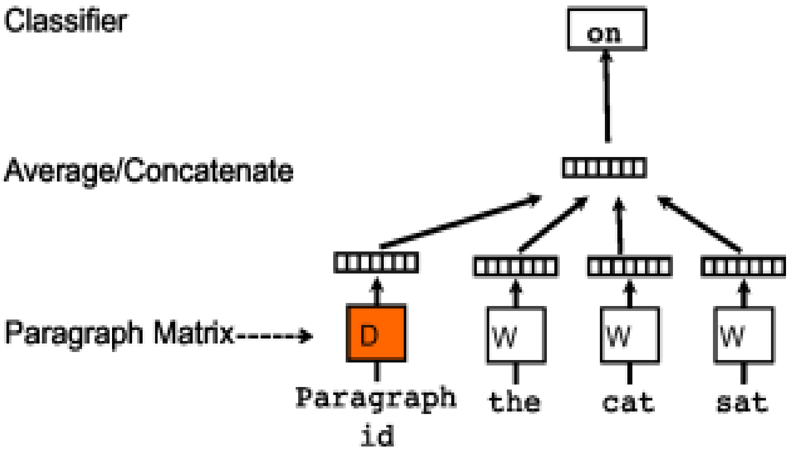
$$v_i = \frac{1}{J_i} \sum_{j=1}^{J_i} c_{i,j}$$

  - Derives a document vector as the centroid of word vectors within the document
  - Bias towards words without significant contribution to representing the semantics of the documents

Doc2Vec



<Word2Vec>



<Paragraph2Vec>

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents

- Extension of Word2Vec: a document is considered as an extra word
- Document (paragraph) id represents one-hot encoded vector of documents
- As a result, documents are also embedded into continuous vector space

Next Generation of Document Representation?

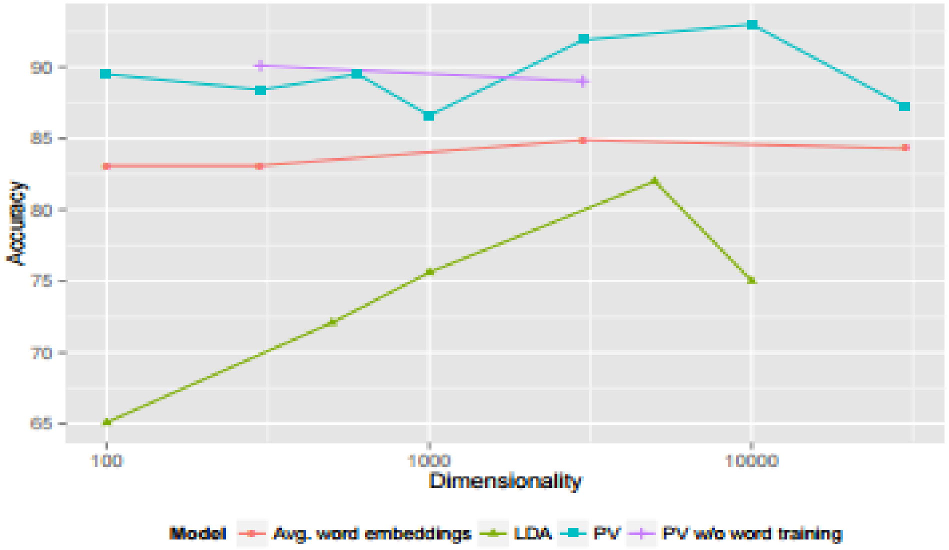
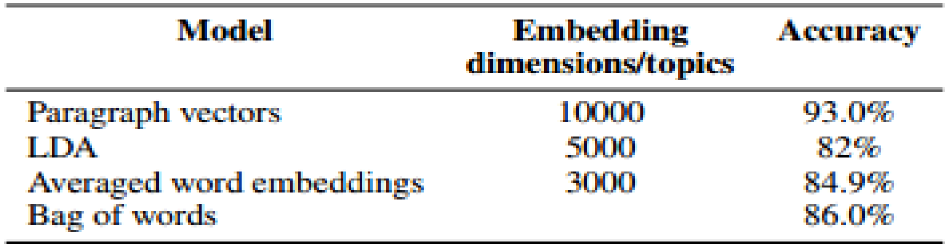


Figure 4: Results of experiments on the hand-built Wikipedia triplet dataset.

