Interpretable Distributed Representation of Documents with Explicitly Explanatory Features:

<Decision Tree>

November 30th, 2015
SNU Data Mining Center
Han Kyul Kim

Paper Outline

1. **Introduction**
   - Distributed Representation of words and documents has established itself as a new standard in text mining and NLP communities.
   - Word2vec clustering method provides…
     1. Vector Interpretability: Can intuitively understand the features and the components of calculated document vectors
     2. Model Explainability: Can comprehend the operating logic behind a trained classifier, through which one can easily compare the calculated result with his domain expertise or enhance his understanding of the given phenomenon
2. **Background**
   - BOW (pros & cons)
   - Word2Vec & Doc2Vec (pros & cons)
3. **Proposed Method**
4. **Data Set**
5. **Result**
   - Vector Interpretability: Document Clustering Task (Clustering)
   - Model Explainability: Document Classification (Decision Tree)
6. **Conclusion**

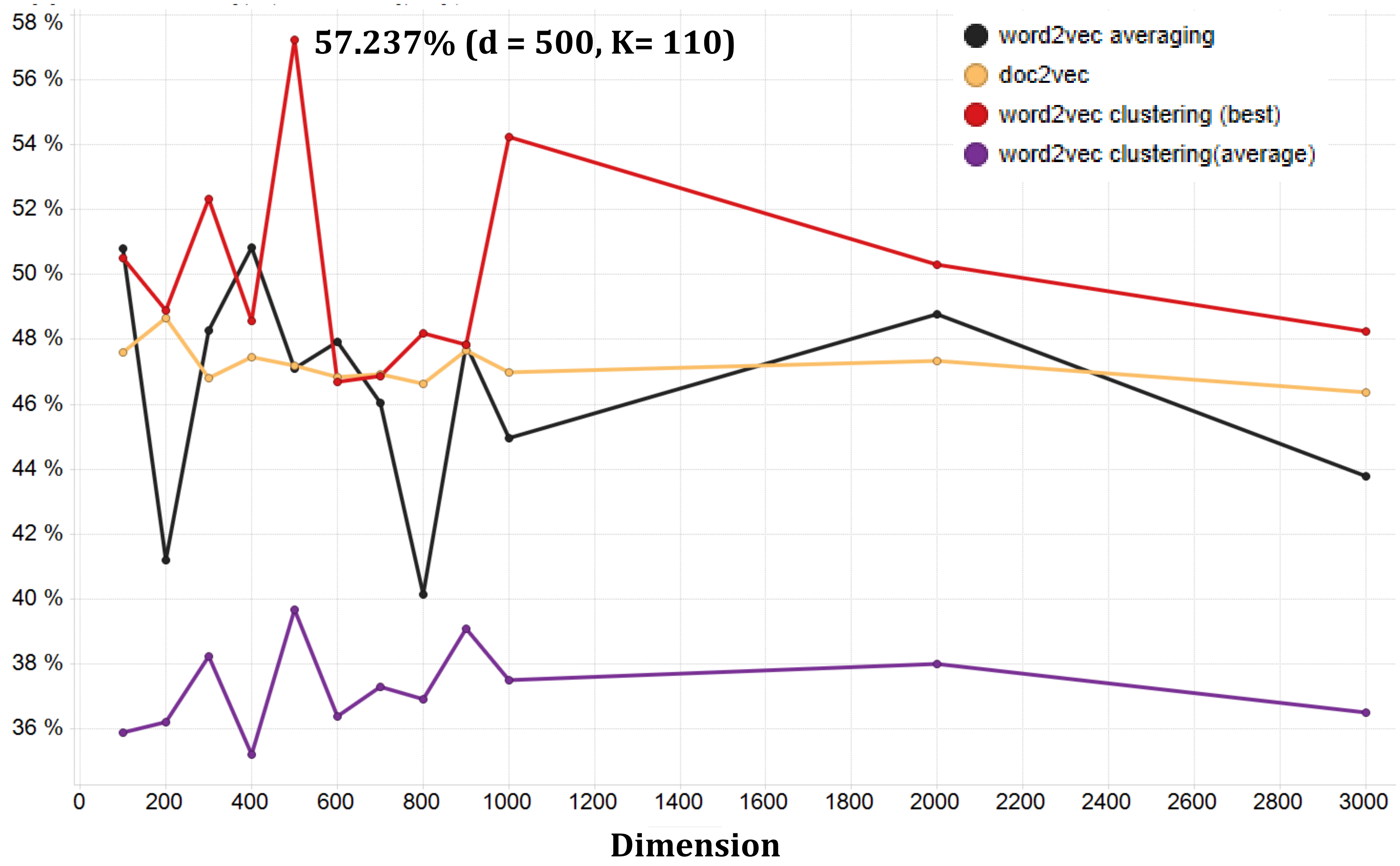
Dataset: <Reuters>

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	Total Number of Documents	Training Set	Test Set
Entertainment	25,500	20,500	5,000
Sports	25,500	20,500	5,000
Technology	25,500	20,500	5,000
Market	25,423	20,423	5,000
Politics	25,500	20,500	5,000
Business	25,500	20,500	5,000
World	25,500	20,500	5,000
Health	25,500	20,500	5,000

Experiment Setting

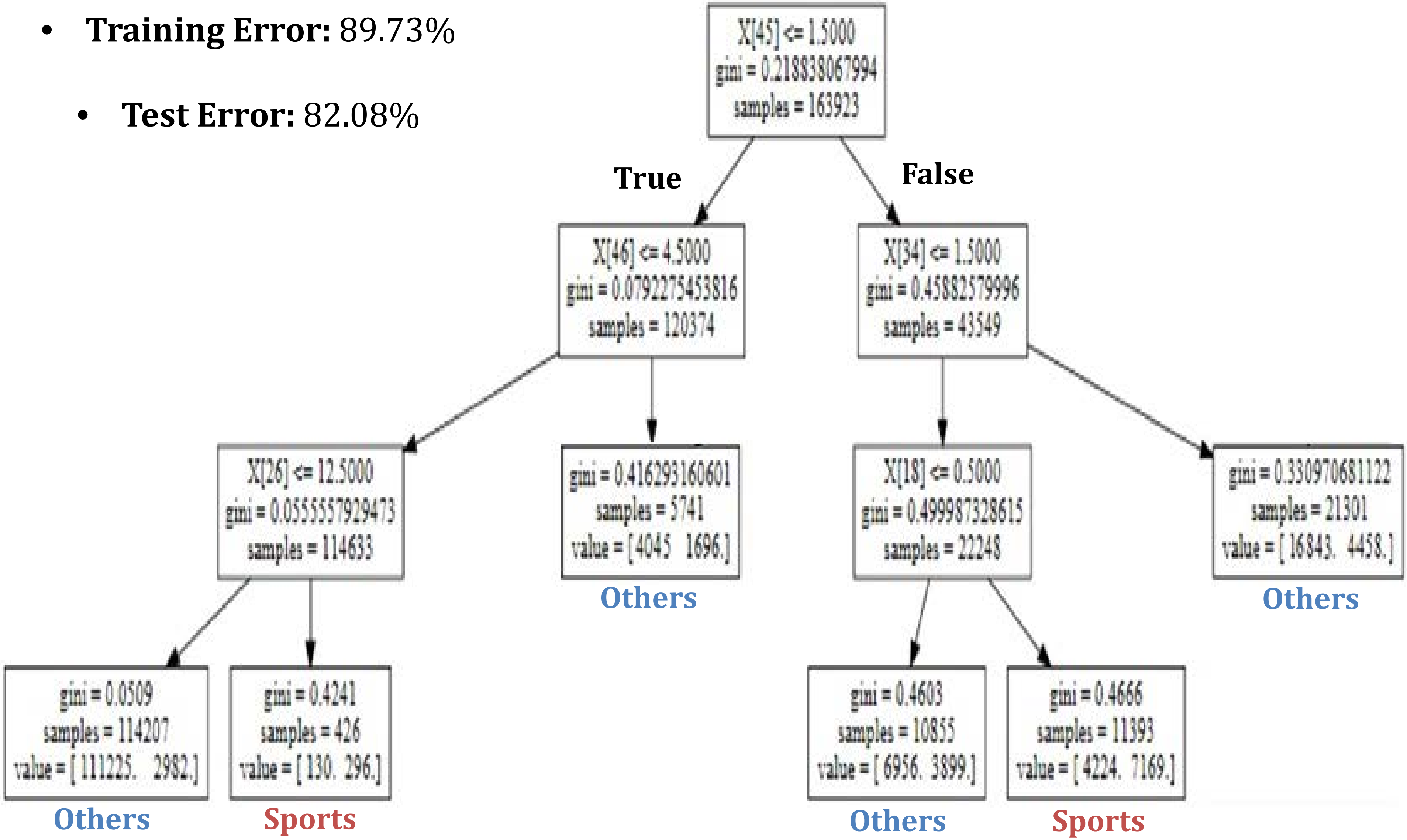


57.237% (d = 500, K= 110)

