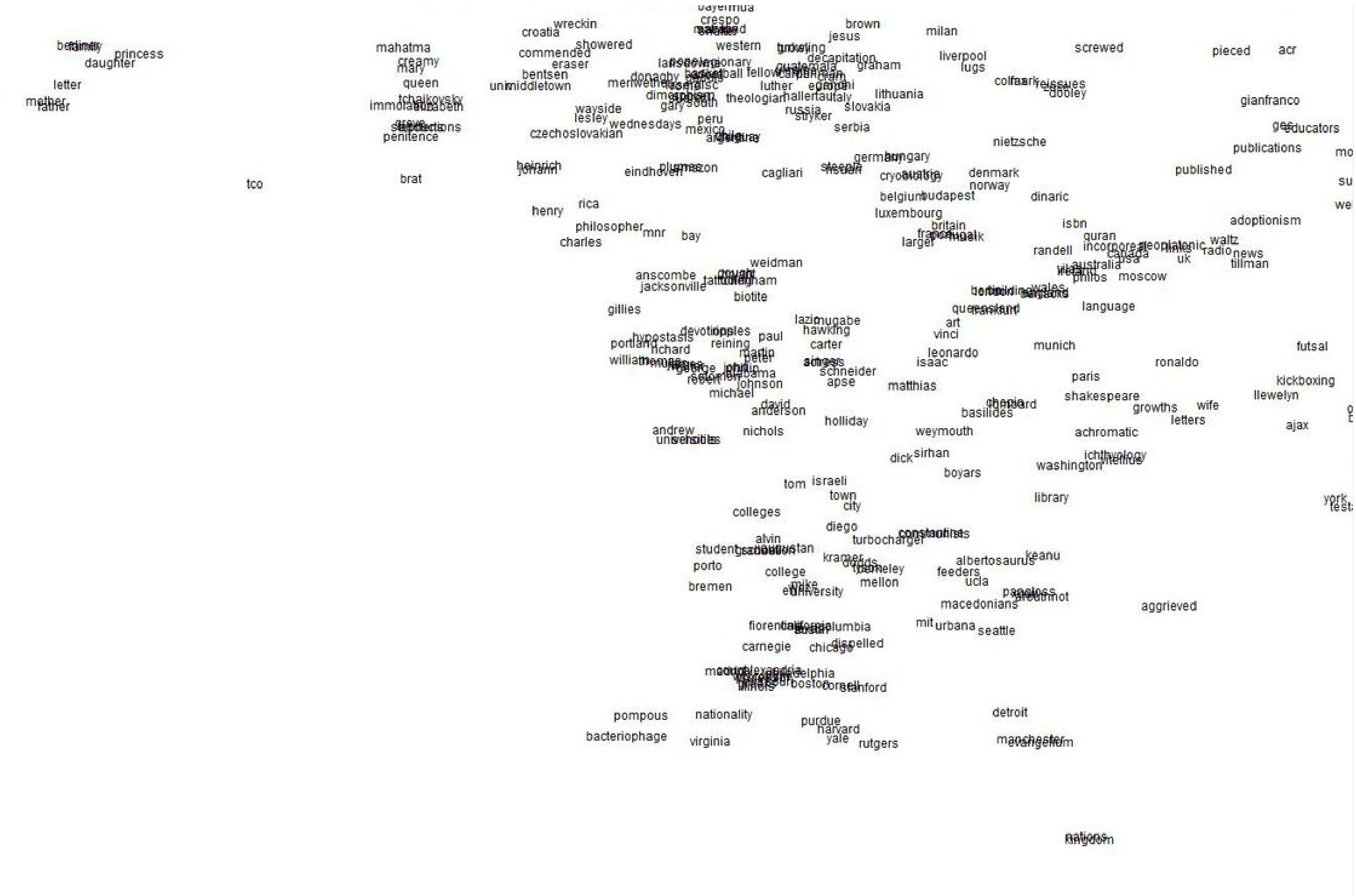
Interpretable Distributed Representation of Documents through Explicitly Explanatory Features:

<Experiment Result>

November 25th, 2015
SNU Data Mining Center
Han Kyul Kim

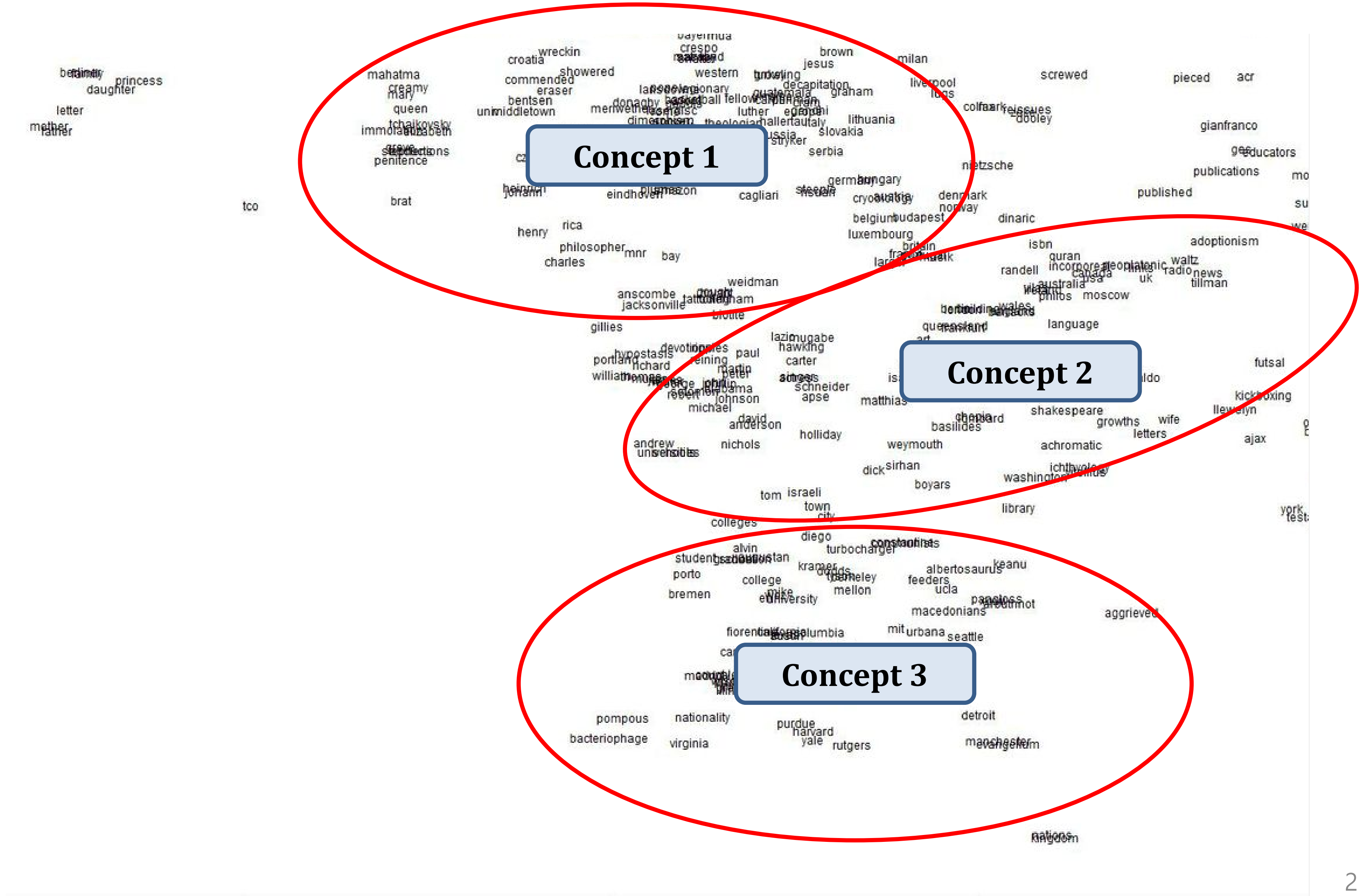
Proposed Method

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…}

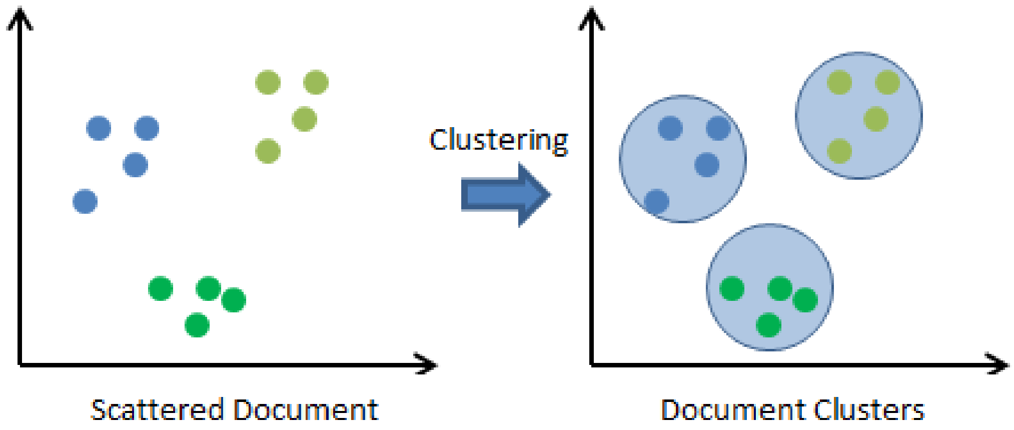
[Document 1]:
Arsenal's annual injury problem is underway. Their thin squad will be put to the test by a Swansea City team looking to build on a vital win at Aston Villa.