Table 3: Performances of different methods on hand-built triplets of Wikipedia articles on the best performing dimensionality.



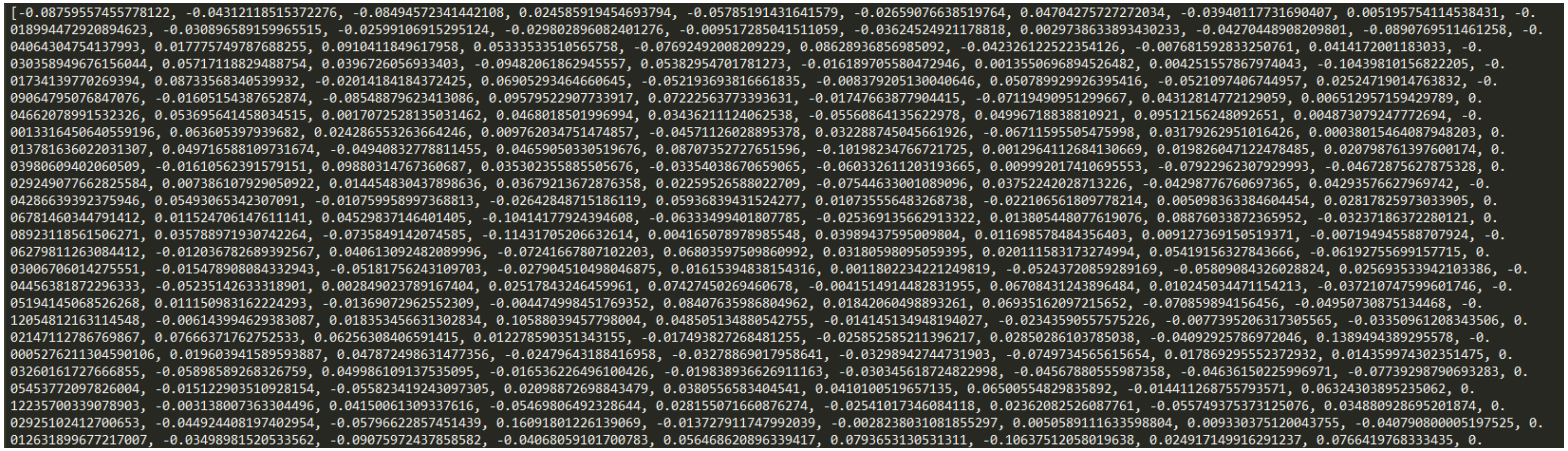
Model	Embedding dimensions/topics	Accuracy
Paragraph vectors	10000	93.0%
LDA	5000	82%
Averaged word embeddings	3000	84.9%
Bag of words		86.0%

Dai, Andrew M., et al. "Document Embedding with Paragraph Vectors." NIPS Deep Learning Workshop. 2014.

Distributed Representation Approach

- <u>Pros:</u> Good Performance!
  - Document vector successfully captures useful yet unknown features for representing documents (validated through document clustering and classification)
  - Dimension of a document vector is restricted to certain size