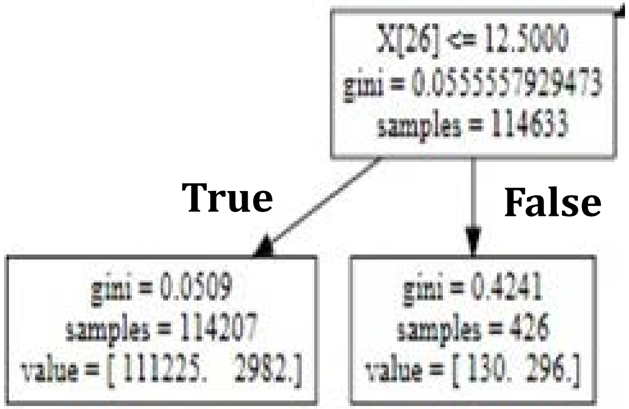
- For simplicity and visualization convenience, decision tree that classifies sports articles from other remaining classes of articles has been constructed (Sports vs. Rest)

Decision Tree Result

- **Training Error:** 89.73%
- **Test Error:** 82.08%



Split Analysis

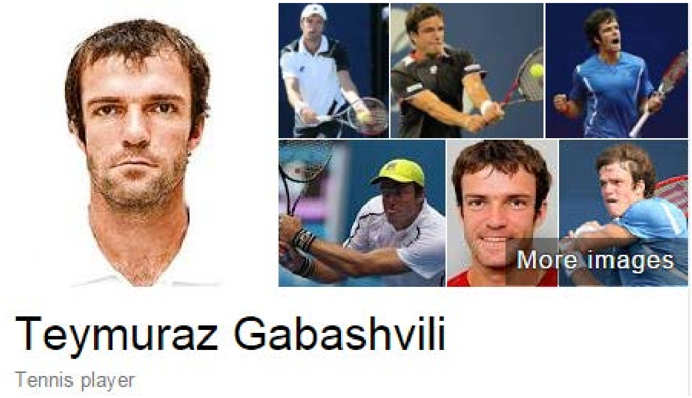


Others

Sports

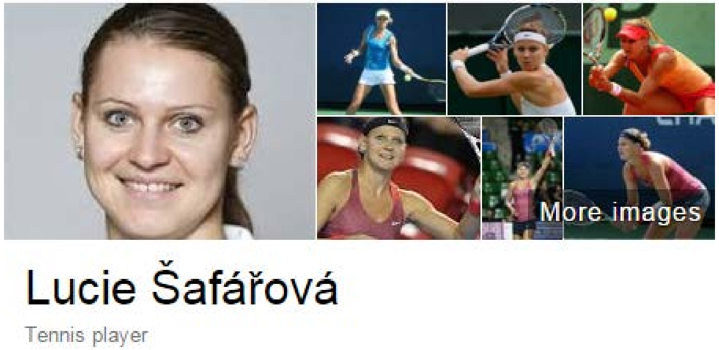
X[26] = Tennis

Word	Distance to Centroid
Gabashvili	0.18321
Kvitova	0.19082
Dushevina	0.19822
Ninth-seeded	0.21496
Bojana	0.21574
Kuznetsova	0.22725
Safarova	0.23322
Fifth-seeded	0.23704
Rodionova	0.23830
Barthel	0.23836



Teymuraz Gabashvili

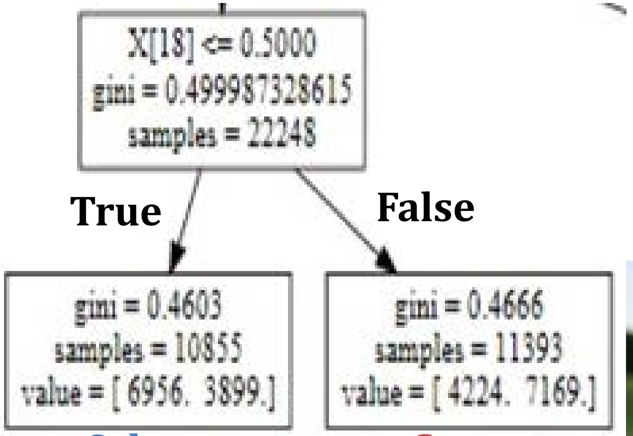
Tennis player



Lucie Šafářová

Tennis player

Split Analysis



X[18] <= 0.5000
gini = 0.499987328615
samples = 22248

True

gini = 0.4603
samples = 10855
value = [6956. 3899.]