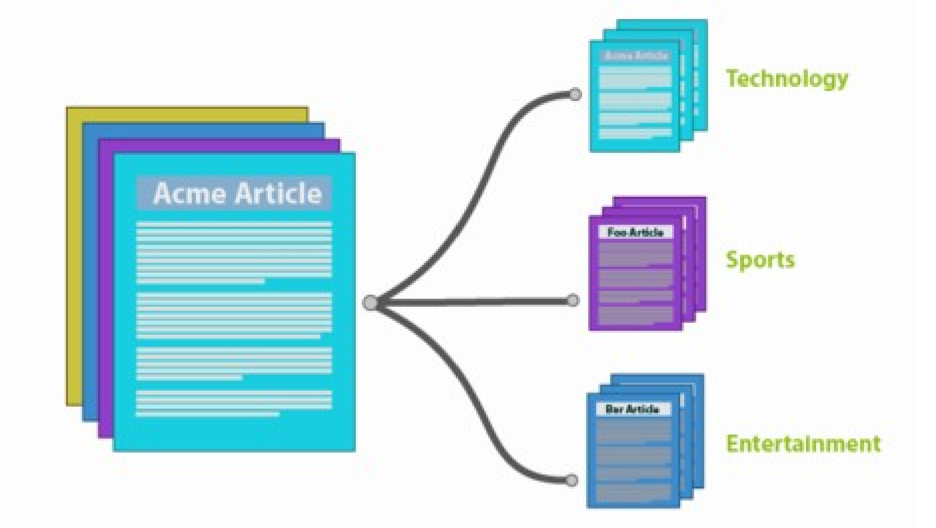
[Document 2]:
Arsenal have a whole host of injury problems to contend with. The Gunners currently sit top of the Premier League's infamous injury table. Eight senior players will be unable to take part at the Liberty Stadium

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

Proposed Framework

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





Word2Vec Averaging

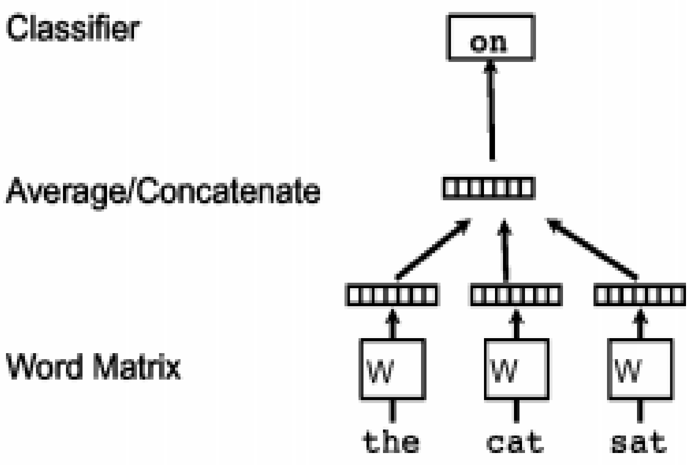
Xing, Chao, et al. "Document classification with distributions of word vectors." Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA). IEEE, 2014.