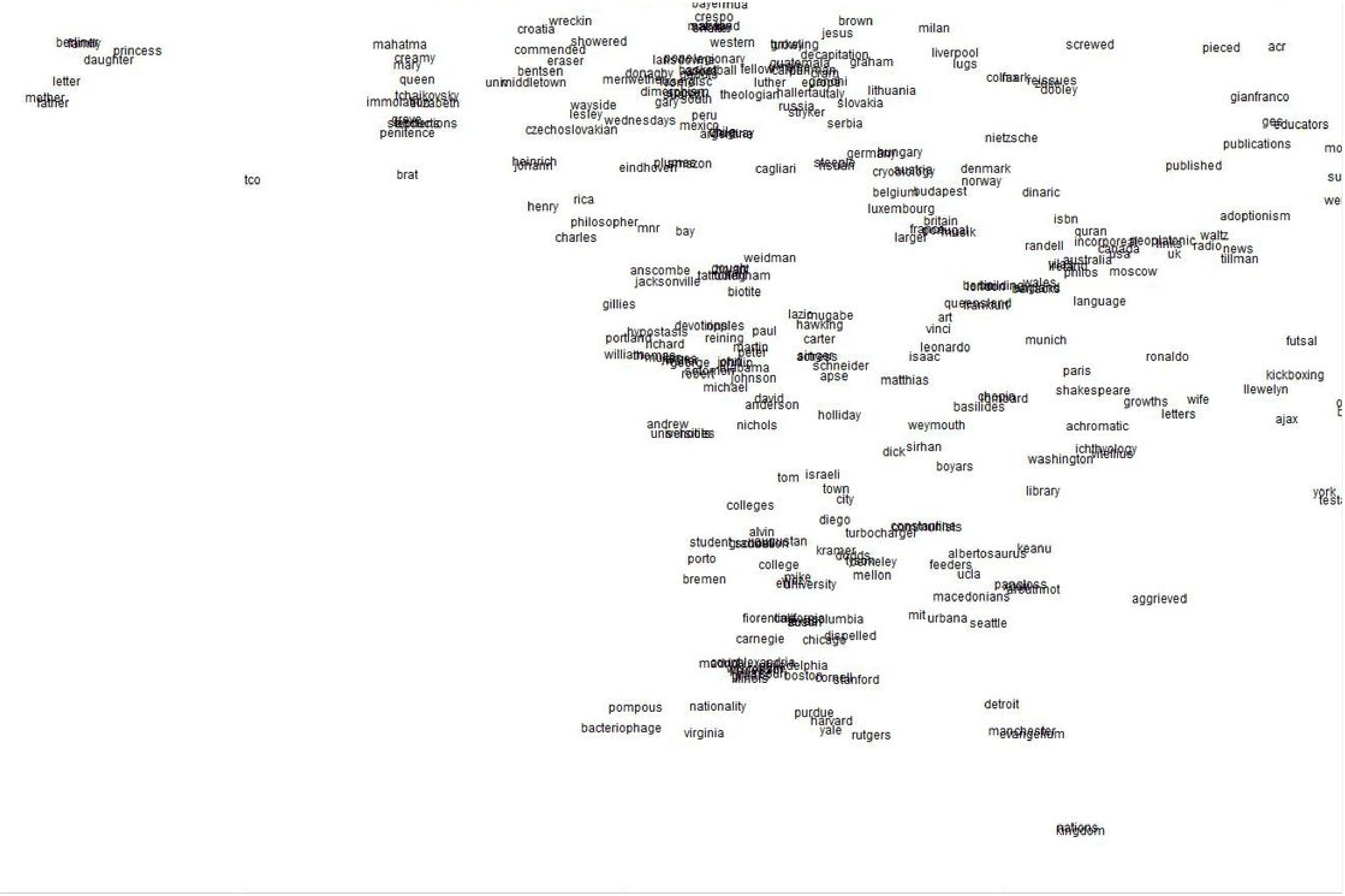
- <u>Cons:</u> Not intuitive!
  - As with other neural network based models, unknown feature are used to represent a document
  - Each value in a document vector doesn't provide any explicit explanation about the document



```
[-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, 0.01777574978768255, 0.0910411849617958, 0.05333533510565758, -0.07692492008209229, 0.08628936856985092, -0.042326122522354126, -0.007681592833250761, 0.0414172001183033, -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, -0.05219369381661835, -0.008379205130040646, 0.050789929926395416, -0.0521097406744957, 0.02524719014763832, -0.
09064795076847076, -0.01605154387652874, -0.08548879623413086, 0.09579522907733917, 0.07222563773393631, -0.01747663877904415, -0.07119490951299667, 0.04312814772129059, 0.006512957159429789, 0.
04662078991532326, 0.053695641458034515, 0.0017072528135031462, 0.0468018501996994, 0.03436211124062538, -0.05560864135622978, 0.04996718838810921, 0.09512156248092651, 0.004873079247772694, -0.
0013316450640559196, 0.063605397939682, 0.024286553263664246, 0.009762034751474857, -0.04571126028895378, 0.032288745045661926, -0.06711595505475998, 0.03179262951016426, 0.00038015464087948203, 0.
013781636022031307, 0.049716588109731674, -0.04940832778811455, 0.04659050330519676, 0.08707352727651596, -0.10198234766721725, 0.0012964112684130669, 0.019826047122478485, 0.020798761397600174, 0.
03980609402060509, -0.01610562391579151, 0.09880314767360687, 0.035302355885505676, -0.03354038670659065, -0.060332611203193665, 0.009992017410695553, -0.07922962307929993, -0.04672875627875328, 0.
029249077662825584, 0.007386107929050922, 0.014454830437898636, 0.03679213672876358, 0.02259526588022709, -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, -0.0041514914482831955, 0.06708431243896484, 0.010245034471154213, -0.037210747599601746, -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 5, -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, 0.049986109137535095, -0.016536226496100426, -0.019838936626911163, -0.030345618724822998, -0.04567880555987358, -0.04636150225996971, -0.07739298790693283, 0.
05453772097826004, -0.015122903510928154, -0.055823419243097305, 0.020988726988434790, 0.0380556583404541, 0.0410100519657135, 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, 0.0050589111633598804, 0.009330375120043755, -0.040790800005197525, 0.
012631899677217007, -0.03498981520533562, -0.09075972437858582, -0.04068059101700783, 0.056468620896339417, 0.0793653130531311, -0.10637512058019638, 0.024917149916291237, 0.0766419768333435, 0.
```