False

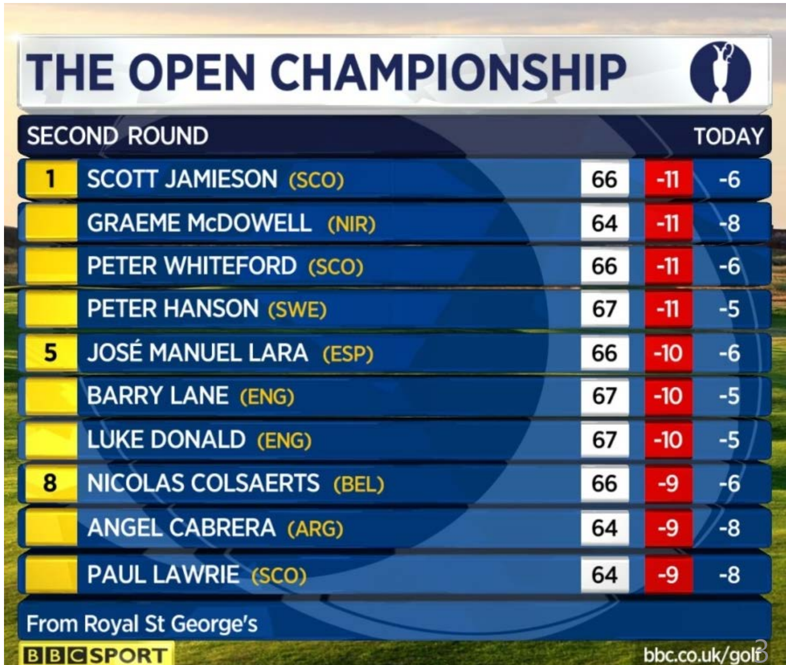
gini = 0.4666
samples = 11393
value = [4224. 7169.]

Others

Sports

X[18] = Golf Terms

Word	Distance to Centroid
Back-nine	0.23369
Double-bogeys	0.23978
Eagling	0.24029
Congressional	0.24441
Six-over	0.24894
Five-over	0.24914
Seven-over	0.25737
One-over	0.25855
Five-birdie	0.26099
Three-putting	0.26230

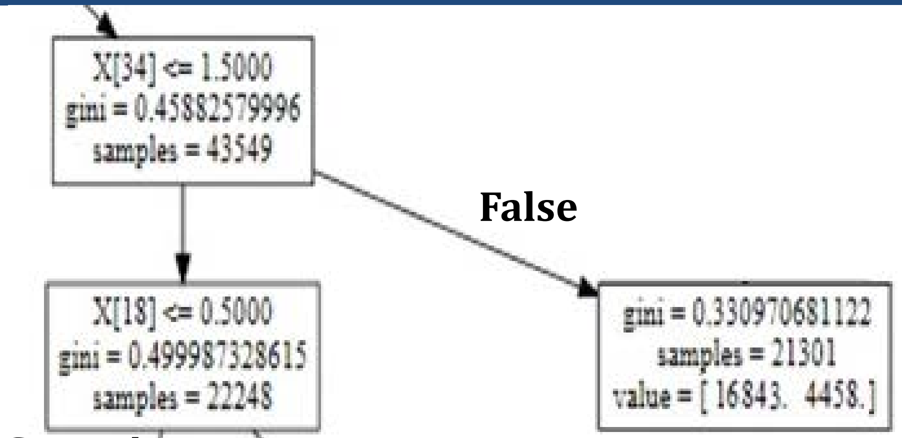


THE OPEN CHAMPIONSHIP

SECOND ROUND				TODAY
1	SCOTT JAMIESON (SCO)	66	-11	-6
	GRAEME McDOWELL (NIR)	64	-11	-8
	PETER WHITEFORD (SCO)	66	-11	-6
	PETER HANSON (SWE)	67	-11	-5
5	JOSÉ MANUEL LARA (ESP)	66	-10	-6
	BARRY LANE (ENG)	67	-10	-5
	LUKE DONALD (ENG)	67	-10	-5
8	NICOLAS COLSAERTS (BEL)	66	-9	-6
	ANGEL CABRERA (ARG)	64	-9	-8
	PAUL LAWRIE (SCO)	64	-9	-8