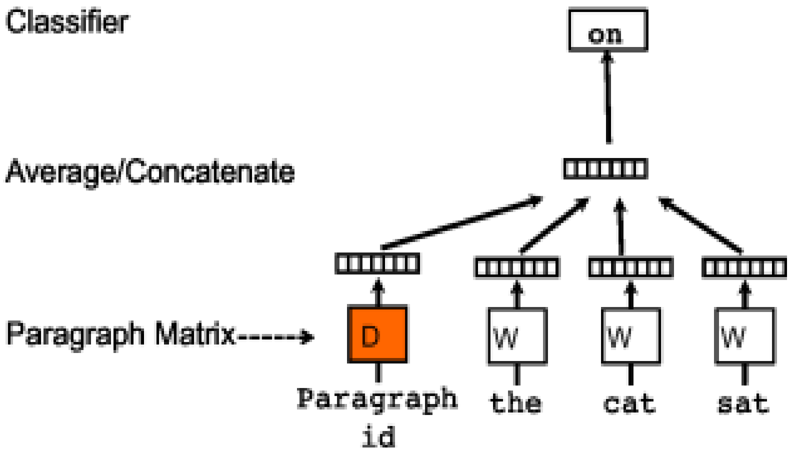
- **Simple average pooling approach:**

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

  - 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>

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

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

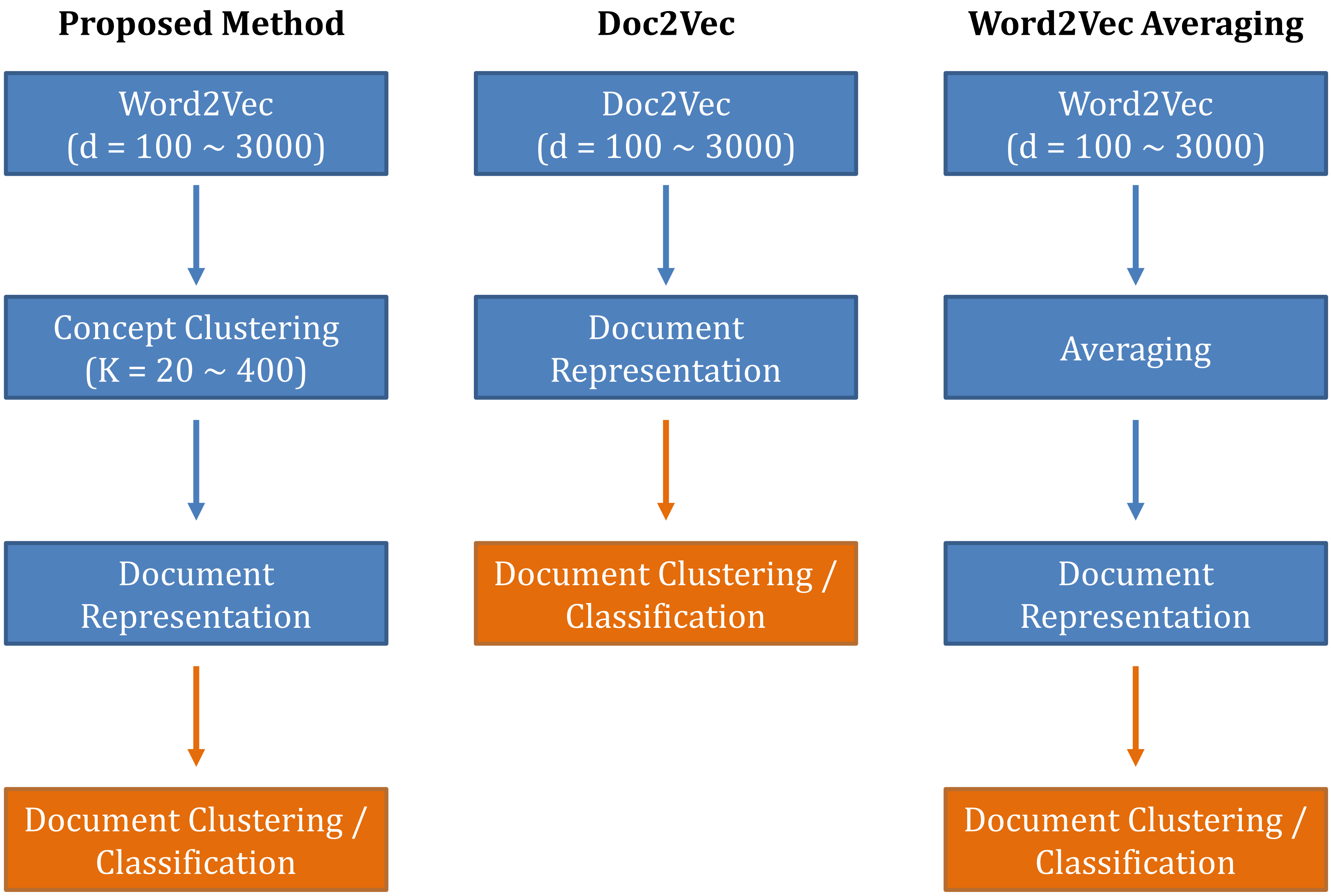
Dataset: <Reuters>

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

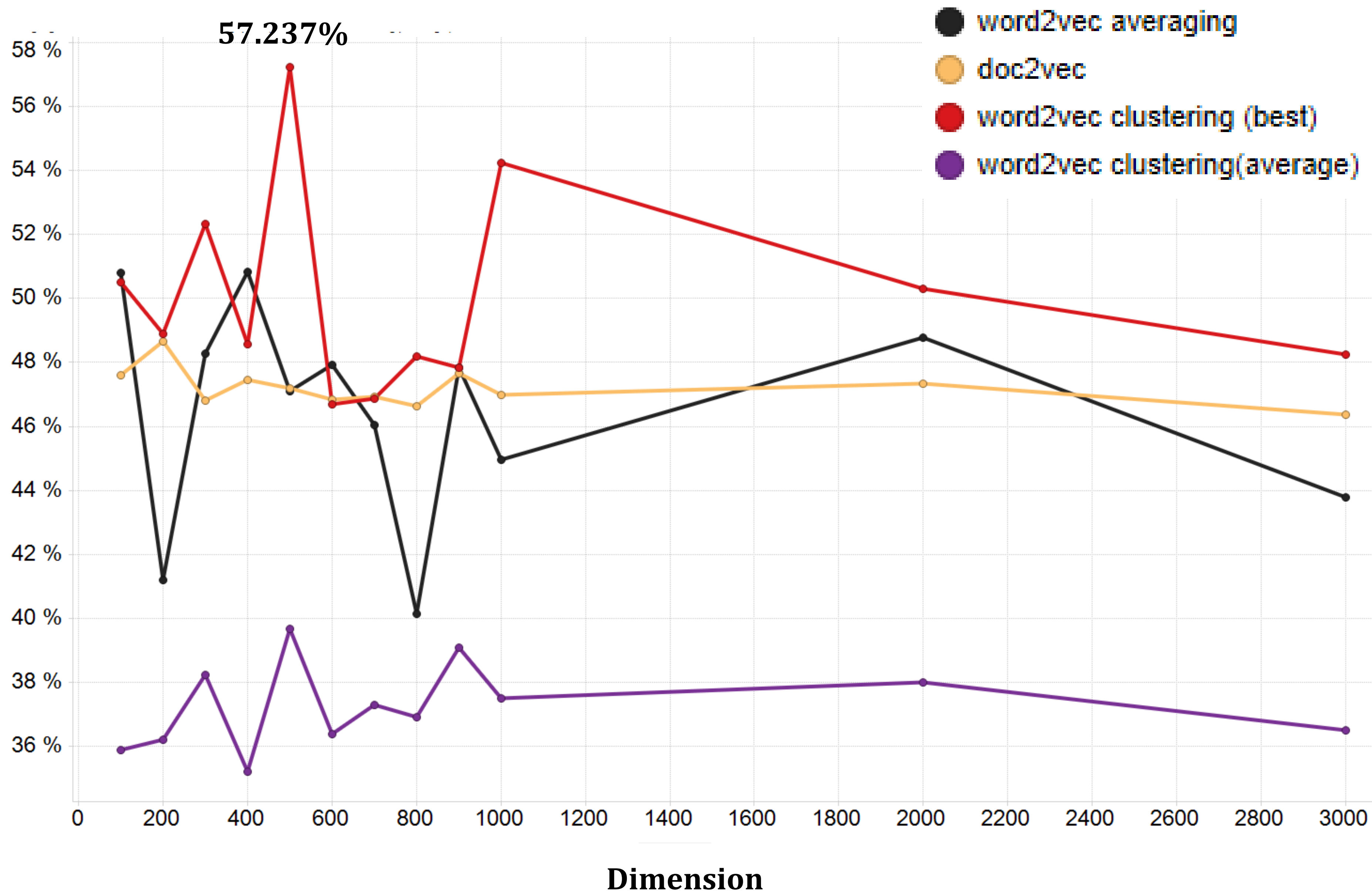
- Divided into 8 different categories
- Total number of sentences: 3,076,016
- Total number of tokens: 89,146,031
- Total number of unique tokens: 65,159