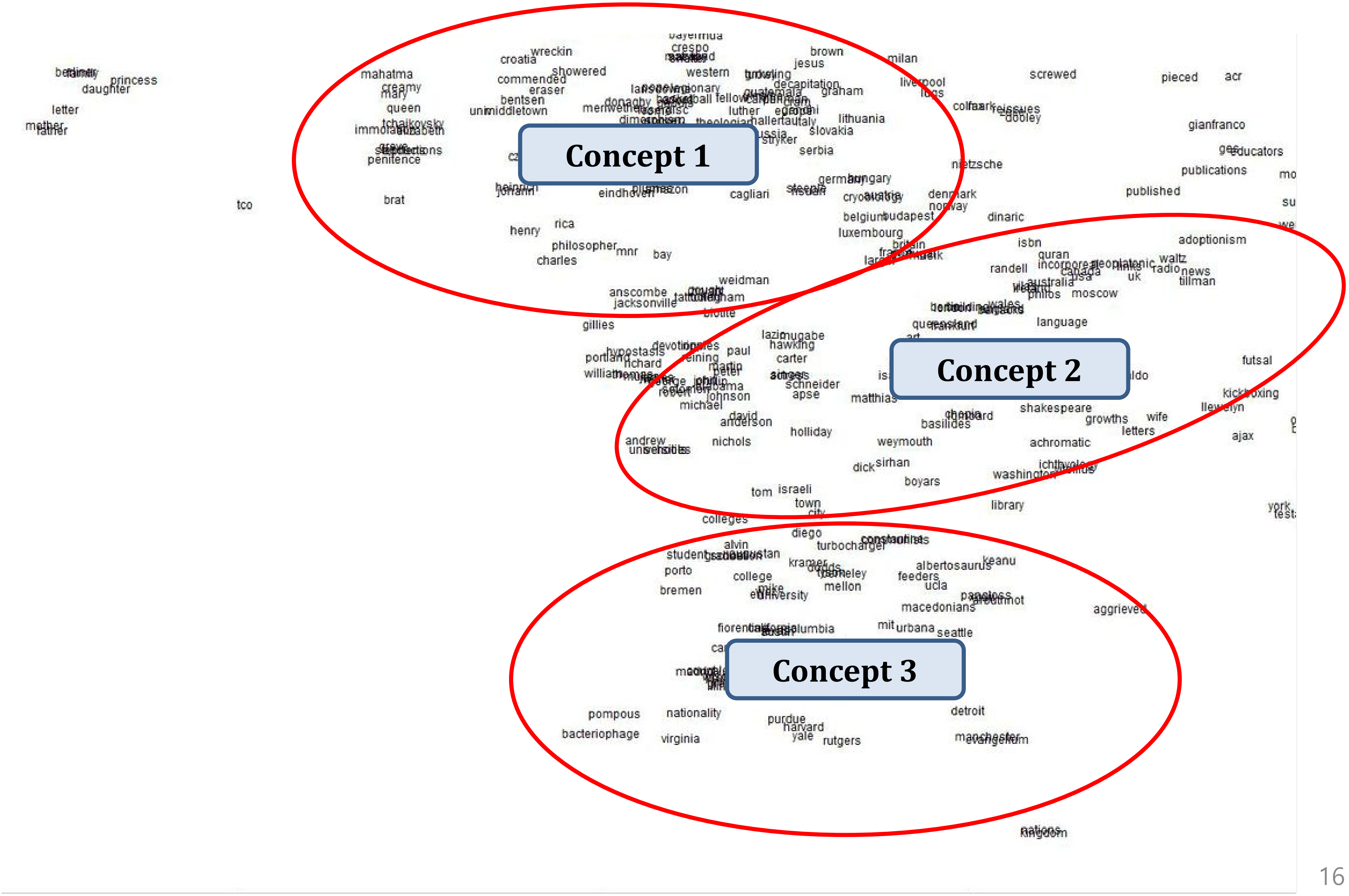
Proposed Method

1. Train Word2Vec from the collection of documents



Proposed Method

2. Cluster word2vec generated vectors to create clusters of concepts



Proposed Method

3. **Represent the documents by counting the number of times that their words belong to these different concept clusters (Similar to BOW approach!)**

Concept Cluster 1 = {Arsenal, Arsenal's, Aston Villa, Swansea City, Gunners…}
Concept Cluster 2 = {Squad, Players…}

[Document 1]:
Arsenal's annual injury **problem** is underway. **Their** thin squad will be put to the test by a Swansea City team looking to build on a vital win at Aston Villa.