From Royal St George's

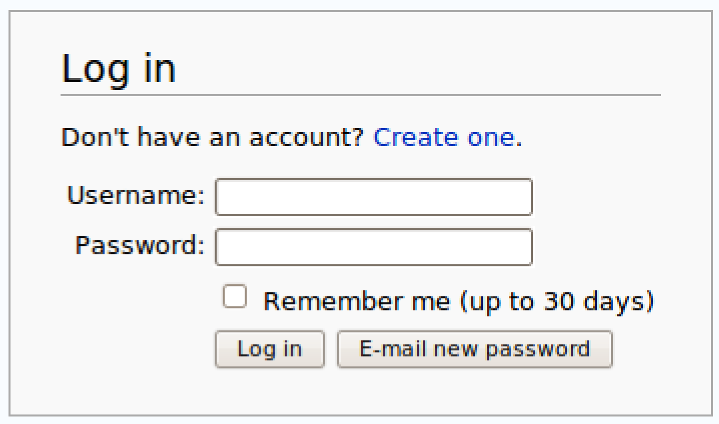
Split Analysis



X[34] = Computer Security

Others

Word	Distance to Centroid
Login	0.33140
Backdoors	0.34021
Usernames	0.34876
Troves	0.35415
Unencrypted	0.35721
Logins	0.36225
Password-protected	0.36632
Accessed	0.37360
Passwords	0.38089
Inboxes	0.39457



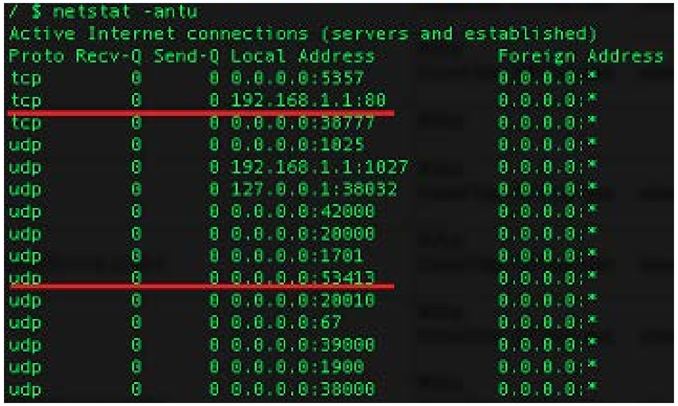
Log in

Don't have an account? Create one.

Username:

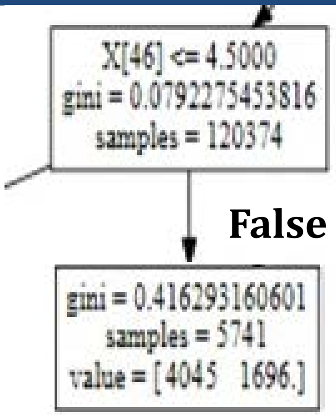
Password:

□ Remember me (up to 30 days)



```
/ $ netstat -antu
Active Internet connections (servers and established)
Proto Recv-Q Send-Q Local Address           Foreign Address
tcp        0      0 0.0.0.0:5357            0.0.0.0:*
tcp        0      0 192.168.1.1:80          0.0.0.0:*
tcp        0      0 0.0.0.0:38777           0.0.0.0:*
udp        0      0 0.0.0.0:1025            0.0.0.0:*
udp        0      0 192.168.1.1:1027        0.0.0.0:*
udp        0      0 127.0.0.1:38032         0.0.0.0:*
udp        0      0 0.0.0.0:42000           0.0.0.0:*
udp        0      0 0.0.0.0:20000           0.0.0.0:*
udp        0      0 0.0.0.0:1701            0.0.0.0:*
udp        0      0 0.0.0.0:53413           0.0.0.0:*
udp        0      0 0.0.0.0:20010           0.0.0.0:*
udp        0      0 0.0.0.0:67              0.0.0.0:*
udp        0      0 0.0.0.0:39000           0.0.0.0:*
udp        0      0 0.0.0.0:1900            0.0.0.0:*
udp        0      0 0.0.0.0:38000           0.0.0.0:*
```

Split Analysis



Others

X[46] = Sports Terms

Word	Distance to Centroid
Drawcards	0.34262
Over-age	0.41479
Multi-sports	0.43338
Multi-sport	0.44926
1908	0.46296
Honours	0.46650
Cups	0.47149
Fourth-best	0.47747
WTAs	0.48097
Player	0.48181

WTA

Women's Tennis Association

Conclusion

- Word2Vec Clustering method provides intuitive explanation behind the model's operating logic

- It can also provide reasons behind the model's performance as it identifies both the components that contribute to the performance of the model, and as well as those that deteriorate the model.

- This insight can provide deeper understanding about the given dataset, and can be used for modifying or developing a text mining model that will solve the true end goal of the given text mining task at hand.

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.