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

Experiment Setting



F1 Score – Document Clustering



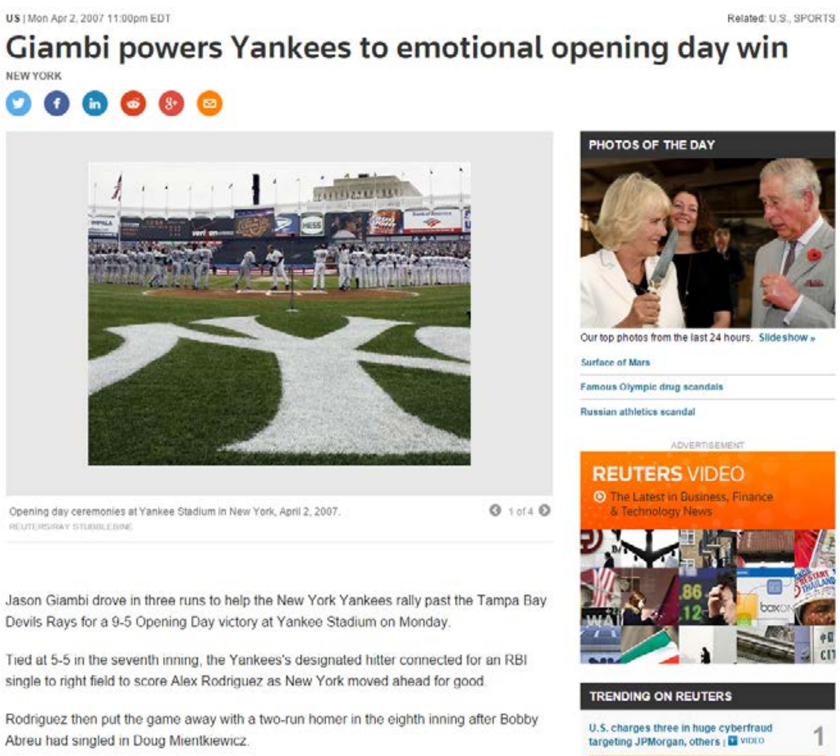
F1 Score – Document Clustering

Dimension	Doc2vec	Word2vec Averaging	Word2vec Clustering (best)	Word2vec Clustering (average)
100	0.4759801	0.5080175	0.505036383	0.358902869
200	0.4866283	0.4119441	0.489006684	0.362013807
300	0.4680473	0.4826981	0.523417232	0.382304123
400	0.4746061	0.5083417	0.485580917	0.352056779
500	0.4718293	0.4711248	0.572365785	0.396641247
600	0.468465	0.4791967	0.466831366	0.363745187
700	0.4692426	0.4603129	0.468562776	0.372937356
800	0.4664247	0.4015193	0.481906004	0.369164546
900	0.4765871	0.4784783	0.478379763	0.390731813
1000	0.4698723	0.4494728	0.542261416	0.374829576
2000	0.4733753	0.4877426	0.502956721	0.379956078
3000	0.4636311	0.4377777	0.482590878	0.364973752

Concept Cluster Documents

Features	X[0]	...	X[33]	...	X[108]	X[109]
Document 1	5	...	**1**	...	**0**	0
Document 2	27	...	**36**	...	**1**	0

[Document 1]



US | Mon Apr 2, 2007 11:00pm EDT — Related: U.S., SPORTS

Giambi powers Yankees to emotional opening day win

NEW YORK

PHOTOS OF THE DAY

Our top photos from the last 24 hours. Slideshow »

Surface of Mars

Famous Olympic drug scandals

Russian athletics scandal

ADVERTISEMENT

REUTERS VIDEO

The Latest in Business, Finance & Technology News

Opening day ceremonies at Yankee Stadium in New York, April 2, 2007.