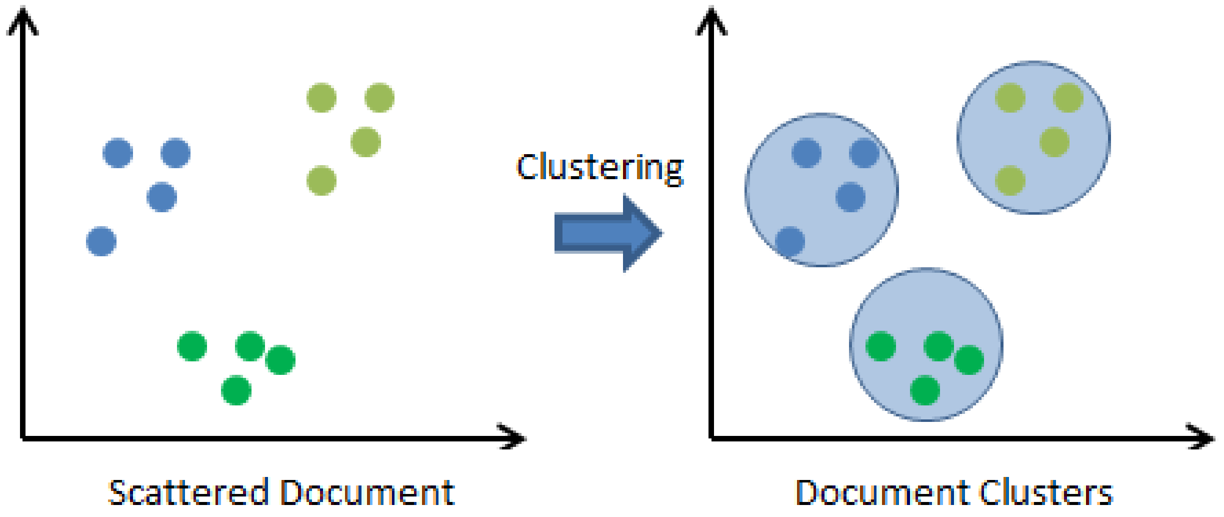
[Document 2]:
Arsenal have a whole host of injury **problems** to contend with. **The Gunners** currently sit top of the Premier League's infamous injury table. Eight senior players will be unable to take part at the Liberty Stadium

Features	Concept Cluster 1	Concept Cluster 2	…
Document 1	3	1	…
Document 2	2	1	…

Proposed Framework

4. **Evaluate the effectiveness of the document representation through document clustering task using the document vectors created from the previous step**

Features	Concept Cluster 1	Concept Cluster 2	...
Document 1	3	1	...
Document 2	2	1	...
...	....	....	...
Document n	0	8	...



Dataset: <Reuters>

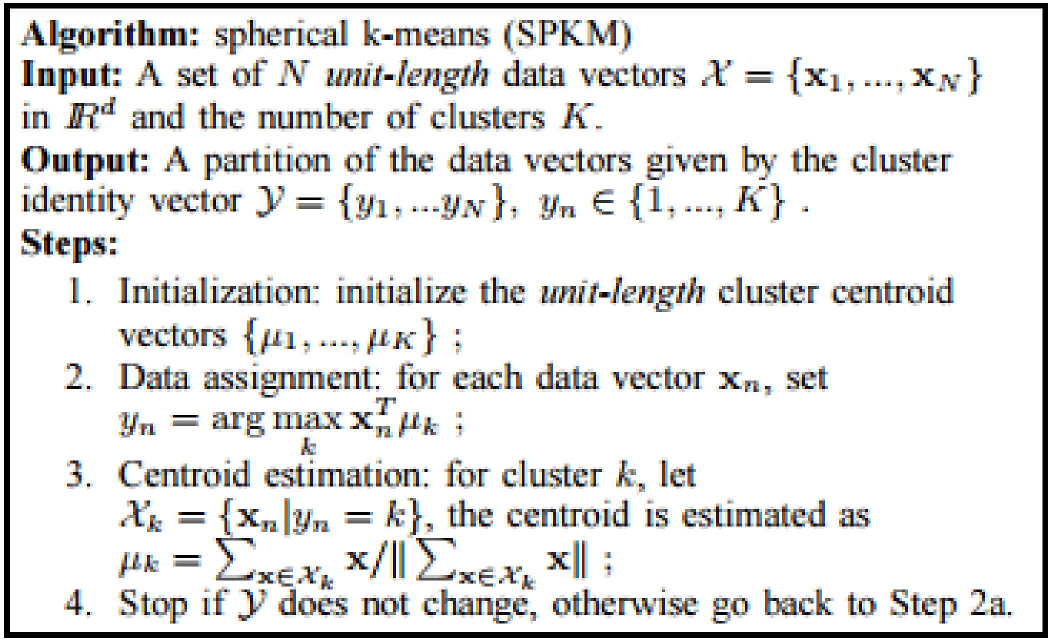
Total Number of Documents: 203,923 (2006. 09. 01 ~ 2015. 06. 06)

- Divided into 8 different categories
- Total number of sentences: 3,076,016
- Total number of tokens: 89,146,031
- Total number of unique tokens: 65,159
- Dimension of Doc2Vec and Word2Vec set as 500

Categories	Number of Documents
Entertainment	25,500
Sports	25,500
Technology	25,500
Market	25,423
Politics	25,500
Business	25,500
World	25,500
Health	25,500