1 of 4

TRENDING ON REUTERS

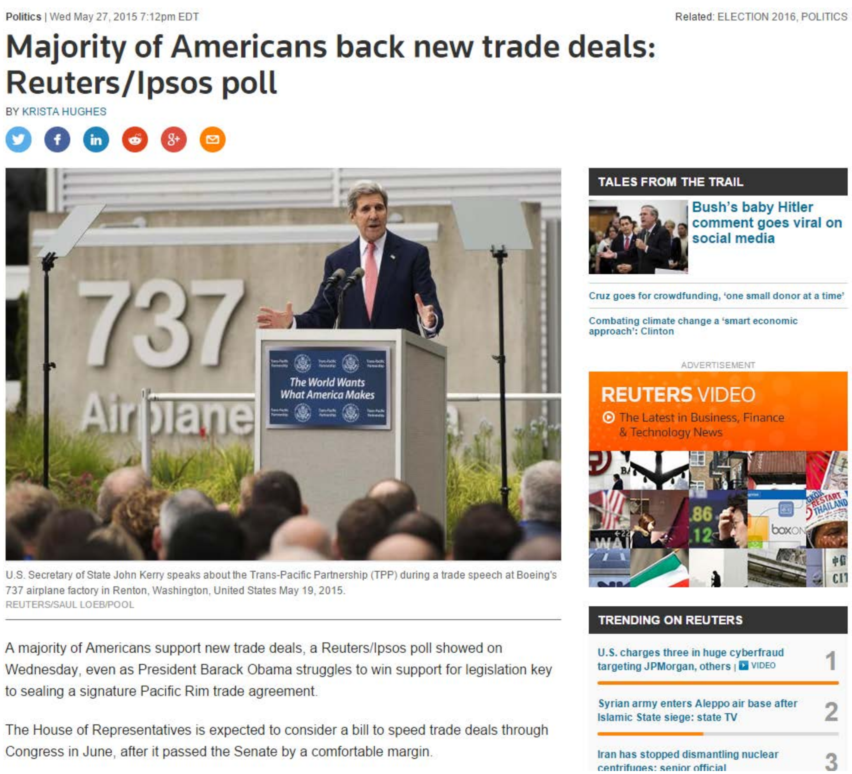
U.S. charges three in huge cyberfraud targeting JPMorgan, others | VIDEO 1

Jason Giambi drove in three runs to help the New York Yankees rally past the Tampa Bay Devils Rays for a 9-5 Opening Day victory at Yankee Stadium on Monday.

Tied at 5-5 in the seventh inning, the Yankees's designated hitter connected for an RBI single to right field to score Alex Rodriguez as New York moved ahead for good.

Rodriguez then put the game away with a two-run homer in the eighth inning after Bobby Abreu had singled in Doug Mientkiewicz.

[Document 2]



Politics | Wed May 27, 2015 7:12pm EDT — Related: ELECTION 2016, POLITICS

Majority of Americans back new trade deals: Reuters/Ipsos poll

BY KRISTA HUGHES

TALES FROM THE TRAIL

Bush's baby Hitler comment goes viral on social media

Cruz goes for crowdfunding, 'one small donor at a time'

Combating climate change a 'smart economic approach': Clinton

ADVERTISEMENT

REUTERS VIDEO

The Latest in Business, Finance & Technology News

U.S. Secretary of State John Kerry speaks about the Trans-Pacific Partnership (TPP) during a trade speech at Boeing's 737 airplane factory in Renton, Washington, United States May 19, 2015.

REUTERS/SAUL LOEB/POOL

TRENDING ON REUTERS

U.S. charges three in huge cyberfraud targeting JPMorgan, others | VIDEO 1

Syrian army enters Aleppo air base after Islamic State siege: state TV 2

Iran has stopped dismantling nuclear centrifuges: senior official 3

A majority of Americans support new trade deals, a Reuters/Ipsos poll showed on Wednesday, even as President Barack Obama struggles to win support for legislation key to sealing a signature Pacific Rim trade agreement.

The House of Representatives is expected to consider a bill to speed trade deals through Congress in June, after it passed the Senate by a comfortable margin.

Contrasting Features

Word	Distance to Centroid
Fretilin	0.298141
hard-left	0.299046
Smer	0.300370
Ovp	0.300925
Greens	0.303287
Socialists	0.305534
Party	0.310117
Peronist	0.321366
Kke	0.324051
Pis	0.333701
Congress-led	0.336214
Centrists	0.340830
Pro-eu	0.343883

- **<u>Political Party</u>**
- Concept Frequency: **<u>Doc 1: 5 vs. Doc 2: 27</u>**

Contrasting Features

Word	Distance to Centroid
Six-nation	0.341851
Negotiations	0.358357
Final-status	0.358551
Talks	0.369950
Accord	0.384951
Two-track	0.388305
Agreement	0.388699
Working-level	0.401054
Long-stalled	0.411923
Trilateral	0.416301
Deal	0.417467
Disarmament	0.423539
Israeli-Syrian	0.424372