Spherical K Means

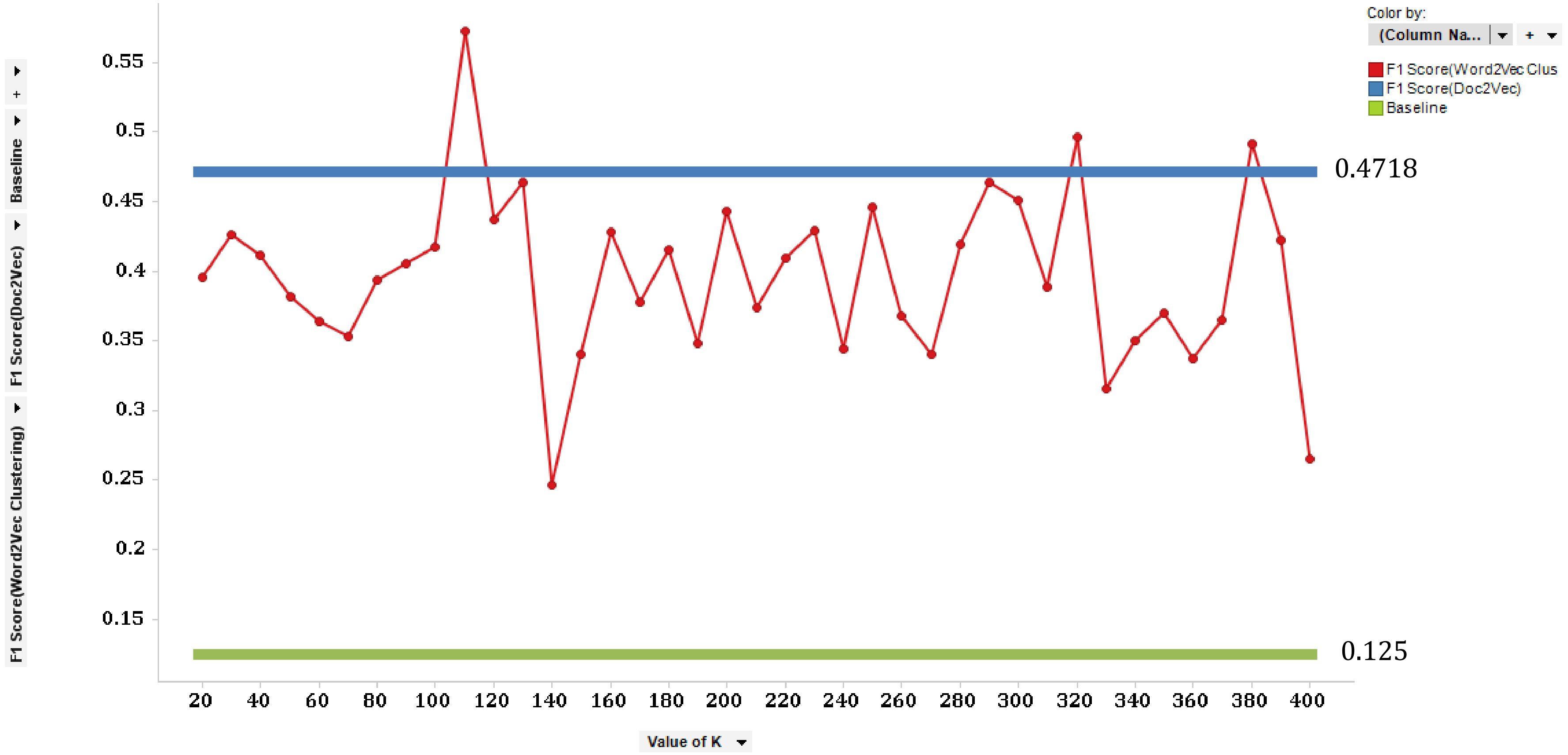
- To deal with these issues of high dimensional clustering, spherical k means algorithm has been used
- Essentially same as k means algorithm but with cosine similarity as a measure of proximity instead of Euclidean distance
- Chosen best result in terms of inertia



```
Algorithm: spherical k-means (SPKM)
Input: A set of $N$ unit-length data vectors $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$
in $I\!R^d$ 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.

Result



- Depending on the value of K used for clustering word vectors, the performance of document vectors vary greatly

- If appropriate value of K is chosen (in this case, K = 110), proposed method can outperform Doc2Vec in document clustering, indicating better document representation

Word Clusters

- Pharmaceutical Terminologies

two-drug	3
Genzymes	3
Suppressant	3
Vectibix	3
Adderall	3
Intravenously	3
Crestor	3
non-prescription	3
Antivirals	3
Sandoz	3
Discontinuing	3
anti-retroviral	3
anti-viral	3

- Military Terminologies

dni	47
Expeditionary	47
Panettas	47
maj.	47
Marshals	47
Operationally	47
Chiefs	47
pentagon.	47
Commandant	47
Stationing	47
Troop	47
Defences	47
Lieutenant	47

Word Clusters

- Garment Related Terminologies

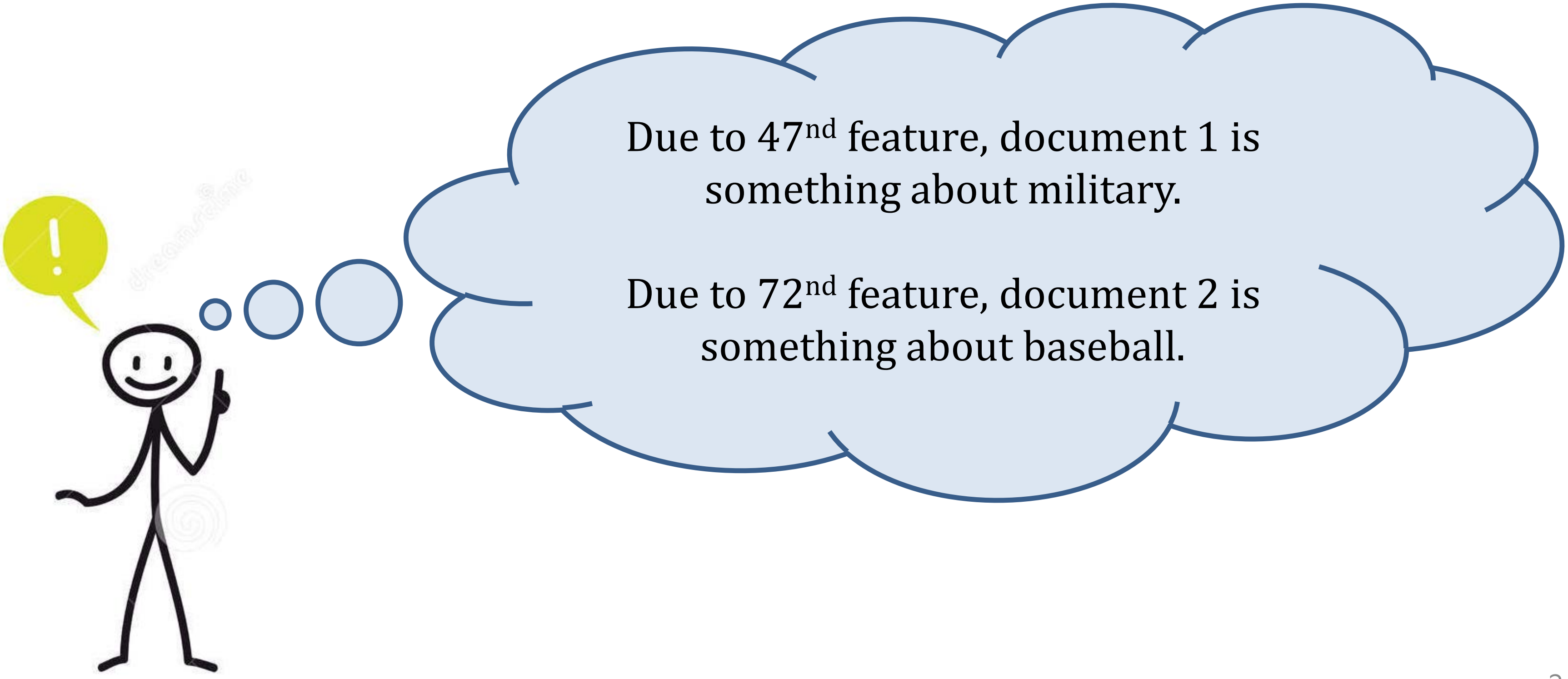
handbags	60
lagerfeld	60
lipstick	60
disheveled	60
flowing	60
bra	60
figurines	60
embellished	60
hoodie	60
garlands	60
masculine	60
tuxedos	60
garish	60

- Baseball Terminologies

dickey	72
beltre	72
peavy	72
runs	72
ervin	72
pinch-hitter	72
cincinnatis	72
batting	72
jacoby	72
cubs	72
sabathia	72
prado	72
hits	72

Word Clusters

Features	X[0]	…	X[47]	…	X[72]	…
Document 1	2	…	**23**	…	**0**	…
Document 2	2	…	**0**	…	**33**	…



Due to 47nd feature, document 1 is something about military.

Due to 72nd feature, document 2 is something about baseball.

Conclusion

- The proposed method incorporates both the effective performance of the distributed representation and intuitive explanatory power of BOW representation

- For certain value of K used for clustering word2vec vectors, the proposed method, at least in terms of document clustering, can represent the documents at a similar performance level as Doc2Vec

- As the performance of Doc2Vec can be sensitive to the number of dimensions (the number of hidden nodes), the proposed method needs to be compared with various Doc2Vec models resulting from different number of dimensions

- Performance of the proposed method should also be checked for document classification task in order to substantiate the effective document representation of the proposed method

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.

Dai, Andrew M., et al. "Document Embedding with Paragraph Vectors." NIPS Deep Learning Workshop. 2014.

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.

R. Liu, D. Wang, and C. Xing, "Document classification based on word vectors." ISCSLP, 2014

Rong, Xin. "word2vec Parameter Learning Explained." arXiv preprint arXiv:1411.2738 (2014).
5. IEEE, 2005.

Reference

Turney, Peter D., and Patrick Pantel. "From frequency to meaning: Vector space models of semantics." Journal of artificial intelligence research 37.1 (2010): 141-188.

Xing, Chao, et al. "Document classification with distributions of word vectors." Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA). IEEE, 2014.

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