- **Negotiation & Treaty**

- Concept Frequency: **Doc 1: 1 vs. Doc 2: 36**

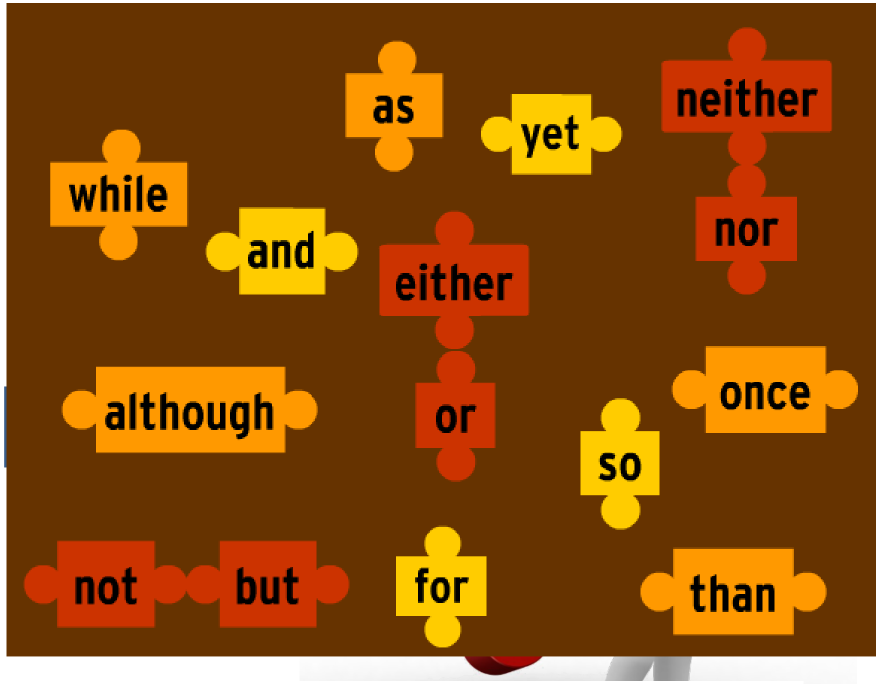
Contrasting Features

Word	Distance to Centroid
Astros	0.209113
playoff-bound	0.216279
Phillies	0.231677
last-place	0.232075
Timberwolves	0.237807
Mariners	0.242180
Flyers	0.245423
Thrashers	0.247595
Sabres	0.250336
Devils	0.252015
Blackhawks	0.255871
Orioles	0.256698
Athletics	0.260109

- **Sport Teams**
- Concept Frequency: **Doc 1: 14 vs. Doc 2: 0**

Contrasting Features

Word	Distance to Centroid
While	0.378267
But	0.384359
However	0.387299
Although	0.388328
Only	0.417179
Now	0.421535
Then	0.424409
Also	0.425922
Another	0.439093
The	0.449224
May	0.449749
Leaving	0.451124
That	0.451503



- **Conjunctions**

- Concept Frequency: **Doc 1: 146 vs. Doc 2: 198**

Contrasting Features

Word	Distance to Centroid
Contravenes	0.368736
Non-discrimination	0.389781
Contravened	0.395901
Unenforceable	0.396915
Prohibit	0.400195
Obliging	0.402253
Supersede	0.405173
Enshrine	0.408061
Codify	0.411534
Prohibiting	0.411764
Contravene	0.415466
Specifies	0.417802
Reclassifying	0.418312

- **To oppose/revise (legal context)**

- Concept Frequency:
  Doc 1: 0 vs. Doc 2: 18

Contrasting Features

Word	Distance to Centroid
Fourth-inning	0.188195
Aybar	0.201127
Pinch-hit	0.217082
Pinch-hitter	0.221174
Hitless	0.227714
First-inning	0.236647
DH	0.240897
Two-out	0.241593
Okajima	0.249996
No-hit	0.250199
Delmon	0.253375
Kozma	0.255309
Eighth-inning	0.255412

- **<u>Baseball Terminologies</u>**

  - Concept Frequency: **<u>Doc 1: 68 vs. Doc 2: 1</u>**

Contrasting Features

Word	Distance to Centroid
Sirnak	0.190246
Barzeh	0.216446
Qaboun	0.218347
Sidon	0.218943
Mukalla	0.226163
Mosul	0.231129
Hama	0.232689
Adhamiya'	0.233669
Ramadi	0.235161
Jobar	0.241562
Vabroud	0.242106
Kerbala	0.242618
Gunbattles	0.243645

- **<u>Middle Eastern Cities</u>**

- Concept Frequency:
  <u>Doc 1: 37 vs. Doc 2: 11</u>

Document Classification

- Create triplets of documents (2 from same class and 1 from different class)
- Using cosine similarity as distance metric, we want to classify a document within a triplet that is from a different class (most distant from the other two documents)
- Tested on 280,000 triplets

Example:

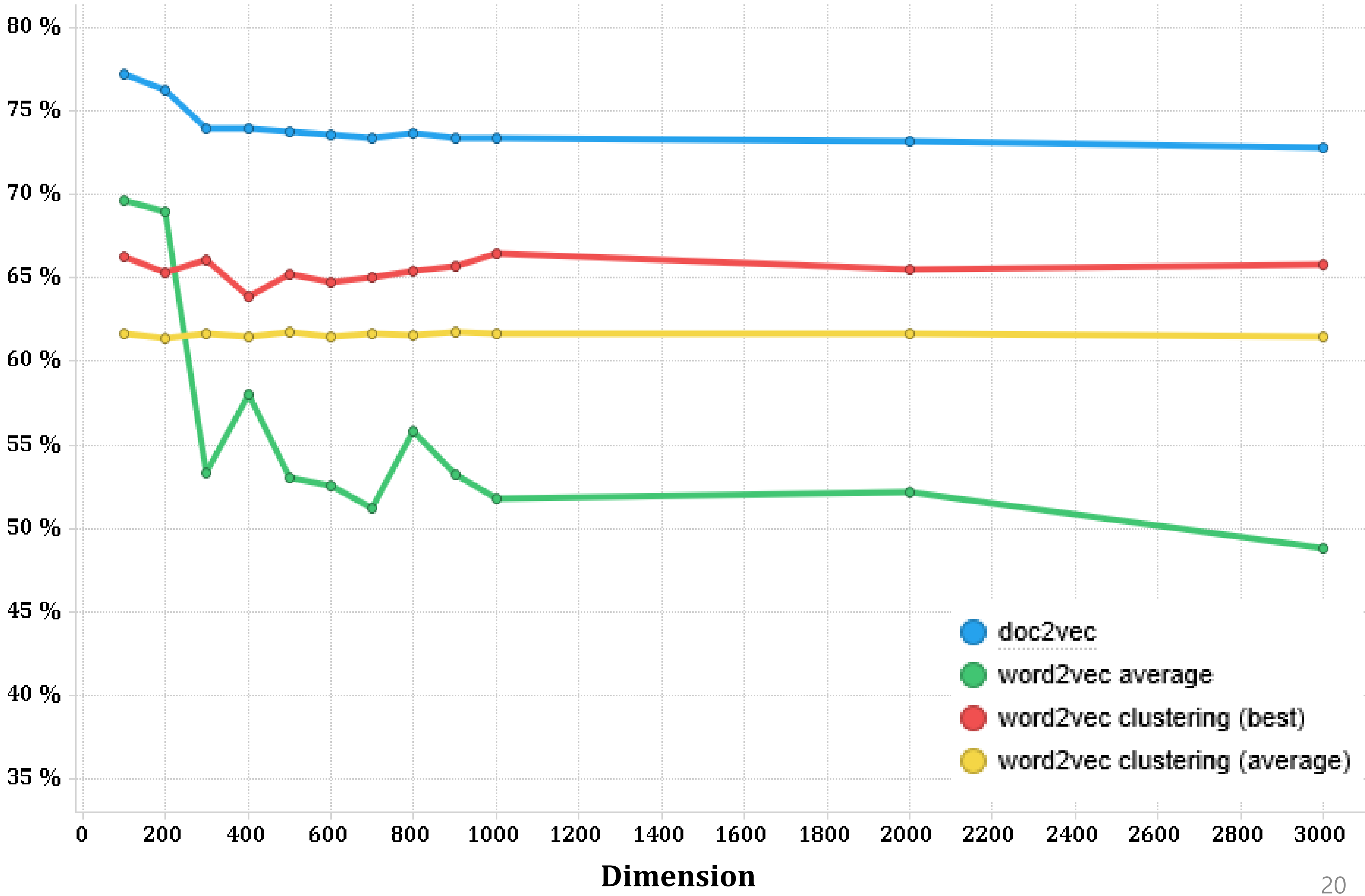
(Document ID, Class Label)

[(9699, 'businessNews'), (3817, 'businessNews'), (38841, 'entertainmentNews')]

→ (38841, 'entertainmentNews') **(Correct)**

→ (3817, 'businessNews') **(Incorrect)**

F1 Score -Document Classification



F1 Score -Document Classification

Dimension	Doc2vec	Word2vec Averaging	Word2vec Clustering (best)	Word2vec Clustering (average)
100	0.771861	0.696642857	0.662660714	0.616532601
200	0.762332	0.68935	0.6529	0.614233608
300	0.739907	0.533425	0.660646429	0.616769414
400	0.739675	0.579921429	0.639146429	0.614428297
500	0.737425	0.529971429	0.652560714	0.617219597
600	0.735929	0.526042857	0.647025	0.614812363
700	0.733518	0.512546429	0.650364286	0.616948168
800	0.736554	0.558592857	0.654253571	0.615760897
900	0.734243	0.531871429	0.657182143	0.617507509
1000	0.733321	0.517432143	0.664453571	0.616804762
2000	0.731429	0.521503571	0.655253571	0.616660714
3000	0.727796	0.488457143	0.657810714	0.61490348

Conclusion

- Word2Vec Clustering method provides interpretable power to distributed representation of documents

- It is a hybrid method that incorporates the advantages of BOW and Doc2Vec Approach

- It can…
  - provide explanations on what each component of document vector indicates

  - further provide concrete explanation behind the results generated from additional text mining techniques based on word2vec clustering method

  - Test whether attempted hyperparameter of text mining model is appropriate or not

Paper Outline

1. **Introduction**
    - Growing importance of text mining
    - Need for interpretability for applicability (representation itself is not the end)
2. **Background**
    - BOW
    - Extension of BOW (LSA)
    - Word2Vec & Doc2Vec
3. **Proposed Method**
4. **Data Set & Task Description**
5. **Result**
    - Quantitative – F1 Score
    - Qualitative – Actual examples
6. **Conclusion**

Reference

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

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

3. 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.

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

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

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

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

Reference

8. 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.

9. 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.

10. Zhong, Shi. "Efficient online spherical k-means clustering." Neural Networks, 2005. IJCNN05. Proceedings. 2005 IEEE International Joint Conference on. Vol. 5. IEEE, 2005.